结合对象属性识别的图像场景图生成方法研究

周浩，罗廷金，崔国恒

引用本文

周浩，罗廷金，崔国恒. 结合对象属性识别的图像场景图生成方法研究[J]. 计算机科学，2024，51(11)：205-212.

ZHOU Hao, LUO Tingjin, CUI Guoheng. Scene Graph Generation Combined with Object Attribute Recognition [J]. Computer Science, 2024, 51(11): 205-212.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900013

结合对象属性识别的图像场景图生成方法研究

周　浩[1]　罗廷金[2]　崔国恒[1]
1 海军工程大学作战运筹与规划系　武汉 430033
2 国防科技大学理学院　长沙 410073
(zhouhao3075@hotmail.com)

摘　要　场景图生成在视觉场景深度理解任务中发挥着重要的作用。现有的场景图生成方法主要关注场景中对象的位置、类别以及对象之间的关系,而忽略了对象属性蕴含的丰富场景语义信息。为了将图像属性语义融入场景图,提出了一种结合对象属性识别的图像场景图生成方法。首先针对属性识别的多标签分类问题,提出了一种基于混合分类器的属性分类损失函数来进行属性识别,通过结合二值交叉熵函数训练的二分类器和改进的团组交叉熵函数训练的多分类器来实现单个属性分类的查准率和多个属性预测的查全率全面提升。其次,通过将属性识别分支与原有场景图框架进行融合,将提取的属性信息作为额外的上下文语义与对象特征进行融合后辅助对象之间关系的识别。最后,模型在VG150数据集上与多个基准模型进行了对比实验,结果表明所提模型的对象属性预测和关系识别均取得了更优的结果。

关键词:场景图生成;对象属性识别;属性融合;关系预测;多标签分类;团组交叉熵函数
中图分类号　TP391

Scene Graph Generation Combined with Object Attribute Recognition

ZHOU Hao[1],LUO Tingjin[2] and CUI Guoheng[1]
1 Department of Operational Research and Planning,Naval University of Engineering,Wuhan 430033,China
2 College of Science,National University of Defense Technology,Changsha 410073,China

Abstract　Scene graph generation(SGG) plays an important role in deep visual understanding tasks. Existing SGG methods mainly focus on the locations and categories of objects,as well as the relationship between objects,while ignoring that the object attributes also contain rich semantic information. This paper proposes a SGG model integrating with the object attributes. Firstly,to achieve multi-label object attribution recognition,we propose the composite classifiers that combine the multi-class classification trained by improved group cross entropy loss and binary classification trained by binary cross entropy loss,which can improve the accuracy and recall of multiple attribute predictions. Then,the branch of attribution recognition is fused into the SGG framework. As a kind of context information,the attribution features are fed into the relationship branch for better relationship classification. Finally,compared with several baseline models,our method has achieved better performance in both object attribute prediction and relationship recognition on VG150 dataset.

Keywords　Scene graph generation,Object attribute recognition,Attribute fusion,Relationship classifications,Multi-label learning,Group cross entropy function

1　引言

随着人工智能和深度学习技术的发展,当前计算机在目标检测与识别等任务上已经取得了巨大的成功,但在视觉场景的深度理解上仍面临一定的挑战[1]。场景图生成任务不仅能从场景中提取对象类别信息,还能学习和分析对象之间关系,并通过结构化的文本形式对图像中多个对象和多种语义进行表达,从而加深计算机对视觉场景的深度理解,因此,被广泛应用于多种高层视觉推理任务中,如视觉问答[2-3]、图像描述[4]等。

当前,场景图生成的研究主要集中在如何提高关系谓词预测的准确率上,其研究方法主要有两类。第一类方法侧重于

到稿日期:2023-09-04　返修日期:2024-04-06
基金项目:国家自然科学青年科学基金(62302516)国家自然科学基金项目(62376281);湖北省自然科学基金项目(2022CFC049);湖南省湖湘青年人才项目(2021RC3070).
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(62302516),National Natural Science Foundation of China(62376281),Natural Science Foundation of Hubei Province,China(2022CFC049) and NSF for Huxiang Young Talents Program of Hunan Province(2021RC3070).
通信作者:罗廷金(tingjinluo@hotmail.com)

拟合场景图生成中的关系数据分布[5-8]，主要通过加强上下文特征的学习、引入额外的信息或者结合先验知识等方式来生成完整的场景图。例如，Zellers 等[5]提出了 Motifs 的神经网络结构，将场景图生成分解为对象候选区域生成、对象分类识别以及关系分类识别 3 个阶段，将对象类别和关系之间的统计依赖关系隐式地蕴含在模型的框架之中，其已成为现有主流场景图框架。第二类方法侧重于通过软化数据集中不平衡的关系分布来生成降偏的场景图[9-12]，通常采用重加权和重采样等再平衡的方法、构建关系语义的层次结构树的方法、因果分析的方法等来平衡不同类别的训练数量。Tang 等[9]提出了首个在场景图任务中处理关系类别不平衡分布的模型，并有效地提升了场景图中尾部关系类别的识别准确率。

