doi:10.3969/j.issn.1672-1292.2022.02.008

基于语义连通图的场景图生成算法

姜有亮[1],张锋军[2],沈沛意[1,3],张　亮[1,3]

(1.西安电子科技大学计算机科学与技术学院,陕西 西安 710071)
(2.中国电子科技网络信息安全有限公司,四川 成都 610041)
(3.西安电子科技大学西安市智能软件工程重点实验室,陕西 西安 710071)

[摘要]　提出了基于语义连通图的场景图生成算法. 将关系检测过程分为关系建议和关系推理两步;以目标检测算法得到的候选对象为节点集合,构建一个全连接图;使用物体的类别信息和相对空间关系计算物体之间存在关系的概率;通过设置阈值来删除图中的无效连接,得到稀疏的语义连通图;使用图神经网络聚合物体节点的特征进行聚合,融合上下文信息. 根据语义连通图的连接关系,结合更新后的主语和宾语特征以及两个物体联合区域的特征,构建关系特征,预测图中的每条边对应的关系类别.

[关键词]　场景图生成,图卷积神经网络,目标检测,视觉关系检测,场景语义理解
[中图分类号]TP311　[文献标志码]A　[文章编号]1672-1292(2022)02-0048-08

Scene Graph Generation Based on Semantic Connected Graph

Jiang Youliang[1], Zhang Fengjun[2], Shen Peiyi[1,3], Zhang Liang[1,3]

(1.School of Computer Science and Technology, Xidian University, Xi'an 710071, China)
(2.China Electronics Technology Cyber Security Co., Ltd., Chengdu 610041, China)
(3.Xi'an Key Laboratory of Intelligent Software Engineering, Xidian University, Xi'an 710071, China)

Abstract: A scene graph generation algorithm based on semantic connected graph is proposed. Relationship detection process can be divided into two steps: relationship advice and reasoning. The detected object candidates are used as nodes to build one fully connected diagram. Object category and relative space information are used to calculate the relationship probability between objects. A threshold is utilized to remove the invalid connection and build the sparse semantic connected graph. A graph neural network method is used to aggregate the node feature representation with contextual information. At last, the relation category corresponding to each edge of the graph is classified according to the connectivity of the semantic connectivity graph by combining the updated feature representations of the subject and object, and the characteristics of the joint region of the two objects.

Key words: scene graph generation, graph convolution network, object detection, visual relationship detection, scene semantic understanding

深度学习技术在目标检测、图像语义分割等基本视觉理解任务上取得了显著的成果,但对于对视觉信息的整体感知和有效表达仍然不够. 人们希望计算机可以理解图像中更深层次的语义信息. Johnson 等人提出了场景语义结构图(scene graph,简称场景图)[1],这是一种对特定场景中语义信息的结构化文本表示,其中,节点表示物体,边表示物体之间的关系. 场景图生成是将输入的图像解析成一种结构化的文本表示,其核心任务是检测视觉关系,即检测关系三元组(主体、关系、宾语)[2-4]. 目前主流的场景图生成方法遵循两阶段流程[5],第一阶段由目标检测算法得到图像中的物体集合,并根据物体的信息提取关系特征,然后执行一个分类任务以确定每个物体对之间的关系.

场景图生成的基础任务是目标检测. 现有的目标检测算法有传统检测算法和基于深度学习的检测算

收稿日期:2021-08-31.
基金项目:国家自然科学基金项目(62072358)、国家重点研发计划项目(2020YFF0304900,2019YFB1311600)、陕西省重点研发计划(2018ZDXM-GY-036).
通讯作者:张亮,教授,博士生导师,研究方向:场景感知与理解、人机交互、嵌入式系统. E-mail:liangzhang@xidian.edu.cn

法. 传统目标检测算法多基于滑动窗口的框架或是根据特征点进行匹配,已难以满足人们对目标检测效果的要求. 随着深度学习在图像分类任务上取得巨大进展,基于深度学习的目标检测算法逐渐成为主流,并取得了极大的成功. 目前基于深度学习模型的目标检测算法可分为两大类:一种是基于区域建议的目标检测算法,其将目标检测问题划分为两个阶段,第一阶段产生候选区域(region proposals),包含目标大概的位置信息,第二阶段对候选区域进行物体分类和位置精修,该算法的典型代表有 R-CNN[6]、Fast R-CNN[7]和 Faster R-CNN[8]等;另一种是基于回归学习的目标检测算法,其不需要区域建议阶段,可直接预测物体的类别概率和位置坐标值,比较典型的算法有 YOLOv1-v3[9-10]、SSD[11]和 RetinaNet[12]等.

