计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190300002

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

庄志刚　许青林
广东工业大学计算机学院　广州 510000
(ZZG2016@yeah.net)

摘　要　场景图为描述图像内容的结构图(Graph),其在生成过程中存在两个问题:1)二步式场景图生成方法造成有益信息流失,使得任务难度提高;2)视觉关系长尾分布使得模型发生过拟合、关系推理错误率上升。针对这两个问题,文中提出结合多尺度特征图和环型关系推理的场景图生成模型 SGiF(Scene Graph in Features)。首先,计算多尺度特征图上的每一特征点存在视觉关系的可能性,并将存在可能性高的特征点特征提取出来;然后,从被提取出的特征中解码得到主宾组合,根据解码结果的类别差异,对结果进行去重,以此得到场景图结构;最后,根据场景图结构检测包含目标关系边在内的环路,将环路上的其他边作为计算调整因子的输入,以该因子调整原关系推理结果,并最终完成场景图的生成。实验设置 SGGen 和 PredCls 作为验证项,在大型场景图生成数据集 VG(Visual Genome)子集上的实验结果表明,通过使用多尺度特征图,相比二步式基线,SGiF 的视觉关系检测命中率提升了 7.1%,且通过使用环型关系推理,相比非环型关系推理基线,SGiF 的关系推理命中率提升了 2.18%,从而证明了 SGiF 的有效性。

关键词:场景图生成;多尺度特征图;环型关系推理;卷积神经网络;图像理解
中图法分类号　TP389.1

Scene Graph Generation Model Combining Multi-scale Feature Map and Ring-type Relationship Reasoning

ZHUANG Zhi-gang and XU Qing-lin
School of Computer,Guangdong University of Technology,Guangzhou 510000,China

Abstract　The scene graph is a graph describing image content. There are two problems in its generation:one is the loss of useful information caused by two-step scene graph generation method,which promotes the difficulty of this working,and the second is the model overfitting due to the long-tail distribution of visual relationship,which increases the error rate of relationship reasoning. To solve these two problems,a scene graph generation model SGiF (Scene Graph in Features) based on multi-scale feature map and ring-type relationship reasoning was proposed. Firstly,the possibility of visual relationship is calculated for each feature point on the multi-scale feature map and the features with high possibility are extracted. Then,the subject-object combination is decoded from extracted features. According to the difference of the decoding result category,the result will be deduplicated and the scene graph structure will be obtained. Finally,the ring including the targeted relationship edge is detected according to the graph structure,then the other edges of this ring are used as input of the calculation about factor to adjust the original relationship reasoning result,at last,the scene graph generation work is completed. In this paper,SGGen and PredCls were used as verification items. The experimental results on the subset of large dataset VG (Visual Genome) used for scene graph generation show that,by using multi-scale feature map,SGiF improves the hit rate of visual relationship detection by 7.1% compared with the two-step baseline,and by using the ring-type relationship reasoning,SGiF improves the accuracy of relational reasoning by 2.18% compared with the baseline with non-ring relational reasoning,thus proving the effectiveness of SGiF.

Keywords　Scene graph generation,Multi-scale feature map,Ring-type relationship reasoning,Convolution neural networks,Image understanding

1　引言

得益于深度学习的发展,计算机理解图像的能力得到了极大的提高,如图像分类(Image Classification)[1-3]、目标检测(Object Detection)[4-12]以及实例分割(Instance Segmentation)[13-14]等基础计算机视觉任务的精确度已超越以往的表

到稿日期:2019-03-04　返修日期:2019-06-28　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:广东省科技计划项目(2016B030306003)
This work was supported by the Science and Technology Planning Project of Guangdong Province,China (2016B030306003).
通信作者:许青林(gzj2ee@126.com)