然而，当前的场景图生成模型尚没有将对象属性识别作为其主要任务，也没有充分利用对象属性中的上下文语义来辅助关系谓词的预测。对象、关系和属性是计算机深度理解视觉场景需要学习的非常重要的 3 种语义信息，而当前的场景图生成模型仅提取视觉场景中的对象和关系语义，无法学习对象的属性信息。属性是对物体或对象的一种描述[13]。为了达到更高的视觉理解层次，计算机不应该仅停留在对象关系类别层次的理解，还应该理解它们所包含的属性特征[14]。但普适性自然场景下对象的属性分类仍然是一个极具挑战的任务[15]。如图 1 所示，一方面，自然场景中的对象和属性本质上是不同的，对象是物理实体而属性是语义描述，通常同一属性在不同的物体上会呈现出不同的视觉内容，例如“空的街道”和“空的瓶子”等；另一方面，自然场景中同一对象会有多个不同的属性，例如“高的树”和“秃的树”。这种对象和属性之间多对多的对应关系极大地增加了对象属性学习的难度，因为模型不仅要理解属性的语义，更要结合上下文给同一对象特征分配多个合适的属性语义。因此，在对象和关系语义信息的基础上，将对象属性信息纳入场景图生成的任务，不仅仅是将属性作为上下文信息以辅助对象间关系识别，更是从对象、关系、属性三元组的角度完善了整个视觉场景语义的抽取，能够完成对视觉场景更深层次的语义理解。

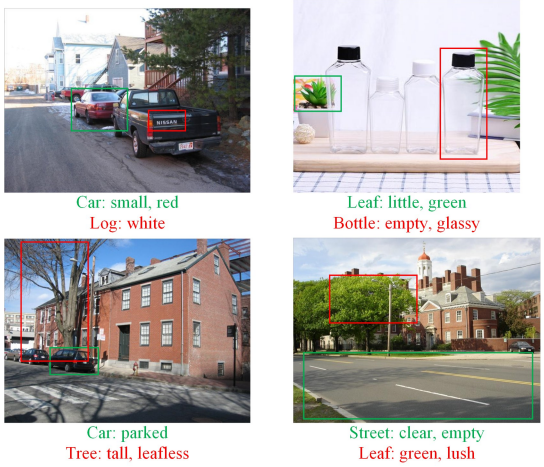

图 1 自然场景数据集中的对象属性识别所面临的挑战示例

Fig. 1 Example of challenges faced by object attributes recognition in natural scene

本文提出了结合对象属性识别的图像场景图生成模型，在自然场景对象属性识别任务的基础上，将对象属性上下文特征融入场景图模型以提高对象关系预测，并生成对象、关系和属性三要素齐全的图像场景图。首先，针对对象属性识别的多标签学习问题，构造了基于混合分类器的属性分类损失函数，通过结合团组交叉熵损失函数(Group Cross Entropy loss，G-CE loss)训练的多分类器和二元交叉熵损失函数(Binary Cross Entropy loss，BCE loss)训练的二分类器对属性特征进行分类和识别，实现单个属性分类的查准率和多个属性预测的查全率全面提升。然后，将对象属性识别分支嵌入原有场景图任务框架进行场景图生成，通过将其所学习的属性特征作为上下文特征融入对象特征，并与其他对象特征进行组合形成对象-属性对特征，之后将其作为输入通过 BiLSTM-MTree 网络对关系特征进行精炼，以提高场景图生成中的关系识别准确率。最后，与多个基准模型相比，所提模型在 VG150 数据集的属性识别和关系识别均取得了更优的结果。

2 相关工作

2.1 场景图生成

场景图生成的核心是视觉关系的检测和识别。当前场景图生成方法主要包括两类：分别是侧重于拟合关系数据分布的场景图生成和侧重于软化不平衡关系的降偏场景图生成。拟合关系分布主要通过关系特征精炼或引入先验统计特征的方式提高模型对关系谓词的预测。Zellers 等[5]提出了 Motifs 的神经网络结构，将场景图生成分解为对象候选区域生成、对象分类识别以及关系分类识别 3 个阶段。在对象分类和关系预测两个阶段使用双向 LSTM 构建了密集图来计算全局上下文。Tang 等[16]将 RoIAlign 和空间特征连接起来，构建了视觉上下文动态树模型(VCTree)。Chen 等[7]提出了知识嵌入路由网络(KERN)，试图利用对象对及其间关系的统计先验知识来解决关系分布不均匀的问题。

软化不平衡的关系分布主要围绕软化长尾分布、探索谓词语义结构和因果学习来处理关系类别不平衡[17-18]。为了缓解关系的频率受到长尾分布问题的影响，Lin 等[8]提出在 GPS-Net 中使用 log-softmax 函数来软化关系的频率分布。Tang 等[9]赋予机器反事实因果关系推理的能力，以追求无偏预测中的总直接效应(TDE)，该模型结合有偏的训练和无偏的预测来提高反事实推理能力。Zhou 等[19]从双重不平衡的角度出发，利用因果干预手段去除不平衡产生的混淆因子，并设计了偏阻损失函数以同时去除场景图中的多种偏见。

2.2 属性识别

对象属性识别的研究目前主要集中于行人属性识别[20]和人脸属性识别[21]。行人属性识别的解决方法通常遵从多标签学习或者多任务学习的框架。Liu 等[22]提出了一种定位引导网络(LGNet)，它可以对不同属性对应的区域进行定位。Zhao 等[23]提出了互补的端到端递归卷积和递归注意力模型，递归卷积模型利用卷积 LSTM 单元挖掘不同属性组之间的相关性，递归注意力模型利用空间局部性和注意力相关性来提高行人属性识别性能。Yang 等[24]提出了一个分层特征嵌入框架(HFE)，通过结合属性和行人身份 ID 信息来学习细粒度特征嵌入。受到深度神经网络在对象分类任务上取得

显著成功的鼓舞，一些研究尝试通过组合两个识别对象和属性的判别模型来识别属性-对象对。Nan 等[14]为了探索属性和对象的内在关系以及内在的属性-对象表示，提出了一种具有编解码机制的生成模型。该模型能够在统一的端到端网络中连接相关视觉和语言信息，并挖掘属性和对象的关系，从而提高零样本下属性-对象对的检测精度。