图卷积神经网络(graph convolutional neural network,GCN)的出现是为了在非欧几里得结构数据上进行卷积操作. 目前图上的卷积定义可分为两类[13]:一是基于频域或谱域(spectral domain)的图卷积,通过傅里叶变换将结点映射到频域空间,通过在频域空间上做乘积来实现时域上的卷积,最后再将做完乘积的特征映射回时域空间;另一种是基于空间域(spatial domain)的图卷积,通过聚合邻居节点的信号对节点的特征做变换. 基于空间域的图卷积神经网络模型是近年来研究的热点,代表性的模型有 GCNConv[14]、GAT[15]和 GraphSAGE[16]等.

本文提出一种基于语义连通图的场景图生成算法,将关系检测过程分为关系建议和关系推理两步. 以目标检测算法得到的候选对象为节点集合,构建一个全连接图,并使用物体的类别信息和相对空间关系计算物体之间存在关系的概率,通过设置阈值来删除图中无效连接,得到稀疏的语义连通图. 为了融合物体的上下文信息,使用图神经网络对图节点的特征进行聚合. 最后根据语义连通图的连接关系,结合更新后的主语和宾语特征,以及两个物体联合区域的特征构建关系特征,预测语义图中的每条边对应的关系类别.

1 问题研究和意义

1.1 基本定义

场景图生成可以看成两阶段的语义检索过程,首先使用目标检测算法生成节点,然后检测物体间的视觉关系得到边.

定义 1 给定一张图像 I,其对应的场景图用 S 表示,$B=\{b_1,\cdots,b_n\}\subseteq\mathbf{R}^4$ 是候选区域的集合,元素 b_i 表示第 i 个区域的边界框. $O=\{o_1,\cdots,o_n\}\subseteq\mathbf{N}$ 是对象的集合,元素 o_i 表示区域 b_i 对应的物体的类别标签. $R=\{r_{1\to2},r_{1\to3},\cdots,r_{n\to n-1}\}$ 是关系的集合,元素 $r_{i\to j}$ 对应一个视觉关系三元组 $t_{i\to j}=\{s_i,r_{i\to j},o_j\}$,其中 s_i 和 o_j 分别表示关系的主语和宾语. 场景图生成过程可以分解为 3 部分[17]:

$$p(S|I)=p(B|I)p(O|B,I)p(R|O,B,I). \tag{1}$$

式(1)中,边界框组件 $p(B_S|S)$ 生成一组候选区域,这些区域包含了输入图像中的大部分关键对象. 对象组件 $p(O_S|B_S,S)$ 预测每个区域中物体的类别. 这两部分可通过广泛使用的 Faster R-CNN 检测器来实现. 关系组件 $p(R_S|O_S,B_S,S)$ 推断每个物体对之间的关系.

1.2 语义连通图

场景图生成的核心是检测视觉关系. 检测一对物体间的关系有两种思路,一种是直接看作分类问题,即假设数据集中共有 K 种关系谓词,则有 $K+1$ 种分类结果,其中+1 表示两个物体间没有关系;另一种是先预测两个物体之间是否存在关系,若存在关系,则再进行 K 种关系的分类. 对于第一种方式,将任意两个对象关联为一个可能的关系会形成一个全连接图,通过删除一些在语义上弱依赖的对象之间的连接便形成稀疏的语义连通图. 语义连通图中的边表示两端的物体间的语义关联度更强,更易发生关系. 稀疏的语义连通图计算量小,精简了节点的邻接域,减少了干扰,使用图网络提取特征更合理高效.

2 基于语义连通图的场景图生成算法

2.1 网络整体结构

基于语义连通图的场景图生成算法的整体网络结构如图 1 所示,主要分为以下几个步骤:

(1)目标检测和特征提取:使用 Faster R-CNN 网络检测物体,并使用 Faster R-CNN 的输出构建物体节点的初始特征.

(2)关系提议:将任意两个对象关联为一个可能的关系以形成一个全连接图,然后使用一个关系建议模块删除不会发生关系的两个物体之间的连接,从而形成一个稀疏的语义连通图.

(3)特征增强:在语义图上使用图卷积网络模型对结点的特征进行更新和增强,使得结点融合上下文环境信息.

(4)联合区域视觉特征提取:对于有连接关系的物体对,将其联合边界框映射到特征图上,得到关系 RoI,使用 CNN 提取关系 RoI 对应的视觉短语特征.

(5)物体和关系预测:根据细化的对象特征重新预测物体类别,将视觉短语特征、主语和宾语特征进行哈达玛乘积运算得到关系的特征,使用分类器对关系进行分类.