现。然而，对于视觉问题回答（Visual Question Answering，VQA）[15-16]、图像标注（Image Annotaion）[17-18]以及基于内容的图像检索（Content-based ImageRetrieval，CBIR）[19-20]等图像转换文本（Image2Text）任务而言，其不仅仅需要检测、分类图像中的物体，还需要理解物体间的关系，此时便需要场景图（Scene Graph，SG）的辅助。场景图为结构化的图像内容描述方式，是一种以图像物体为顶点、物体间关系为连接边的图（Graph），具有与应用场景无关且方便计算机使用的优点。目前，常见的场景图生成方法被统称为基于目标检测[4-12]的二步式场景图生成方法，其由文献[21]提出，具体流程为：首先使用目标检测模型检测顶点（物体），然后根据成对顶点组合推理连接关系边。这种基于目标检测的二步式场景图生成方法由于具有过程简明、模块耦合度低的优点，成为了主流场景图生成研究工作[22-25]的选择。

然而，文献[26]认为，功能内聚的目标检测模型专注于物体检测，忽略了物体间交互动作以及状态等有益于关系推理的信息收集，导致有益信息流失，增大了场景图的生成难度。

除此之外，在场景图生成过程中，视觉关系长尾分布导致关系推理模型发生过拟合。视觉关系长尾分布是指同属一对顶点的大量训练样本集中于少数几类关系上，而其他大多数关系仅占有少部分训练样本的情形。对于依赖数据的关系推理模型而言，这将会导致模型过拟合于少数关系类别，致使关系推理错误率上升。

面对上述两个问题，本文受文献[26]的启发，提出了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——SGiF，其包含两部分重要工作：

1）使用多尺度特征图取代目标检测模型以检测场景图构成元素。这里的场景图构成元素为文献[21]中提到的视觉关系，是形为（主语，谓语，宾语）的三元组，可作为场景图的子结构。通过使用卷积神经网络中的多尺度特征图，SGiF 不仅能达到避免有益信息流失的目的，而且可以在一定程度上避免文献[26]所提到的场景图构成元素重叠问题。

2）使用环型关系推理方法，即通过使用场景图中的环路背景信息来辅助基于顶点对的关系推理过程。参考使用图中环型结构推理未知关系边的方法[27-28]，SGiF 检测由目标关系边组成的环路结构，并将其作为关系推理过程所使用的背景信息。SGiF 一方面可以提高基于顶点对的关系推理正确率，另一方面可以提高数据样本的丰富度，一定程度上缓解了数据分布不均的问题。

2 相关工作

2.1 特征图应用

随着深度学习的发展，具有自动学习能力且鲁棒性强的卷积神经网络（Convolution Neural Networks，CNNs）成为现代主流的图像特征提取方法。通常 CNNs 的前向计算过程会产生多个特征图，这些特征图遵循这样的规律：低层特征图分辨率低、视觉细节较多，而高层特征图分辨率高、语义信息较多。按照使用特征图数量的不同，将基于 CNNs 的应用分作为单层特征图应用与多层特征图应用。以目标检测为例，RCNN 系列[10-12]，YOLOV1[4]以及 YOLOV2[5]等单层特征图应用仅使用 CNNs 中的一层特征图，这种做法的优点是足够简单，计算量少，但缺点是单层特征图信息量有限，限制了模型能力，具体体现为这些目标检测模型大多不擅长处理图像中存在小型目标或者目标密集的情况。为弥补这一缺点，YOLOV3[6]，SSD[7]，FPN[8]以及 RetinaNet[9]等多层特征图目标检测应用使用 CNNs 中的多层特征图，这种做法一方面通过使用具有较多视觉细节的底层特征图来提高小型目标的检测精度，另一方面通过增加预测结果数量，形成以多胜少的优势，以此来提高模型的精确度。

同为多层特征图的应用，SGiF 更强调多种不同尺度特征图的应用。除保留有益信息的根本目的外，使用多尺度特征图的另一个原因在于使用相同尺度特征图的做法难以保证检测目标不出现重叠问题，为此 SGiF 中必须使用多尺度特征图来降低重叠现象的发生概率。