Wei 等[15]为了识别不可见的属性-对象对，提出了一种新的对抗性细粒度合成学习模型。此外，Yamaguchi 等[25]研究服装领域的属性识别，主要考虑对象间或属性间兼容的识别问题。该模型考虑人体、服装、属性之间特定位置的外观特征，并基于一个条件随机场寻求最可能的服装搭配组合。

2.3　多标签学习

当前主流的解决多标签学习的方法主要有 3 类[26-27]：1）利用图像中对象局部信息融合并生成多个预测分布[28]；2）利用视觉注意力机制来生成注意力区域序列，根据区域序列顺序预测类别分布；3）利用标签之间的依赖信息，这种方式是目前多标签图像分类问题的主要研究方向，通过探索标签之间的相关性来提升多标签分类的准确度[29]。对于图像分类而言，对象标签往往存在某些类别之间的自然依赖性。例如“鼠标”和“键盘”通常一起出现，当已经正确分类出“鼠标”类别时，那么这张图片中存在“键盘”类别的可能性也会相应提高[30]。然而，在对象属性分类中，这种类别之间的相互依赖性并不十分明显。例如“黑色的”汽车和“老旧的”汽车，虽然都是描述同一对象，但是“黑色”和“老旧”之间并不存在任何因果联系。

随着深度学习技术的发展，基于损失函数设计的多标签分类方法也取得了一定的成功。为了惩罚被错误分类的正样本，Weston 等[31]提出了 WARP 损失函数，将更高的权重暴力地分配给目标类别。Li 等[32]提出了基于 LSEP 损失函数和目标类别数量预测模块来解决无法直接输出预测结果的问题。Sohn 等[33]从度量学习的角度提出了多分类的 N-pair 损失函数，通过多个负样本的方式来优化训练。Su 等[34]通过改进 softmax 函数提出了无需阈值划分的 ZLPR 损失函数，结合标签之间的依赖性关系促使正标签类别团组得分都大于 0。不同于上述方法，针对属性识别，若仅单一地使用团组标签类别分类方式，则模型的收敛难度更大且预测结果具有一定的不确定性。因此，本模型在团组分类的基础上，针对所有属性类别，额外引进 BCE 损失函数进行二分类预测。这种组合设计不仅能够消除单一分类预测结果的随机性，同时还能够兼顾单个正类别标签和多个正类别标签的预测准确率。

3　结合对象属性识别的图像场景图生成

图 2 给出了本文所提出的结合对象属性识别的图像场景图生成模型框架。在本模型中，首先提出了基于混合分类器的属性分类损失函数以缓解对象属性识别所面临的多标签问题，然后将对象属性与场景图框架进行融合，并将对象属性特征作为一种额外的上下文特征融入关系特征学习以提升关系预测的性能。

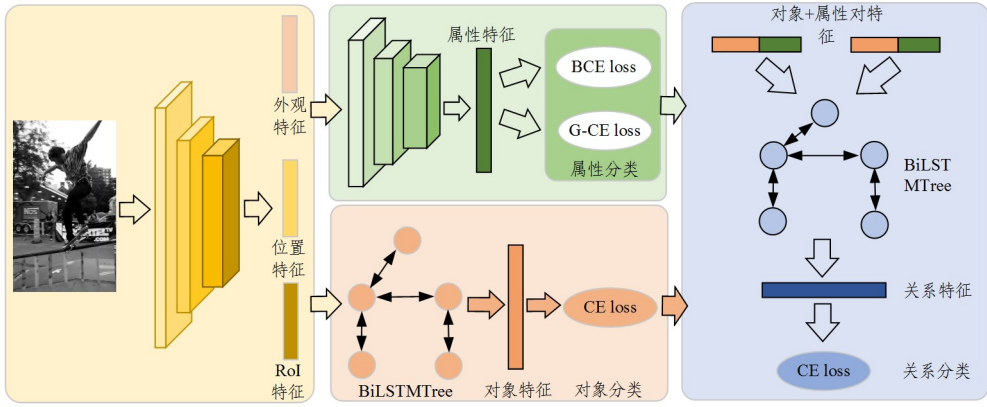

图 2　结合对象属性识别的图像场景图生成模型框架

Fig. 2　Framework of scene graph generation model combined with object attribute recognition

3.1　基于混合分类器的对象属性分类识别

在一般的图像分类等多标签学习任务中，通常可以借助各个类别之间标签的统计共现性来提升多标签学习的预测结果。然而，属性信息属于语义描述，不同属性类别（颜色、姿态、尺寸、情绪等）之间的相互依赖性通常更低，例如对于汽车，“黑色的”和“老旧的”两种属性之间并没有必然的联系。因此，不同于其他多标签学习方法建立标签之间的内在联系，本文主要从对象属性分类的训练方式和损失函数的设计两方面来解决属性的多标签分类问题。

给定一张图像 I，包含对象 $O=\{o_1,o_2,\cdots,o_n\}$。对于每一个对象 o_i 有 k 个属性 $A_{o_i}=\{a_1,a_2,\cdots,a_k\}$，其中 $a_i\in\{T_1,T_2,\cdots,T_N\}$ 表示数据集中总共有 N 类属性标签。最朴素的多标签分类任务使用的损失函数是 sigmoid 二元函数分类器和二值交叉熵损失函数 BCE Loss，其基本思想是将多标签分类问题转换为多个单标签二分类问题。将模型的 N 维特征输出 $\{x_1,x_2,\cdots,x_N\}$ 分别输入给 N 个 sigmoid 函数分类器。对于第 i 维特征，其属于第 i 类属性标签的预测概率为：

$$p_i=\frac{1}{1+\mathrm{e}^{-x_i}} \tag{1}$$

在训练阶段，模型采用二值交叉熵损失函数来计算预测值和真实值之间的损失。对于 N 个分类器，其平均损失为：

$$loss^{bce}_{\text{multilabel}}=-\frac{1}{N}\sum_{i=1}^{N}(y_i\log p_i+(1-y_i)\log(1-p_i)) \tag{2}$$