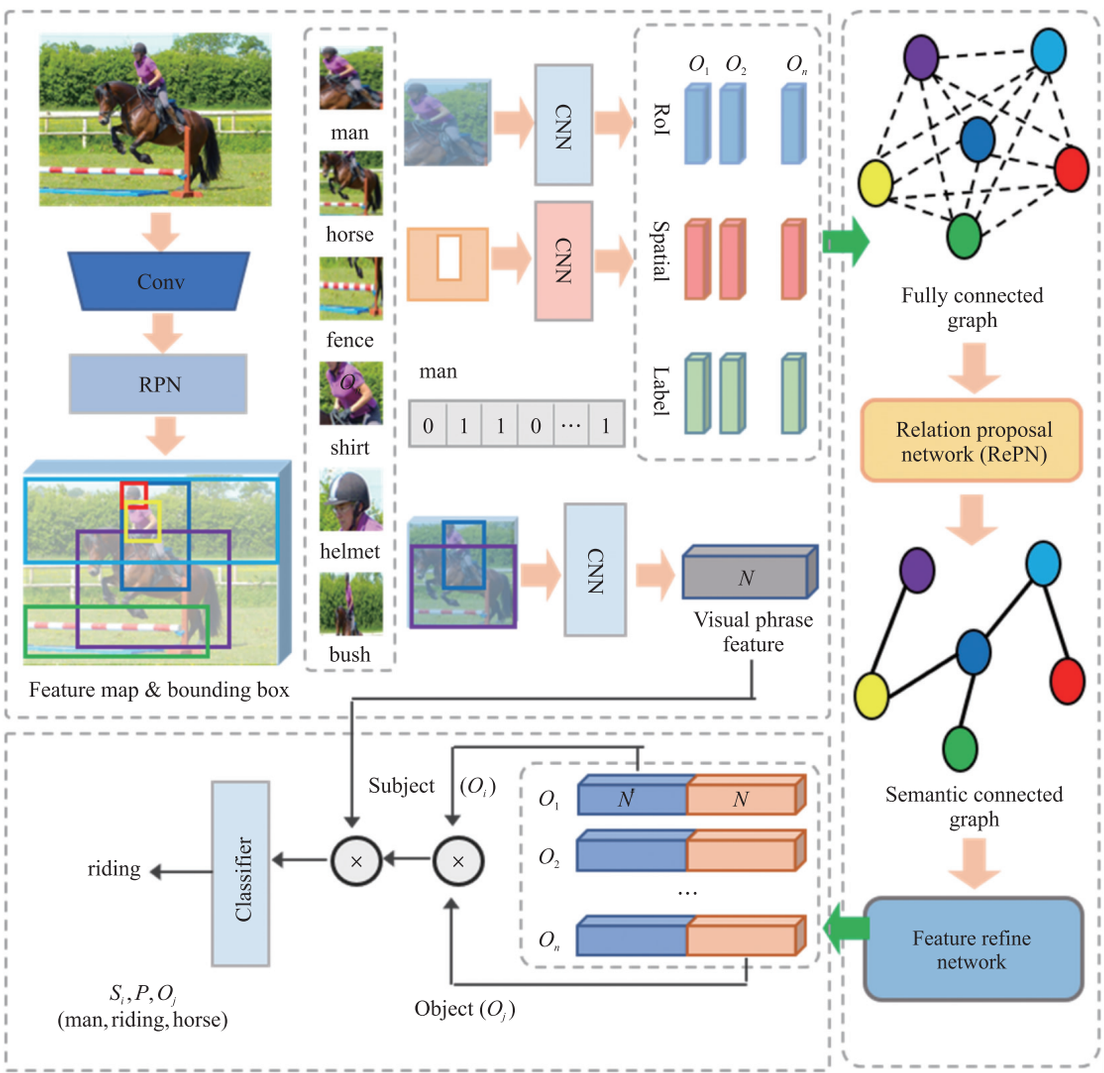

图 1　整体网络结构示意图

Fig. 1　Overview of the model

2.2　特征提取网络

本文使用 Faster R-CNN 网络作为基础的目标检测模型. Faster R-CNN 的输出包含物体的类别信息(类别概率向量)和物体边界框(bounding box,BBox)(代表物体在图像上的位置信息). 对于 Faster R-CNN 检测出的每个物体,可以得到 3 种初始信息:

(1)视觉特征:物体的视觉特征使用 Faster R-CNN 后端的两层全连接层输出的 4 096 维特征,$v_i \in \mathbf{R}^{4096}$.

(2)空间特征:物体的边界框由(x,y,w,h)4 个值表示,分别为中心点的横纵坐标及矩形框的长和宽,用正弦位置编码对位置坐标进行转化,然后使用一组可训练的参数以学习的方式转化相对位置特征,将

4 维向量表示的位置特征映射到 128 维,即 $\boldsymbol{s}_i \in \mathbf{R}^{128}$.