2.2 场景图生成中的关系推理

文献[21]的研究工作大大降低了场景图的生成难度，并促使场景图生成的研究工作将注意力转向更高精度的关系推理研究上。一部分场景图生成工作集中于视觉关系，如文献[22]受到 TransE[29]思想的启发认为：顶点间关系约等于两顶点特征的偏移量，并由此推出 VTransE 模型，该模型的核心内容为 $W_s x_s + t_p \approx W_o x_o$，其中参数 W_s 和 W_o 为将顶点视觉特征 x_s 和 x_o 映射至关系空间的权重，t_p 为关系的语义特征。此外，文献[23]在关注顶点空间关系及其闭合区域视觉特征对生成任务的影响之余，以条件随机场（Conditional Random Field，CRF）的形式解释视觉关系中主谓宾三者的依赖关系，并推出了 DR-Net 模型来完成视觉关系推理工作。另一部分场景图生成研究工作利用图结构辅助工作完成，如文献[24]认为图（Graph）中的信息具有在顶点和连接边之间相互传播的特性。以顶点为例，顶点更新过程中接受其连接边内容作为输入，并按照权重选择性地更新原本的内容，该过程被称作 Message Pool，而连接边的更新过程与顶点类似，不同点在于所接收的输入为连接边两端的顶点，在整个场景图生成过程中，顶点与连接边将迭代更新若干次，直至所有顶点与连接边的状态稳定，以此完成场景图生成工作。相对文献[24]强调图结构的利用，文献[25]更强调贴近数据集，其通过实验证明数据集 VG[30]中的场景图存在一种共同的子结构 Motif，并表明令模型学会记忆子结构 Motif 将有助于场景图的生成。其具体做法是使用一种以 LSTM（Long Short Time Memory）为基础的编码解码网络（将其命名为 MotifNet）来记忆子结构 Motif 并提高场景图的生成精确度。

不同于上述针对视觉关系的场景图研究工作，为提高关系推理的精确率，SGiF 在原有的顶点对关系推理的基础上增加环路背景信息调整功能，具体是使用环路结构计算调整因子，并以此因子调整原关系推理结果，从而提升关系推理的精确度。

3 SGiF

SGiF 的工作过程可分为多尺度特征图检测场景图元素以及环型关系推理两个阶段。第一个阶段如图 1 所示，SGiF 接收图像作为输入，经过特征提取、主宾组合推理以及元素去重 3 个步骤，得到场景图结构，此时需要注意连接边尚未进行

推理工作，因此，第二阶段中 SGiF 根据所得的场景图结构，检测由所有待推理连接边所组成的环路，将环路作为关系推理输入，以此求得所有连接边并最终完成场景图的生成，如图 2 所示。

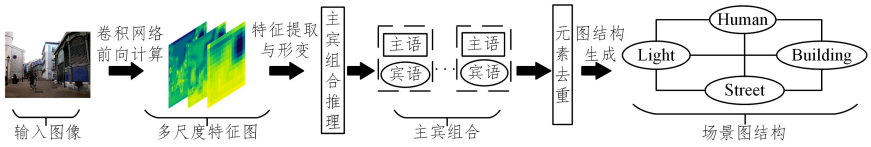

图 1 多尺度特征图检测场景图元素流程

Fig. 1 Flow of multi-scale feature map detecting scene graph element

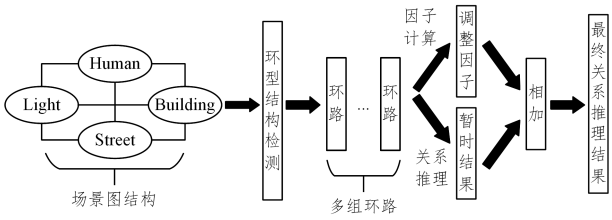

图 2 环型关系推理流程

Fig. 2 Flow of ring-type relationship reasonling

3.1 多尺度特征图检测场景图元素

从基于目标检测的二步式场景图生成方法中可了解到：使用目标检测模型所得到的信息缺乏对物体间状态的描述，导致场景图生成难度上升。针对该问题，SGiF 借鉴文献[7-9]的思想，提出使用多尺度特征图来检测视觉关系，其在最大化保留图像细节信息的同时，能降低元素发生重叠的可能性，其具体流程如下：