其中，y_i 为真实值，通常取 0 或者 1。在对象属性的多标签分类任务中，sigmoid 函数分类器和二值交叉熵损失函数的模型

训练方式会带来不平衡的问题。例如在 VG150 数据集中包含了 200 类属性类别和 1 类无属性类别，而平均每个对象的属性标签小于 5 个。因此，在计算损失函数时，loss 值会被大量的无属性类别所主导，从而导致模型倾向于将对象属性预测为无属性类别。

为了缓解这种训练方式所带来的不平衡，本文额外引入 softmax 多元函数分类器和改进的团组交叉熵损失函数来指导模型的训练。对于模型的 N 维特征输出 $\{x_1, x_2, \cdots, x_N\}$，其经过 softmax 函数分类器的类别预测概率为：

$$p_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \tag{3}$$

在 softmax 函数分类器和交叉熵损失函数的模型训练下，一般多分类的损失可以表示为：

$$loss^{ce} = -y_i \log p_i = -\log \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

$$= \log \sum_{j=1}^{N} e^{x_j - x_i} = \log\left(1 + \sum_{j=1, j\neq i}^{N} e^{x_j - x_i}\right) \tag{4}$$

根据 max 函数的光滑近似，有：

$$\log\left(1 + \sum_{j=1, j\neq i}^{N} e^{x_j - x_i}\right) \approx \max(0, x_1 - x_i, \cdots, x_{i-1} - x_i, x_{i+1} - x_i, \cdots, x_N - x_i) \tag{5}$$

从上述公式可以得出，基于 softmax 函数的多类交叉熵损失函数的目标是实现目标类的概率得分大于所有非目标类的概率得分。那么对于多标签分类，相当于在 N 个类别中选取 k 个属性，损失函数的目的是使这 k 个标签所对应的概率得分大于其他概率得分，因此改进后的团组交叉熵损失函数可以表示为：

$$loss^{ce}_{\text{multilabel}} = \log\left(1 + \sum_{i\in pos, j\in neg} e^{x_j - x_i}\right)$$

$$= \log\left(1 + \sum_{j\in neg} e^{x_j} \sum_{i\in pos} e^{-x_i}\right) \tag{6}$$

其中，$i \in pos$ 为正样本集合，也就是对象的真实属性标签；$j \in neg$ 为负样本集合。

总的来说，如图 3 所示，基于 sigmoid 函数的二值交叉熵损失函数的目的是使对象的单个真实属性标签的预测概率更高，但是会受到不平衡分布的影响。而基于 softmax 函数的多分类团组交叉熵损失函数的目的是使对象的多个真实属性标签的整体预测概率高于其他负样本属性标签，不会导致不平衡分布。两种多标签训练的损失函数是互补的。因此，本模型通过融合这两种损失函数的优势来学习对象属性分类，其损失函数可以表示为：

$$loss_{\text{attr}} = loss^{bce}_{\text{multilabel}} + loss^{ce}_{\text{multilabel}} \tag{7}$$

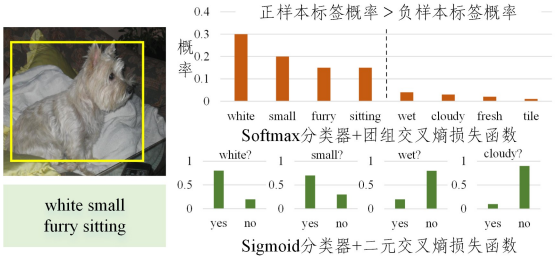

图 3 属性多标签学习中损失函数的作用示意图

Fig. 3 Role of loss function in attribute multi-label learning

通过融合两种损失函数，模型在训练中既能保证对象属性单个标签预测的查准率，又能提升整体多标签预测的查全率。

3.2 融合对象属性特征的对象关系识别

现有的场景图生成模型通常通过引入对象空间特征和对象类别特征等额外的上下文信息来增强关系预测的性能，例如以 Motifs[5] 和 VCTree[16] 模型为基础骨干网络的场景图生成模型。除此之外，对象属性作为一种上下文特征，能够提供部分对象的姿态信息，例如弯曲的、直立的、缩缩的等。这些对象姿态信息蕴含了对象的部件所处的状态，能够在关系识别中有力地促进模型对对象之间关系的精准识别。因此，本文在 VCTree 骨干网络的基础上，创新性地将对象属性分支与场景图框架进行融合，通过对象属性特征和对象特征拼接形成对象-属性，并将其特征输入 BiLSTMTree 网络模型进行关系特征学习与精炼，最终实现对象之间关系的分类与识别。

对于输入图像 I，首先通过以 ResNeXt-101-FPN 为主干网络的 Fast RCNN 模型提取多层次的对象外观特征 $\boldsymbol{f}^{\text{appear}}$，并通过 RPN 网络生成对象候选框集合 B。然后通过 RoI 网络提取对象的 RoI 特征 $\boldsymbol{f}^{\text{roi}}$，通过类似结构的属性网络提取对象属性特征 $\boldsymbol{f}^{\text{attr}}$，它们在网络中都是 2048 维的向量。对象属性特征提取的过程可以表示为：

$$\boldsymbol{f}^{\text{attr}} = g_a(\boldsymbol{f}^{\text{roi}}, B) \tag{8}$$

其中，$g_a(,)$ 表示属性特征提取网络。在多标签分类任务的指导下，通过对象属性特征对对象属性进行分类与预测。

$$\{a_1, a_2, \cdots, a_k\} = multilabel(\boldsymbol{f}^{\text{attr}}) \tag{9}$$

对于对象分类，本文将在 Faster RCNN 网络所提取的对象特征的基础上，进一步通过 BiLSTMTree[16] 对对象特征进行精炼，从而可以得到精炼后的对象特征 $\boldsymbol{f}^{\text{obj}}$：

$$\boldsymbol{f}^{\text{obj}} = BiLSTMTree(\boldsymbol{f}^{\text{roi}}, B) \tag{10}$$