(3)语义特征:语义特征是物体类别标签的词嵌入(word embedding),采用预训练的 Word2vec 将类别标签转化为对应词向量 $\boldsymbol{l}_i \in \mathbf{R}^{200}$.

2.3 关系建议网络

在所有 $O(n^2)$ 个物体对中,只有一小部分物体对可能具有关系. 大量的连接会使训练和推理难以进行,且冗余的关系建议会降低召回性能. 因此需要构建一个语义指导的关系建议网络,以删除全连接图中语义依赖性弱的物体之间的连接,从而构建高效的语义连通图. 本文使用预训练的 Word2vec 模型学习单词嵌入,从而得到对象之间的语义依赖. 词嵌入矩阵为 $\boldsymbol{W}_e \in \mathbf{R}^{C\times m}$,每一行为一个物体类别标签的词嵌入向量,则物体 i 的语义嵌入可表示为:

$$\boldsymbol{e}_i = \boldsymbol{s}_i \cdot \boldsymbol{W}_e, \tag{2}$$

式中,$\boldsymbol{s}_i \in \mathbf{R}^C$ 为物体 i 的预测类别分布向量,是 Faster R-CNN 后端用于分类的全连接层的输出,C 为数据集中物体类别数量. 式(2)表示的词嵌入是一种软嵌入,考虑了 Faster R-CNN 模型给出的对象类别预测的不确定性,能够减轻对象分类误差带来的负面影响.

对于一个对象对 (o_i, o_j),将其关系表示为主语嵌入、空间信息和宾语嵌入的连接:$r_{i,j} = [e_i, \hat{b}_i, \hat{b}_j, e_j] \in \mathbf{R}^{2m+8}$. 其中,$\hat{b}_i = (\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i)$ 为物体 o_i 的边界框 $b_i = (x_i, y_i, w_i, h_i)$ 的参数相对于联合边界框 $b_{i,j} = (x, y, w, h)$ 计算得到的相对位置特征,计算方式为:

$$\begin{aligned} \hat{x}_i &= (x_i - x)/w, \\ \hat{y}_i &= (y_i - y)/h, \\ \hat{w}_i &= \log(w_i/w), \\ \hat{h}_i &= \log(h_i/h). \end{aligned} \tag{3}$$

然后将关系表示 $r_{i,j}$ 送入多层感知机(multilayer perceptron,MLP). MLP 的输出由 Sigmoid 函数正则化到[0,1]区间,即得到语义依赖分数 $SC_{i,j}$,该分数表示对象对 (o_i, o_j) 之间形成一个有意义的关系的可能性. 选择语义依赖得分在 top-K 且大于阈值的对象对. 由于非极大值抑制(non-maximum suppression,NMS)[18]会降低关系提议的召回率,本文使用 top-K 来限制关系建议的最大数量,以提高训练的有效性,并使用分数阈值来减少冗余.

2.4 特征更新网络

在语义连通图上使用图卷积网络 GCNConv[14] 和 GAT[15] 对物体的特征进行融合和更新. 假设节点 i 的初始特征向量为 $\boldsymbol{x}_i$,每一次更新用时间步长 t 表示,则在 t 时刻,每个节点有一个隐状态 h_i^t. 使用特征向量 $\boldsymbol{x}_i$ 初始化时刻 $t=0$ 时的隐状态 $h_i^0 = \phi_0(\boldsymbol{x}_i)$,其中 ϕ_0 是一种将 $\boldsymbol{x}_i$ 映射到低维向量空间的变换,由全连接层实现. 在每个时间步 t,每个节点根据图结构聚合来自其邻居的消息,得到更新后的节点表示.

(1)GCNConv:GCNConv 从图谱理论角度定义图结构上的卷积操作,其信息聚合的方式为:

$$h_i^t = W \sum_{j \in N(i) \cup \{i\}} \frac{1}{\sqrt{\hat{d}_j \hat{d}_i}} h_j^{t-1}, \tag{4}$$

$$\hat{d}_i = 1 + \sum_{j \in N(i)} e_{j,i}, \tag{5}$$

式中,$e_{j,i}$ 表示源节点 i 到目标节点 j 之间的边的权重.