将图像 $I\in R^{W*H*C}$ 作为特征提取网络（即卷积神经网络）的输入，得到若干特征图 $F=\{f_i\mid f_i\in R^{w_i'*h_i'*c_i'},i\in(0,K]\}$。考虑到在大多情况下，图像中的物体空间分布总是稀疏的，为保证效率，需要为特征图上的每个特征点计算视觉关系存在的可能性。通过使用一个由一层卷积操作层和一层 Sigmoid 激活层组成的评分网络，将特征图转变成评分图 $S=\{s_i\mid s_i\in R^{w_i'*h_i'*1},i\in(0,K]\}$。设置视觉关系存在可能性的阈值为 $\vartheta\in(0,1)$，将评分大于或等于 ϑ 的特征从特征图中提取出来，并通过线性映射操作来统一特征维度，此时可得一组输入特征 $X=\{x_i\mid x_i\in R^{c''},i\in(0,N]\}$。由于输入特征中包含视觉关系信息，因此可使用主宾推理模型从输入特征中解码出多对主宾组合 $SO=\{(s_i,o_i)\mid i\in(0,N]\}$，需要注意到，一个特征 x_i 仅对应一对 (s_i,o_i)，且 s_i 与 o_i 均为顶点，其内包含顶点类别与坐标信息，对于同一对 (s_i,o_i) 而言，其需要满足条件 $s_i\neq o_i$。最后，根据所得的主宾组合 SO，生成相对应的场景图结构。图结构的生成过程为：首先进行顶点去重和聚合，对于同一类别顶点采用非极大值抑制法（Non-Maximum Suppression，NMS），以聚合相同顶点并分离相异顶点；然后进行主宾组合去重，前项工作可能导致多对主宾组合之间发生重合，此时优先选择重合部分中视觉关系存在可能性与主宾两顶点分类置信度的乘积最大的一项，以此保证去重的效果，并最终完成场景图结构的生成。

3.2 环型关系推理

视觉关系长尾分布促使基于顶点对的关系推理模型 $f(x_a,x_b)$ 过拟合于少数关系类别，使得关系推理错误率大大提升。SGiF 解决该问题的措施是引入环路作为背景信息以辅助关系推理过程。修改关系推理模型为 $f(x_a,x_b,B_{x_ax_b})$，其中 $B_{x_ax_b}$ 是包含 x_a x_b 边在内的 K 元环路，为降低复杂度，可令 $K=3$。此时令关系推理输入 (x_a,x_b) 为一对主宾组合 (s_i,o_i)，考虑到 (s_i,o_i) 是从输入特征 x_i 中解码得到的，因此为减少信息的损耗，可将关系推理模型改作 $f(x_{so},x_{ok},x_{ks})$，其中的 x_{so},x_{ok},x_{ks} 属于输入特征 X，且 x_{so} 代指目标关系边，其余两项代指三元环路上的另外两边。模型 $f(x_{so},x_{ok},x_{ks})$ 的具体工作过程如下。

首先基于顶点对的关系推理，将输入 x_{so} 线性映射至关系空间，过程如式(1)所示：

$$y_{so}=W_{so}\cdot x_{so}+b_{so} \tag{1}$$

其中，W_{so} 和 b_{so} 为基于顶点对的关系推理过程所用的权重与偏置，需要注意的是，y_{so} 为暂时性关系推理结果，整个环型关系推理工作尚未完成。

然后对另外两项输入特征 x_{ok}，x_{sk} 进行连接操作（Concatenate），得到一个新特征，并以此计算调整因子，过程如式(2)所示：

$$y_{adj}=W_{adj}\cdot[x_{ok};x_{sk}]+b_{adj} \tag{2}$$

其中，W_{adj} 和 b_{adj} 为调整因子计算过程所用的权重与偏置，y_{adj} 为调整因子，其与 y_{so} 具有相同维度。令 y_{adj} 和 y_{so} 相加，可使得 y_{so} 概率分布发生改变，以起到调整的作用。