为了构建 BiLSTMTree 中的二叉树结构，模型计算所有对象对之间的任务相关性得分，并通过 prim 算法从得分矩阵中计算得到二元分支的最大生成树。基于 BiLSTMTree 进行特征融合不仅能够编码视觉场景中不同对象之间的平行或者层次结构关系以及提高特征传递的效率，还能够针对不同图像或者任务，在对象之间传递更多特定内容/任务的特征消息。

同时，对象属性特征作为一种上下文特征，有利于场景图生成中的对象分类和关系分类。因此，本文将对象属性分支嵌入原有的场景图生成框架来辅助提升对象和关系的预测。在场景图分支中，精炼后的对象特征和对象属性特征拼接形成对象-属性特征：

$$\boldsymbol{f}^{\text{fusion}} = C(\boldsymbol{f}^{\text{obj}}, \boldsymbol{f}^{\text{attr}}) \tag{11}$$

其中，$C(,)$ 表示特征的拼接。场景图中的关系分支在获取融合的对象-属性特征之后，模型将构造二叉树结构并通过 BiTreeLSTM 网络学习关系分类的特征 $\boldsymbol{f}^{\text{rel}}_{i\to j}$：

$$\boldsymbol{f}^{\text{rel}}_{i\to j} = BiLSTMTree(\boldsymbol{f}^{\text{fusion}}_i, f^{\text{fusion}}_j) \tag{12}$$

最后，场景图中的对象分类和关系预测的训练都在一般交叉熵损失函数的指导下进行训练。

4　实验

4.1　实验设置

4.1.1 数据集

本文在视觉基因组数据集(VG)[35]上测试对象属性学习模型。VG数据集不仅包含对象和关系的标注信息，还包括对象属性、区域描述等标注信息，是一个广泛应用于视觉语言任务的数据集。原始VG数据集包含108077张图片，其中共标注对象属性280万条。平均每张图片有35个对象标注和26个属性标注，且每个对象可能没有或者有多个相关的属性信息。为了将属性特征嵌入场景图生成框架，本文主要在VG数据集的变种数据集VG150上进行模型评估和测试。VG150数据集中包含最频繁的150类对象类别和200类属性，属性的类型主要包括：颜色(如黄色)、尺寸(如大、小)、姿态(如弯曲的)、状态(如透明的)和情绪(如高兴的)等。

4.1.2　任务设置和测度

对于场景图生成，本文按照文献[5]和文献[9]中的方法进行任务设置，主要包括：1)谓词分类(Predicate Classification，PredCls)，给定对象类别和对象边框，模型预测所有对象对之间的关系标签。2)场景图分类(Scene Graph Classification，SGCls)，给定对象边框，模型预测对象类别和所有对象对之间的关系标签。与其他模型类似，本文使用Recall@K(R@K，$K=20,50,100$)，mean Recall@K(mR@K，$K=20,50,100$)和Zero-Shot Recall@K(ZSR@K，$K=20,50,100$)。Recall@K是场景图生成任务中传统的测度，它计算前K个预测结果中被正确召回的真实实例的比例。mean Recall@K平等地对待每一个关系类别，计算每个关系类别Recall@K的平均值，相对能够更好地反映尾部类别的性能。Zero-Shot Recall@K主要反映那些从未在训练集中出现过的“主语对象-关系-宾语对象”三元组的性能，它能够很好地反映模型的泛化性和稳定性，也是因果学习中一个重要的评估测度。

对于属性识别，目前在VG数据集上并没有独立的任务设置和相关评价指标。因此，基于对象分类和关系识别的任务和评估测度，因此本文主要通过准确率和召回率两个指标来测试模型的性能。由于同一个目标可能对应多个不同属性，本文在计算模型的属性识别准确性时主要针对Top1的预测结果进行计算。只要Top1的预测结果在多标签属性范围内，则认为模型分类识别正确。为了评估模型对多标签属性预测的能力，本文主要采用Recall@K($K=5,10$)计算模型前5个预测结果或前10个预测结果中对所有属性标签的查全率。具体地，准确率的计算式为：

$$Pre_{\mathrm{Top1}}=\frac{1}{M}\sum_{m=1}^{M}num(R_{\mathrm{Top1}}\cap Lables) \tag{13}$$

召回率的计算方法为：

$$Recall_{\mathrm{Top}K}=\frac{1}{M}\sum_{m=1}^{M}\frac{num(R_{\mathrm{Top}K}\cap GroundTruth)}{num(GroundTruth)} \tag{14}$$

除此之外，本文还通过定性分析对对象属性分类和属性特征对场景图生成的影响进行研究。

4.1.3　实验细节

对于场景图关系识别部分的模型分类，与Tang等的方法[9]类似，本节采用以ResNeXt-101-FPN[36]为主干网络的Faster R-CNN框架作为基本目标检测器来生成对象候选边框和提取对象的RoI特征。模型在SGD算法的优化下训练，其中动量设置为0.9，batch size大小为12，初始学习率为0.001。属性分类的模型训练与Faster R-CNN框架训练同时进行，对象检测器在文献[9]中的参数设置下重新训练。

4.2　实验结果定量分析

4.2.1　与其他场景图模型的性能对比

为了证明所提模型的有效性，本节首先将本文模型与最近其他场景图生成模型进行对比，主要包括IMP+模型[37]、FREQ模型[5]、Motifs模型[5]、VTransE模型[38]、KERN模型[7]、VCTree模型[16]和GPS-Net模型[8]等，对比结果如表1所列。