(2)GAT:GAT 将自注意力机制引入图网络,注意力机制可看作为在将一个节点的邻居的特征聚合到这个节点时为每个邻居节点分配的权重,其节点特征的更新方式为:

$$h_i^t = \alpha_{i,i} W h_i^{t-1} + \sum_{j \in N(i)} \alpha_{i,j} W h_j^{t-1}. \tag{6}$$

其中,注意力系数 α_{ij} 由下式计算:

$$\alpha_{ij} = \frac{\exp(\mathrm{LeakyReLU}(\boldsymbol{\Lambda}^{\mathrm{T}}[Wh_i^{t-1}, Wh_j^{t-1}]))}{\sum_{k \in N(i) \cup \{i\}} \exp(\mathrm{LeakyReLU}(\boldsymbol{\Lambda}^{\mathrm{T}}[Wh_k^{t-1}, Wh_j^{t-1}]))}, \tag{7}$$

式中,[.]表示拼接操作;$\boldsymbol{\Lambda}$ 为向量变换,由全连接层实现.

如图 2 所示,特征更新网络共有 3 个分支,主分支为两层 GCNConv 网络,节点的输入为语义特征和视觉特征的拼接. 同时,由于希望图网络可以学习到物体之间的语义相关性,设计了单独的分支将语义特征送入一层 GAT 网络进行更新. 对于某一物体来说,与其相邻的物体与之发生关系的概率是不一样的,因而对其影响也不相同,GAT 刚好可以学习到不同邻居之间的权重. 此外,再增加一个分支单独融合视觉特征,和主分支同步进行,两个分支的输出向量求哈达玛积得到融合后的特征. 最终特征更新网络模块输出的节点特征为更新后的语义特征 $\tilde{\boldsymbol{l}}_i \in \mathbf{R}^{128}$、融合后的节点特征 $h_i \in \mathbf{R}^{128}$和初始的位置特征 $\boldsymbol{l}_i \in \mathbf{R}^{128}$的拼接.

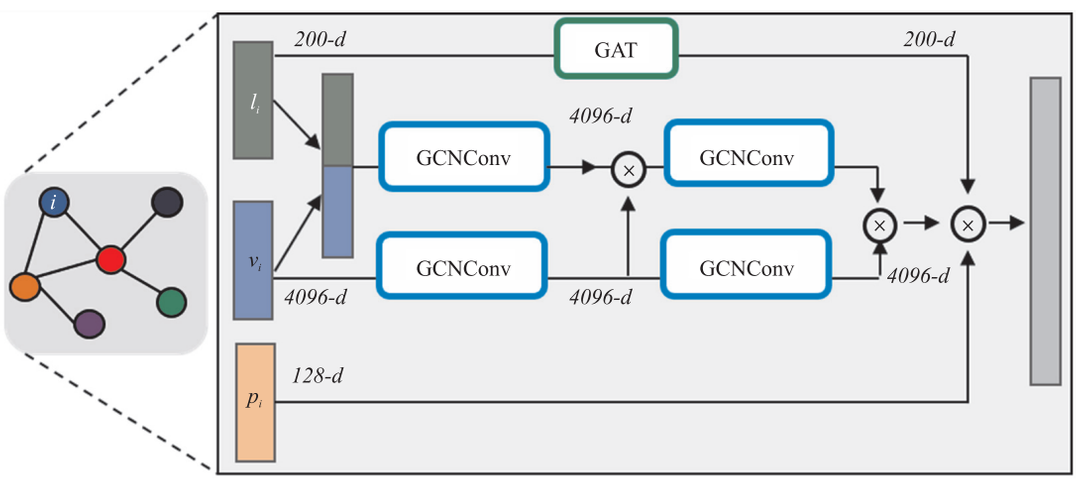

图 2　特征更新网络结构图

Fig. 2　Illustration of the feature refining module

2.5　关系推理网络

由于视觉关系是有方向性的,对于物体对(o_i, o_j)来说,o_i 作主语和 o_j 作主语,形成的关系往往是不一样的,而在数据集中,往往只标注一种主-客体组合下的关系. 因此,为了提高关系的召回率,需要预测两种组合下的关系. 特征更新网络输出的物体节点特征为$\tilde{f}_i^o \in \mathbf{R}^{4424}$,通过全连接层将其映射到 8 192 维的特征向量,前 4 096 维是物体 o_i 作为主语时的特征,后 4 096 维是 o_i 作为宾语时的特征. 对于主-客体对(o_i, o_j),o_i 的主语特征和 o_j 的宾语特征求哈达玛积得到$\tilde{f}_{ij}^r \in \mathbf{R}^{4096}$,$\tilde{f}_{ij}^r$再与视觉短语特征 v_{ij}求哈达玛积得到关系推理特征 x_{ij}. 关系推理模块如图 3 所示. 假设数据集中的物体类别数为 N,关系类别数为 K,则物体特征经过全连接层得到物体类别分数 $SC_i^o \in \mathbf{R}^{N+1}$,关系特征经过全连接层得到关系类别分数 $SC_{i,j}^r \in \mathbf{R}^{K+1}$.