最后考虑到环型关系推理模型的鲁棒性问题，在与暂时结果 y_{so} 相加之前，需要先对调整因子 y_{adj} 进行 Dropout 处理[31]，过程如式(3)所示：

$$y_{relationship}=y_{so}+dropout(y_{adj}) \tag{3}$$

其中，$y_{relationship}$ 为最终关系推理结果，直到此时才算完成整个环型关系推理工作。

需要说明的是，在进行环型关系推理之前，SGiF 需要从场景图结构中检测环路结构，此时对目标关系边的环路检测会出现两种极端情况：1）图中存在多条由目标关系边组成的环路结构；2）图中不存在由目标关系边组成的环路结构。前者的解决方法较为简单，根据当前模型训练迭代次数与环路条数相除所得余数来决定所用环路，而后者则需要以零向量代替输入特征 x_{ok}，x_{sk}，此时计算得出调整因子为零，环型关系推理将退化为基于顶点对的关系推理。

3.3 SGiF 模型训练

为确保环型关系推理不退化为基于顶点对的关系推理，环型结构检测部分至少能根据目标关系边找到一条环路，这意味着在进行环型关系推理之前，SGiF 首先需要生成符合要求的场景图结构。因此，需要为 SGiF 设计更独特的训练方案。

先设定 SGiF 模型损失值为 $loss_{all}$，其包含如式(4)所示的 3 部分内容。

$$loss_{all}=\alpha\cdot loss_{ex}+\beta\cdot loss_{so}+\gamma\cdot loss_{pred} \tag{4}$$

其中，$loss_{ex}$，$loss_{so}$，$loss_{pred}$ 分别为视觉关系存在评分部分、主宾组合部分以及关系推理部分的损失值，而 α,β,γ 为这 3 部分损失值的控制系数。

视觉关系存在评分部分的损失值 $loss_{ex}$ 的计算如式(5)所示：

$$loss_{ex}=-\frac{1}{F}\sum_{i=1}^{F}\frac{1}{N_i}\sum_{j=1}^{N_i}t_j\log(s_j) \tag{5}$$

其中，$s_j\in(0,1)$为评分图 S 上的某一特征点评分，$t_j\in\{0,1\}$ 为该特征点上是否存在视觉关系的指标，当该特征点所指代的原图像区域与真值的 IoU(Intersection over Union)大于设定阈值时，该点的 $t_j=1$，反之则为 0。F 和 N_i 分别为评分图总数和该评分图特征点总数。

主宾组合部分的损失值主要计算顶点分类以及坐标回归的误差。某一对 (s_i,o_i) 主宾组合损失值 $loss_{so}^{i}$ 的计算式如式(6)所示：

$$loss_{so}^{i}=-st_i^{cls}\log(s_i^{cls})+(st_i^{box}-s_i^{box})^2-ot_i^{cls}\log(o_i^{cls})+(ot_i^{box}-o_i^{box})^2 \tag{6}$$

其中，st 与 ot 均为真值。全体主宾组合损失值 $loss_{so}$ 如式(7)所示：

$$loss_{so}=\frac{1}{N}\sum_{i=1}^{N}loss_{so}^{i} \tag{7}$$

其中，N 为主宾组合总数。

关系推理部分损失值 $loss_{pred}$ 的计算如式(8)所示：

$$loss_{pred}=-\frac{1}{N}\sum_{i=1}^{N}t_i\log(p_i) \tag{8}$$

其中，t_i 为真值，p_i 为关系推理结果。

为满足环型结构检测部分至少能找到一条符合条件的环路的要求，需要将整个 SGiF 训练过程分作前后两个阶段，第一阶段可称为非环型训练阶段，该阶段不使用环型关系推理，即令零向量代替输入特征 x_{ok}，x_{sk}，根据 3.2 节，此时计算调整因子的权重与偏置将不会更新，以确保关系推理模型部分冻结，再令损失值式(4)中的 3 个控制系数 α,β,γ 满足关系式 $\alpha=\beta>\gamma$，从而令 SGiF 趋向于场景图结构生成。当主宾组合平均命中率(即命中真值的主宾组合数量与全体预测主宾组合数量之比)达到预定阈值时，可认为此时的 SGiF 模型满足上述要求，进入环型训练的第二阶段，此时设式(4)中的 3 个控制系数 α,β,γ 满足关系式 $\alpha=\beta=\gamma$，以此进行完整的模型训练。