表1　与其他场景图生成模型在R@K测度上的对比结果

Table 1　Comparison results of our model with other SGG models on R@K

(%)

方法	PredCls				SGCls			
	R@20	*R*@50	*R*@100	*mean*	*R*@20	*R*@50	*R*@100	*mean*
IMP+[37]	52.7	59.3	61.3	57.8	31.7	34.6	35.4	33.9
FREQ[5]	53.6	60.6	62.2	58.8	29.3	32.3	32.9	31.5
Motifs[5]	58.5	65.2	67.1	63.6	32.9	35.8	36.5	35.1
VTransE[38]	59.0	65.7	67.6	64.1	35.4	38.6	39.4	37.8
KERN[7]	—	65.8	67.6	—	—	36.7	37.4	—
VCTree[16]	59.8	66.2	68.1	64.7	37.0	40.5	41.4	39.6
GPS-Net[8]	60.7	66.9	68.8	65.5	36.1	39.2	40.1	38.5
本文模型	**60.9**	**67.3**	**69.2**	**65.8**	**42.4**	**46.2**	**47.3**	**45.3**

由表1可以发现，引入对象属性信息，本文模型在场景图生成任务中取得了更好的性能。例如在PreCls任务上，本文模型在平均R@K测度上取得了65.8%的性能，比VCTree和GPS-Net模型分别提高了1.1%和0.3%，说明属性上下文信息对对象之间关系的识别具有促进作用。特别对于更具挑战性的SGCls任务，相比VCTree和GPS-Net模型，本文模型在平均mR@K测度上分别提高了5.7%和6.8%。

4.2.2　属性特征对场景图生成的影响

为了充分探索属性特征对场景图生成的影响，本文将属性分支分别嵌入Motifs，VTransE和VCTree模型，结果如表2所列。

表 2 在 R@K 测度上对象属性对场景图生成的影响

Table 2 Effect of object attributes on SGG on R@K

(%)

方法	PredCls				SGCls			
	R@20	*R*@50	*R*@100	*mean*	*R*@20	*R*@50	*R*@100	*mean*
Motifs	58.5	65.2	67.1	63.6	32.9	35.8	36.5	35.1
Motifs+属性	**60.0**	**66.2**	**67.6**	**64.6**	**36.3**	**39.6**	**41.8**	**39.2**
VTransE	59.0	65.7	67.6	64.1	35.4	38.6	39.4	37.8
VTransE+属性	**59.4**	**65.9**	**68.3**	**64.5**	**39.0**	**42.7**	**44.6**	**42.1**
VCTree	59.8	66.2	68.1	64.7	37.0	40.5	41.4	39.6
VCTree+属性	**60.9**	**67.3**	**69.2**	**65.8**	**42.4**	**46.2**	**47.3**	**45.3**

表 2 列出了模型分别在有属性分支和无属性分支情况下 PredCls 和 SGCls 任务上各个测度的性能。从表中可以发现，加入对象属性分支能够提升模型在 SGCls 任务 3 个测度上的性能。与原始 Motifs 模型、VTransE 模型和 VCTree 模型相比，本文模型在平均 R@K 测度的性能分别提高了 4.1%，4.3%和 5.7%。在 PredCls 任务上，数据集中关系标签的有限理性等原因导致各个模型在此任务上的 R@K 测度性能达到了相同的瓶颈。而加入属性分支之后，各模型在 PredCls 任务的 R@K 测度上的性能仍能取得提升或保持同一水平。因此，对象属性特征作为对象的重要语义特征之一，能够加深模型对视觉场景的理解以及强化对象的特征，并能够促进场景图生成任务中的对象分类以及关系预测。

为进一步分析对象属性对尾部类别和模型泛化性能的影响，本文在表 3 中列出了模型在 mR@K 和 zR@K 测度上与原始 VCTree 模型的对比结果。

表 3 模型在 mR@K 和 zR@K 测度上与 VCTree 模型的对比结果

Table 3 Comparison results of our model with VCTree model on mR@K and zR@K

(%)

任务	测度	VCTree				本文模型			
		K=20	*K*=50	*K*=100	mean	*K*=20	*K*=50	*K*=100	mean
PredCls	mR@K	11.7	14.9	16.1	14.2	13.5	16.3	17.8	**15.9**
	zR@K	4.6	10.8	14.3	9.9	11.6	18.2	21.3	**17.0**
SGCls	mR@K	6.2	7.5	7.9	7.2	8.5	10.3	11.0	**9.9**
	zR@K	0.5	1.9	2.9	1.8	1.5	3.1	4.2	**2.9**

从表 3 中可以发现，对于更能反映模型对尾部关系类别学习的 mR@K 测度，加入属性特征能够提升 mR@K 测度的性能，例如在 PredCls 任务上，mR@20 测度性能从 11.7%提高到了 13.5%，在 SGCls 任务上，mR@100 测度性能从7.9%提高到了 11.0%。而对于能够反映模型泛化性能的 zR@K 测度来说，加入属性特征能够极大地提升其性能，在 PredCls 任务上，zR@100 测度性能从 14.3%提高到了 21.3%，相对提高了约 49.0%。这一巨大的提升说明模型在引入额外的对象属性语义特征之后，能够增强模型的鲁棒性和泛化性，也能够有效地辨别未见到过的对象—关系对。

为了全面衡量模型的性能，部分模型取消了一个对象对只能预测一个关系的约束。因此，表 4 中进一步列出了本模型在无约束条件下 n-R@K，n-mR@K 和 n-zR@K 测度上的结果。