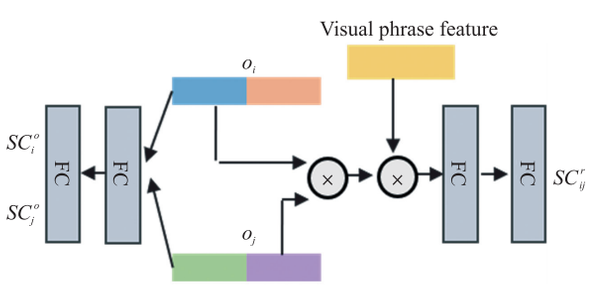

图 3　关系推理网络结构图

Fig. 3　Illustration of the relation reasoning module

3　实验与结果

3.1　数据集

Visual Genome[19]数据集是目前拥有最多的物体和视觉关系的开放数据集. 由于数据集中的注释是由众包工人完成的,很大一部分对象注释的质量很差,且有重叠的边界框和含糊不清的对象名称. 为了消除这些干扰,很多研究人员已经探索了多种半自动的方法(如类合并和过滤)来清除对象和关系注释,构建出很多过滤后的 VG 版本. 本文的实验使用 VG150[20] 和 VG-MSDN[21] 两个数据集. VG150 对每张图像修正了 22 个边界框或对象名称,删除了 7.4 个边界框,合并了 5.4 个重复的边界框. 基准测试方法使用最常用的 150 个对象类别和 50 个谓词关系类别进行实验评估. 因此,每张图像对应一个大约由 11.5 个对象和 6.2 个关系构成的场景图 VG150 保留了 VG 数据集全部的 108 077 张图像. VG-MSDN 对不同时态的词进行规范化,同样选取了 150 个最常见的物体种类和 50 个关系类别,同时删除边界框的短边小于 16 像素的物体的标注信息,经处理后剩余 95 998 张图像.

3.2 评价指标

本文评价指标为图像级的召回率 Recall@ K(R@ K)[22],用以计算预测的关系三元组中置信度最高的前 K 个中包含的真实关系组合的比例. 实验中 K 分别取 50 和 100. 具体计算方式为:给定 N 张图像,对于每张图像首先对所有预测的视觉关系按照得分进行排序,然后取得分为前 K 的预测. 视觉关系的得分由包括主语对象的分类得分、宾语对象的分类得分和视觉关系谓词的分类得分相加而得. 对于图像 i,若其包含 $|GT_i|$ 组真实的关系标注,而模型正确预测到的关系组合为 $TP_i = Top_{K_i} \cap GT_i$,则召回率为:

$$\mathrm{R@}K = \frac{1}{N}\sum_{i=1}^{N}\frac{|TP_i|}{|GT_i|}. \tag{8}$$

3.3 任务设置

给定一幅图像,场景图生成任务包括对一组对象进行定位,对其类别标签进行分类,以及预测这些对象之间的关系. 对场景图生成模型的性能评估通常有多种不同的任务设置,本文将在以下 3 种任务中测试所提出的模型:

(1)谓词分类任务(predicate classification,PredCls):给定图像中物体的类别标签和边界框信息,检测物体对之间是否存在关系,并对关系进行分类;

(2)场景图分类任务(scene graph classification,SGCls):给定图像中的一组物体的边界框,预测物体类别,并检测物体之间的关系;

(3)场景图检测任务(scene graph detection,SGDet):给定一张图像,需检测出图像中的物体,并检测物体间的关系.

3.4 实验和结果

本文在 VG150 和 VG-MSDN 数据集上对所提出的场景图生成算法性能进行测试,并与现有算法进行比较,结果如表 1 所示.

表 1 各种模型的结果对比

Table 1 The results of various models

Dataset	Models	PredCls		SGCls		SGDet	
		R@ 50	R@ 100	R@ 50	R@ 100	R@ 50	R@ 100
VG150	IMP[20]	44.75	53.08	21.72	24.38	3.44	4.24
	Graph R-CNN[23]	54.20	59.10	29.60	31.60	11.40	13.70
	ours	64.01	65.78	35.32	36.04	21.89	26.62
VG-MSDN	MSDN[21]	67.03	71.01	24.34	26.50	10.72	14.22
	AVR[24]	64.97	64.97	29.12	32.29	18.33	19.97
	ours	69.03	70.21	27.33	32.37	17.37	17.98

从表 1 可以看出,本文提出的基于语义连通图的场景图生成算法在两个数据集上都取得了较好的结果. 由于 VG150 数据集中选取的是出现频率最高的 150 种物体和 50 种关系,对于 SGCls 和 SGDet 两个任务,同样的算法在 VG150 数据集上的结果更好一些. 在 VG150 数据集上,本文的算法在 3 个任务上都比其他算法在召回率上有大幅提升,提升约 5% ~ 10%,而在 VG-MSDN 数据集上与其他的算法效果不相上下. 在 PredCls 任务上,R@ 50 提升明显,而其他指标稍低,推测在关系建议模块中删除的连接较多,可能误删了一些有效关系.