4 实验

4.1 实验数据与评价指标

实验采用大型场景图生成数据集 VG[30] 作为实验数据，该数据集大约包含 1.08×10^5 张图像、5.4×10^7 条区域描述、1.7×10^7 对视觉问题答案、3.8×10^7 个物体实例、2.8×10^7 个物体属性以及 2.3×10^7 个视觉关系，平均每张图像上包含有个物体实例、个物体属性以及个视觉关系，可被用于目标检测、视觉短语生成、视觉问题回答以及场景图生成等多个计算机视觉任务。

为与基线模型进行对比，在使用 VG 之前需要根据文献[24]中的数据清理方法得到 VG 的子数据集 VG-Subset。子数据集 VG-Subset 中大约包含 8.73×10^4 张图像、5.95×10^4 个视觉关系，且共有 150 个顶点类别与 50 个关系类别。在实验中，随机抽取子集中的 1 000 张图像作为测试集，其余图像作为训练集。

实验评价指标为视觉关系检测命中率 R@K，其思想与累积匹配曲线(Cumulative Match Characteristic，CMC)相似，计算的是置信度最大的前 K 个视觉关系命中真值的比例，其命中标准为视觉关系中的两顶点与真值中的两顶点类别相同、IoU 大于或等于 0.5，以及两者的关系边类别相同。

除此之外，实验对比内容有 PredCls，SCGen 两项，前者为在给定所有顶点信息的前提下，比较模型视觉关系命中率 R@K，其实则为比较关系推理命中率，而后者为在不给定任何信息的前提下，比较视觉关系命中率 R@K。

4.2 实验工具与硬件条件

实验使用深度学习框架 Chainer[32] 和 GPU 矩阵计算库 Cupy[33] 进行模型实现与运行。整个模型训练、测试过程所使用的硬件环境为 8 核 i7-7700、16 G 内存以及 GTX10504G。

4.3 SGiF 模型工作有效性验证

为解决二步式场景图生成方法中的信息流失与视觉关系长尾分布带来模型过拟合的问题，SGiF 模型做出了两部分重要工作：1)使用多尺度特征图检测视觉关系；2)使用环型关系推理提高关系推理精度。为提高实验可靠性，首先分开验证这两部份工作的有效性，然后合并两部份工作来验证 SGiF 整体工作的有效性。

4.3.1 多尺度特征图的有效性验证

本节验证使用多尺度特征图对于提升 SGiF 模型的视觉关系检测命中率 R@50，R@100 的有效性，因此除去来自文献[25]的基线模型，还需要引入使用 RCNN 的 SGiF 模型、使用单尺度特征图的 SGiF 模型以及使用多尺度特征图的 SGiF 模型作为对照组。需注意，所有 SGiF 对照组以一层线性映射层替代环型关系推理部分。

实验中，SGiF 选择 DenseNet121[4] 作为特征提取网络。设定输入图像的大小为 512 * 512，在多尺度特征图的 SGiF 模型中，选择 DenseNet121 中的 DenseBlock2、DenseBlock3 以及 DenseBlock4 块作为特征提取层，而在单尺度特征图的 SGiF 模型中，只选择 DenseNet121 中的 DenseBlock2 块作为特征提取层。需要注意，所用的 DenseNet 模型处于初始状态(即未经预训练)，所有的模型参数将在训练过程中学习得到。实验验证结果如表 1 所列。

表 1 元素检测方式的有效性

Table 1 Effectiveness of element detection methods

实验组	SGGen/%	
	R@50	R@100
基线模型[24](二步式)	3.44	4.24
SGiF(使用 RCNN)	2.80	3.0
SGiF(使用单尺度特征图)	7.85	9.16
SGiF(使用多尺度特征图)	**9.55**	**13.0**

从表 1 可以看出，使用多尺度特征图的 SGiF 在视觉关系命中率(SGGen)上具有最高精度，在 R@50 指标上超越基线

模型 7.1%，使用单尺度特征图的 SGiF 在 R@50 上的精度略次之，而表现最差的是使用 RCNN 的 SGiF 模型，究其原因在于从 RCNN 中得到的信息是片面的，顶点间状态等重要信息被丢弃，使得场景图生成难度提高，一层线性映射模型无法胜任工作。