表 4 模型在无约束条件下 n-R@K，n-mR@K 和 n-zR@K 测度上的结果

Table 4 Results of our model on n-R@K, n-mR@K and n-zR@K without constrains

(%)

任务	测度	本文模型		
		K=20	*K*=50	*K*=100
PredCls	n-R@K	67.9	82.3	89.4
	n-mR@K	19.8	32.9	44.6
	n-zR@K	15.1	29.2	43.1
SGCls	n-R@K	47.8	57.2	61.6
	n-mR@K	13.3	22.1	29.7
	n-zR@K	1.7	5.7	11.8

4.2.3 属性识别性能

为了验证本文所提出的复合分类损失函数对多标签属性分类的影响，本节计算属性模型的 TOP1 准确率和 R@K（$K=5,10$）召回率，并与 softmax 损失函数和二元交叉熵损失函数进行对比，相关结果如表 5 所列。

表 5 不同分类损失函数下的属性性能对比

Table 5 Comparison of attribute performance with different classification loss functions

(%)

	Top1 *acc*	*R*@5	*R*@10
Softmax loss	23.2	71.6	84.4
BCE loss	52.5	72.8	85.2
本文模型	**60.8**	**75.2**	**86.3**

4.3 定性分析和可视化展示

本节主要对对象属性的学习结果以及在场景图中融合属性分支的结果进行可视化展示和定性分析。

图 4 展示了对象属性学习的结果，图中属性的可视化主要选取模型所预测的最可能属性。从图 4 中展示的结果来看，在结合改进 softmax 函数的多分类交叉熵损失函数和 sigmoid 函数的二值交叉熵损失函数的训练下，多标签学习的属性模型所预测的最可能属性类别基本是合理的，特别是在颜色、形状、大小和材质等对象属性方面具有较好的预测性能，例如“green tree”“square plate”“small boy”和“wooden table”。

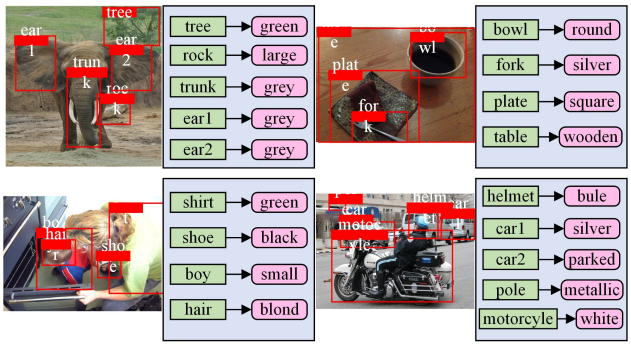

图 4　对象属性学习的可视化结果展示

Fig. 4　Visualization results of object attribute learning

图 5 展示了将对象属性分支融入场景图生成模型的可视化结果，图中对象的属性标签为模型所输出的最可能属性。从图中可以发现，在场景图生成模型中的对象属性预测依然十分出色。模型能够对部分对象所处的状态进行更加详细的描述，并且通过对象属性的标记，模型能够更好地区分同一类别的不同对象实例，例如“sitting man”和“standing man”、“small boy”和“standing boy”等。更重要的是，通过在场景图生成模型中融入对象属性信息，模型能够生成语义更加丰富的结构化场景图。例如，基于结构化的场景图能够清晰地推理得到第一个场景中存在“一个年轻的男人戴着格子的帽子拿了一把紫色的伞”、第二个场景中存在“坐着的男人穿着灰色的衣服拿着黑色的手机”、第三个场景中存在“站着的男孩戴着蓝色的帽子穿着条纹的夹克拿着滑板”，以及第四个场景中存在“狭窄的街道上停着一辆带有黑色篮子的自行车”等。通过融合了对象、属性、关系的结构化场景图，计算机能够加深对场景的理解，并且能够基于结构化语义表示从表层感知向深层认知转化，实现更高层次的视觉推理等任务。

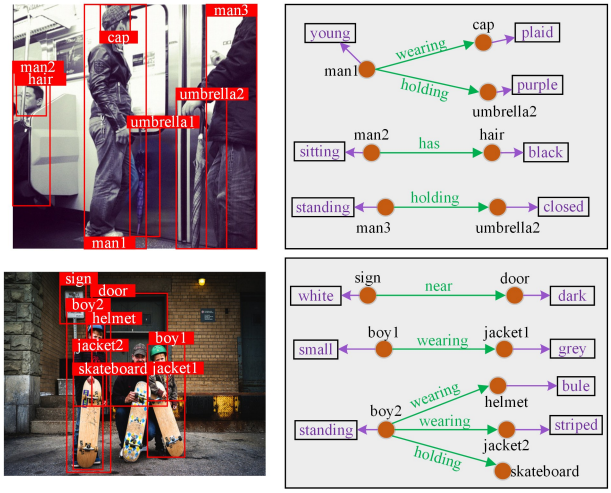

图 5　结合属性识别的场景图生成可视化结果

Fig. 5　Visualization results of SGG combined with attribute recognition

结束语　本文提出了一种结合对象属性识别的场景图生成方法，该方法能够提取场景中的对象、属性和关系语义并形成结构化的表达形式。为了实现模型对自然场景中的对象属性的多标签预测，本文设计了基于混合分类器的属性识别模型来提高查全率和查准率，并将属性分支嵌入场景图生成模型，利用对象属性的上下文特征增强模型对关系预测的能力。模型在 VG150 数据集上的属性预测和关系识别效果均优于其他基准模型，且消融性实验也证明了各模块的有效性。未来，将进一步探索基于对象、关系、属性的场景图生成模型在多模态知识图谱中的生成和应用。

参 考 文 献

[1] HOU H, ZHANG J, LUO T, et al. Debiased Scene Graph Generation for Dual Imbalance Learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4274-4288.