为了更直观地显示算法生成场景图的结果,在 VG150 数据集的测试集中随机选取了 6 张图像进行测试,并将结果可视化表示,如图 4 所示. 上方图像中标记了物体的类别和 Bounding Box,下方是图像对应的场景图,其中绿色标记是算法输出的正确检测结果,红色标记是算法未检测到或是检测错误的物体和关系,检测错误的物体和关系在括号中给出了 ground truth 标签. 可以看出,大多数的物体和关系均可检测到,但物体分类错误的情况仍较多. 从第四张和第五张图像的测试结果可以看出,虽然物体分类错误,但关系仍可检测出来,说明模型在一定程度上还是可以学习到独立于物体的关系特征表示. 同时,检测错误的物体确实存在一定的干扰,如第五张图像中的两个检测错误的“person”,都有雨伞的遮挡.

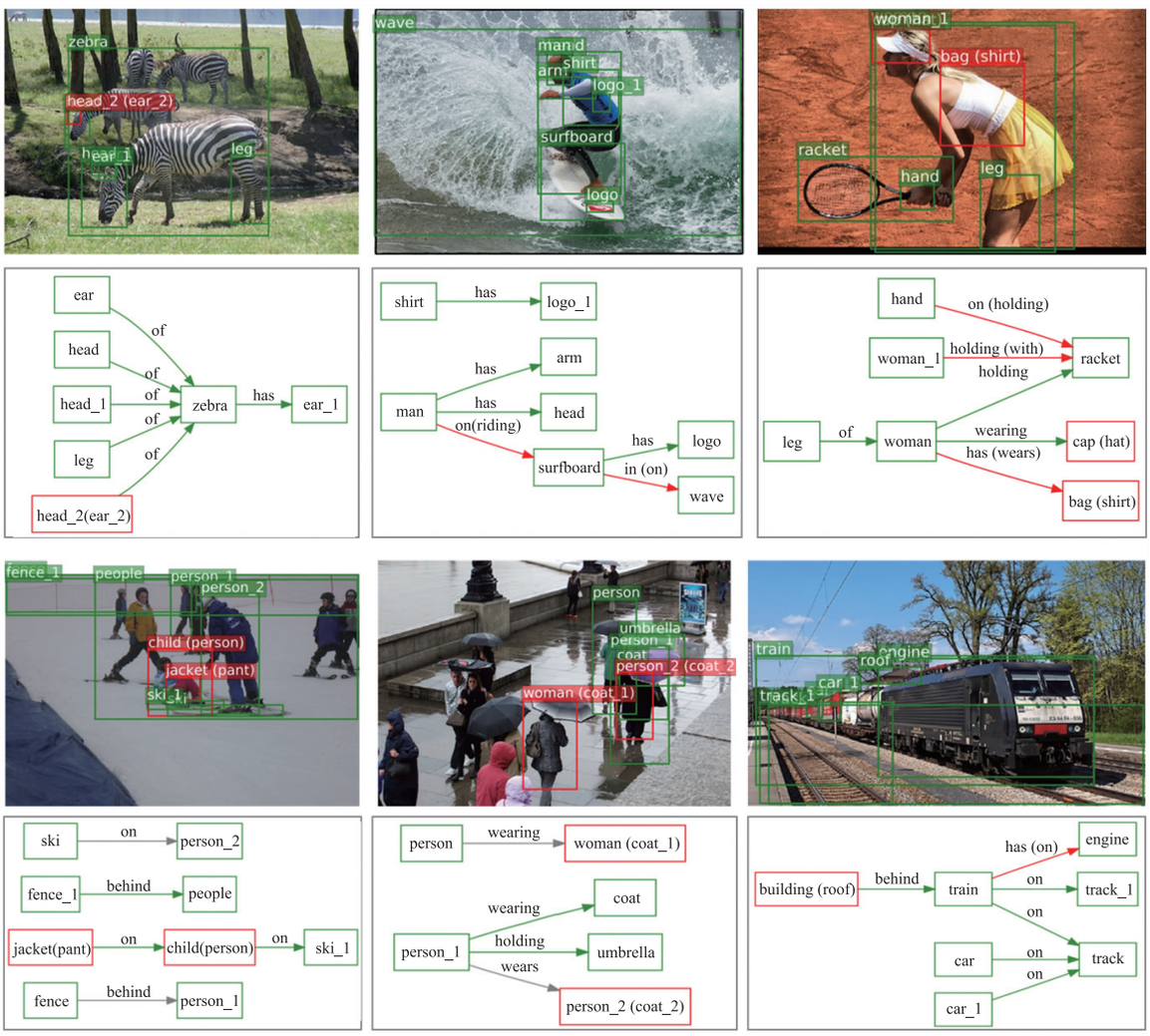

图 4　VG150 数据集上的测试结果示例

Fig. 4　Some visualization test samples of ours model on VG150 dataset

4　结论

本文提出了基于语义连通图的场景图生成算法,借助物体的语义信息预测物体之间是否存在关系,同时使用图卷积神经网络融合物体节点的特征,并使用融合了上下文信息的物体特征组合构建关系特征,对关系类别进行推理. 经实验验证,本文算法在 VG150 和 VG-MSDN 两个数据集上都取得了有竞争力的结果.

[参考文献](References)

[1] JOHNSON J,KRISHNA R,STARK M,et al. Image retrieval using scene graphs[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston,USA:IEEE,2015:3668-3678.

[2] 田鑫,季怡,高海燕,等. 外部信息引导和残差置乱的场景图生成方法[J]. 计算机科学与探索,2021,15(10):1958-1968.

[3] 黄勇韬,严华. 结合注意力机制与特征融合的场景图生成模型[J]. 计算机科学,2020,47(6):133-137.

[4] 庄志刚,许青林. 一种结合多尺度特征图和环型关系推理的场景图生成模型[J]. 计算机科学,2020,47(4):136-141.

[5] LI Y K,OUYANG W L,ZHOU B L,et al. Factorizable net:an efficient subgraph-based framework for scene graph generation[C]//Proceedings of the 2018 European Conference on Computer Vision(ECCV). Munich,Germany:Springer,2018:335-351.

[6] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus,USA:IEEE,

2014:580-587.
[7] GIRSHICK R. Fast R-CNN[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago,Chile:IEEE,2015:1440-1448.
[8] REN S Q,HE K M,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[J]. IEEE Transactions of Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.
[9] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:Unified,real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,USA:IEEE,2016:779-788.
[10] REDMON J,FARHADI A. Yolov3:An incremental improvement[J]. arXiv Preprint arXiv:1804.02767,2018.