4.3.2 环型关系推理的有效性验证

环型关系推理中，环路信息被用于计算调整因子，而该因子又被用于调整暂时性关系推理结果，以此改变该结果的概率分布状态。在实验中，为验证环型关系推理对于提高关系推理命中率的有效性，除去来自文献[25,27]的基线模型，还需要设置环型与非环型的 SGiF 对照组，其中的非环型使用零向量代替原环路背景信息。同时，为避免多尺度特征图对验证结果的影响，所有 SGiF 对比组只使用 DenseBlock2 块的特征图，实验结果如表 2 所列。

表 2 关系推理方式的有效性

Table 2 Effectiveness of relation reasoning methods

实验组	SGGen/%	
	R@50	R@100
基线模型[24](MessagePool)	44.80	53.00
SGiF(使用非环型)	60.10	68.30
基线模型[26](Px2Graph)	68.00	75.20
SGiF(使用环型)	**70.18**	**76.05**

从表 2 中可以看出，使用环型关系推理的 SGiF 比基线模型(Px2Graph)在 R@50 上高出 2.18%，但是需要注意，基线模型(Px2Graph)相比使用非环型关系推理的 SGiF 模型具有更高的精度。回观 Px2Graph 模型为了避免检测元素重叠，其在一个像素点上预测 s_o 个顶点与 s_r 个连接边，并保留置信度超过预定阈值的结果，这种做法造成 Px2Graph 在一个像素点上产生的关系数量远多于在一个特征点上只产生一个视觉关系的 SGiF，自然地，基线模型(Px2Graph)会有更高的关系命中率。

4.3.3 整体工作有效性验证以及结论

为验证 SGiF 模型同时使用多尺度特征图与环型关系推理在场景图生成上的有效性，特设置对照组[23,25,27]进行对比，实验结果如表 3 所列。

表 3 整体工作有效性验证对比

Table 3 Validation and comparison of overall work effectiveness

(单位：%)

实验组	SGGen		PredCls	
	R@50	R@100	R@50	R@100
Message Passing[24]	3.44	4.24	44.75	53.08
VtransE[24]	9.46	10.45	62.63	62.87
Px2Graph[26]	6.70	7.80	68.80	75.20
SGiF	**12.74**	**15.21**	**72.44**	**77.10**

从表 3 中可以看出，通过结合使用多尺度特征图与环型关系推理两项工作，SGiF 模型相对对照组取得了更高的精确度，在测试项 SGGen 的 R@50 验证指标上，SGiF 超越基线模型平均精度近 6.2%，同时在测试项 PredCls 的 R@50 验证指标上，SGiF 超越基线模型平均精度近 13.7%，以此证明了 SGiF 工作在场景图生成任务上的有效性。

结束语 针对二步式场景图生成方法中的信息流失以及视觉关系长尾分布导致的模型过拟合的问题，SGiF 提出使用多尺度特征图代替目标检测模型检测场景图构成元素，以此避免信息流失，同时引入环路背景信息来调整原关系推理结果的概率分布状态，使得关系推理精度上升。最后以对比实验的形式证明了 SGiF 工作的有效性。然而，通过对 SGiF 模型的进一步分析可发现，其亦存在不足之处：1)环型关系推理的使用限制较高，其要求在场景图中最少存在一条包含目标关系边在内的环路，这种限制使得 SGiF 无法在极端条件下(如图中只有一个顶点)使用；2)使用多尺度特征图亦不能完全有效地避免元素重叠问题。于是，针对 SGiF 模型的不足，可以进行两部分改进工作：首先是降低环型关系推理的使用限制，亦或者寻找一种与环型关系推理同样有效但使用限制更低的关系推理方法，然后仿照 Px2Graph 的做法，在一个特征点上预测多组视觉关系，以此缓解元素重叠问题。

参 考 文 献

[1] KAREN S, ANDREW Z. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// International Conference on Learning Representations (ICLR). 2015.