[2] HUDSON D A, MANNING C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering [C]// Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition. 2019: 6700-6709.

[3] ZOU Y, DU S, TENG F, et al. Visual Question Answering Model Based on Multi-modal Deep Feature Fusion[J]. Computer Science, 2023, 50(2): 123-129.

[4] WU Q, SHEN C, WANG P, et al. Image captioning and visual question answering based on attributes and external knowledge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1367-1381.

[5] ZELLERS R, YATSKAR M, THOMSON S, et al. Neural motifs: Scene graph parsing with global context[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5831-5840.

[6] WOO S, KIM D, CHO D, et al. Linknet: Relational embedding for scene graph[C]// Proceedings of the Advances in Neural Information Processing Systems. 2018.

[7] CHEN T, YU W, CHEN R, et al. Knowledge-embedded routing network for scene graph generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6163-6171.

[8] LIN X, DING C, ZENG J, et al. Gps-net: Graph property sensing network for scene graph generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3746-3753.

[9] TANG K, NIU Y, HUANG J, et al. Unbiased scene graph generation from biased training[C]// Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2020: 3716-3725.

[10] LI R, ZHANG S, WAN B, et al. Bipartite graph network with adaptive message passing for unbiased scene graph generation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11109-11119.

[11] YAN S, SHEN C, JIN Z, et al. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 265-273.

[12] TAO L, MI L, LI N, et al. Predicate correlation learning for scene graph generation[J]. IEEE Transactions on Image Processing, 2022, 31: 4173-4185.

[13] SU C, ZHANG S, XING J, et al. Deep attributes driven multi-camera person re-identification[C]// Computer Vision－ECCV

2016:14th European Conference, Amsterdam, The Netherlands, October 11—14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016:475-491.
[14] NAN Z, LIU Y, ZHENG N, et al. Recognizing unseen attribute-object pair with generative model[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:8811-8818.
[15] WEI K, YANG M, WANG H, et al. Adversarial fine-grained composition learning for unseen attribute-object recognition [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:3741-3749.
[16] TANG K, ZHANG H, WU B, et al. Learning to compose dynamic tree structures for visual contexts[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:6619-6628.
[17] LI L, CHEN L, HUANG Y, et al. The devil is in the labels: Noisy label correction for robust scene graph generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:18869-18878.
[18] YU J, CHAI Y, WANG Y, et al. Cogtree: Cognition tree loss for unbiased scene graph generation[C]// Proceedings of the International Joint Conference on Artificial Intelligence. 2021:1274-1280.
[19] ZHOU H, ZHANG J, LUO T, et al, Debiased Scene Graph Generation for Dual Imbalance Learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4):4274-4288.
[20] YANG X, YIN K, HOU S, et al. Person Re-identification Based on Feature Location and Fusion[J]. Computer Science, 2022, 49(3):170-178.
[21] LAI X, CHEN S, YAN Y, et al. Survey on Deep Learning Based Facial Attribute Recognition Methods[J]. Journal of Computer Research and Development, 2021, 58(12):2760-2782.
[22] LIU P, LIU X, YAN J, et al. Localization guided learning for pedestrian attribute recognition[J]. arXiv:1808.09102, 2018.
[23] ZHAO X, SANG L, DING G, et al. Recurrent attention model for pedestrian attribute recognition [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1):9275-9282.
[24] YANG J, FAN J, WANG Y, et al. Hierarchical feature embedding for attribute recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:13055-13064.
[25] YAMAGUCHI K, OKATANI T, SUDO K, et al. Mix and Match: Joint Model for Clothing and Attribute Recognition [C]// BMVC. 2015:4.
[26] TAREKEGN A N, GIACOBINI M, MICHALAK K. A review of methods for imbalanced multi-label classification[J]. Pattern Recognition, 2021, 118:107965.
[27] KIM Y, KIM J M, AKATA Z, et al. Large loss matters in weakly supervised multi-label classification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:14156-14165.
[28] ZHANG Y, WANG Y, LIU X Y, et al. Large-scale multi-label classification using unknown streaming images[J]. Pattern Recognition, 2020, 99:107100.
[29] WEVER M, TORNEDE A, MOHR F, et al. AutoML for multi-label classification: Overview and empirical evaluation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(9):3037-3054.
[30] LI J, LI P, HU X, et al. Learning common and label-specific features for multi-Label classification with correlation information [J]. Pattern Recognition, 2022, 121:108259.
[31] WESTON J, BENGIO S, USUNIER N. Wsabie: Scaling up to large vocabulary image annotation. [C]// Proceedings of the International Joint Conference on Artificial Intelligence. 2011:2764-2770.
[32] LI Y, SONG Y, LUO J. Improving pairwise ranking for multi-label image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:3617-3625.
[33] SOHN K. Improved deep metric learning with multi-class n-pair loss objective[C]// Proceedings of the Advances in Neural Information Processing Systems. 2016.
[34] SU J, ZHU M, MURTADHA A, et al. Zlpr: A novel loss for multi-label classification[J]. arXiv :2208.02955, 2022.
[35] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123:32-73.
[36] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:1492-1500.
[37] XU D, ZHU Y, CHOY C B, et al. Scene graph generation by iterative message passing[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:5410-5419.
[38] ZHANG H, KYAW Z, CHANG S F, et al. Visual translation embedding network for visual relation detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:5532-5540.

ZHOU Hao, born in 1993, Ph.D, lecturer, is a member of CCF (No. T6933M). His main research interests include scene graph generation, image understanding and causal inference.

LUO Tingjin, born in 1989, Ph.D, professor, master supervisor, is a senior member of CCF (No. C4089S). His main research interests include weakly supervised learning, data mining and machine learning etc.

(责任编辑:何杨)