[11] LIU W,ANGUELOV D,ERHAN D,et al. SSD:Single shot multibox detector[C]//Proceedings of the 2016 European Conference on Computer Vision. Amsterdam,The Netherlands:Springer,2016:21-37.
[12] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,32(2):318-327.
[13] WU Z H,PAN S R,CHEN F W,et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,32(1):4-24.
[14] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv Preprint arXiv:1609.02907,2016.
[15] VELIČKOVIĆ P,CUCURULL G,CASANOVA A,et al. Graph attention networks[J]. arXiv Preprint arXiv:1710.10903,2017.
[16] HAMILTON W L,YING R,LESKOVEC J. Inductive representation learning on large graphs[J]. arXiv Preprint arXiv:1706.02216,2017.
[17] CHEN T S,YU W H,CHEN R Q,et al. Knowledge-embedded routing network for scene graph generation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019:6163-6171.
[18] NEUBECK A,VAN GOOL L. Efficient non-maximum suppression[C]//Proceedings of the 18th International Conference on Pattern Recognition(ICPR 2006). Hong Kong,China:IEEE,2006:850-855.
[19] KRISHNA R,ZHU Y K,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.
[20] XU D F,ZHU Y K,CHOY C B,et al. Scene graph generation by iterative message passing[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA:IEEE,2017:3097-3106.
[21] LI Y K,OUYANG W L,ZHOU B L,et al. Scene graph generation from objects,phrases and region captions[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice,Italy:IEEE,2017:1270-1279.
[22] LU C W,KRISHNA R,BERNSTEIN M,et al. Visual relationship detection with language priors[C]//Proceedings of the 2016 European Conference on Computer Vision. Amsterdam,The Netherlands:Springer,2016:852-869.
[23] YANG J W,LU J S,LEE S,et al. Graph R-CNN for scene graph generation[C]//Proceedings of the 2018 European Conference on Computer Vision(ECCV). Munich,Germany:Springer,2018:670-685.
[24] LÜ J M,XIAO Q Z,ZHONG J J. AVR:Attention based salient visual relationship detection[J]. arXiv Preprint arXiv:2003.07012,2020.

[责任编辑:严海琳]