[2] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[3] HUANG G, LIU Z, LAURENS V D M, et al. Densely Connected Convolutional Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 4700-4708.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 779-778.

[5] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 7263-7271.

[6] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement [J]. arXiv: 1804.02767.

[7] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision (ECCV). 2016: 21-37.

[8] LIN T Y, DOLLR, PIOTR, et al. Feature Pyramid Networks for Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2117-2125.

[9] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]// IEEE International Conference on Computer Vision (ICCV). 2017: 2980-2988.

[10] ROSS B G, JEFF D, TREVOR D, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014: 580-587.

[11] GIRSHICK R. Fast R-CNN[C]// IEEE International Conference on Computer Vision (ICCV). 2015: 1440-1448.

[12] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [C]// Neural Information Processing Systems (NIPS). 2015: 91-99.

[13] KAIMING H,GEORGIA G,PIOTR D,et al. Mask R-CNN[C]// IEEE International Conference on Computer Vision (ICCV). 2017:2961-2969.

[14] LI Y,QI H,DAI J,et al. Fully Convolutional Instance-aware Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:2359-2367.

[15] AGRAWAL A,LU J,ANTOL S,et al. VQA:Visual Question Answering [J]. International Journal of Computer Vision,2017,123(1):4-31.

[16] JOHNSON J,HARIHARAN B,LAURENS V D M,et al. CLEVR:A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2901-2910.

[17] ORDONEZ V,KULKARNI G,BERG T. Im2text:Describing images using 1 million captioned photographs[C]// Neural Information Processing Systems (NIPS). 2011:1143-1151.

[18] VINYALS O,TOSHEV A,BENGIO S,et al. Show and Tell:A Neural Image Caption Generator[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015:3156-3164.

[19] CARNEIRO G. Supervised Learning of Semantic Classes for Image Annotation and Retrieval [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2007,29(3):394-410.

[20] VOGEL J,SCHIELE B. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval [J]. International Journal of Computer Vision,2007,72(2):133-157.

[21] LU C,KRISHNA R,BERNSTEIN M,et al. Visual Relationship Detection with Language Priors[C]// European Conference on Computer Vision (ECCV). 2016:852-869.

[22] ZHANG H,KYAW Z,CHANG S F,et al. Visual Translation Embedding Network for Visual Relation Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:5532-5540.

[23] DAI B,ZHANG Y,LIN D. Detecting Visual Relationships with Deep Relational Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:3076-3086.

[24] XU D,ZHU Y,CHOY C B,et al. Scene Graph Generation by Iterative Message Passing[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:5410-5419.

[25] ZELLERS R,YATSKAR M,THOMSON S,et al. Neural Motifs:Scene Graph Parsing with Global Context[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018:5831-5840.

[26] NEWELL A,DENG J. Pixels to Graphs by Associative Embedding[C]// Neural Information Processing Systems (NIPS). 2017:2171-2180.

[27] LIBEN-NOWELL D,KLEINBERG J. The Link Prediction Problem for Social Networks [J]. Journal of the American Society for Information Science and Technology,2003,58(7):1019-1031.

[28] BACKSTROM L,LESKOVEC J. Supervised Random Walks:Predicting and Recommending Links in Social Networks[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. 2011:635-644.

[29] ANTOINE B,NICOLAS U,ALBERTO G D,et al. Translating Embeddings for Modeling Multi-relational Data[C]// Neural Information Processing Systems (NIPS). 2013:2787-2795.

[30] 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(1):32-73.

[31] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout:A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research,2014,15(1):1929-1958.

[32] TOKUI S,OONO K,HIDO S. Chainer:a Next-Generation Open Source Framework for Deep Learning[C]// Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS). 2015.

[33] RYOSUKE O,YUYA U,et al. CuPy:A NumPy-Compatible Library for NVIDIA GPU Calculations[C]// Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS). 2017.

ZHUANG Zhi-gang, born in 1994, postgraduate. His main research interests include scene graph gernation and meta learning.

XU Qin-lin, born in 1963, associate professor, master supervisor. His main research interests include cloud computing and software engineering.