计算机科学
COMPUTER SCIENCE

基于多模态对比学习的场景图生成方法

朱旭东, 赖腾

引用本文

朱旭东, 赖腾. 基于多模态对比学习的场景图生成方法[J]. 计算机科学, 2024, 51(11A): 231200185-5.

ZHU Xudong, LAI Teng. Multimodal Contrastive Learning Based Scene Graph Generation[J]. Computer Science, 2024, 51(11A): 231200185-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231200185

基于多模态对比学习的场景图生成方法

朱旭东　赖　腾
西安建筑科技大学信息与控制工程学院　西安 710055
(zhudongxu@vip.sina.com)

摘　要　场景图生成方法(SGG)主要研究图像中的实体及其关系,广泛应用于视觉理解与图像检索等领域。现有的场景图生成方法受限于视觉特征或单一视觉概念,导致关系识别准确率较低,且需要大量的人工标注。为解决上述问题,文中融合图像和文本特征,提出了一种基于多模态对比学习的场景图生成方法 MCL-SG(Multimodal Contrastive Learning for Scene Graph)。首先,对图像和文本输入进行特征提取,得到图像和文本特征;然后,使用 Transformer Encoder 编码器对特征向量进行编码和融合;最后,采用对比学习的自监督策略,计算图像和文本特征的相似度,通过最小化正样本和负样本之间的相似度差异完成训练,无需人工标注。通过大型场景图生成公开数据集 VG(Visual Genome)的 3 个不同层次子任务(即 SGDet,SGCls 和 PredCls)的实验表明:在 mean Recall@100 指标中,MCL-SG 的场景图检测准确率提升 9.8%,场景图分类准确率提升 14.0%,关系分类准确率提升 8.9%,从而证明了 MCL-SG 的有效性。

关键词:场景图生成;Transformer 模型;多模态;对比学习;目标检测
中图分类号　TP391

Multimodal Contrastive Learning Based Scene Graph Generation

ZHU Xudong and LAI Teng
College of Information and Control Engineering,Xi'an University of Architecture and Technology,Xi'an 710055,China

Abstract　Scene graph generation(SGG) methods play a pivotal role in studying objects and their relationships within images, with widespread applications in visual understanding and image retrieval. However, existing SGG methods are limited by visual features or individual visual concepts such as objects, resulting in a low accuracy of relationship recognition and necessitating a substantial amount of manual annotation. To address the aforementioned issues, this paper integrates image and text features and proposes a multimodal contrastive learning based scene graph generation method, multimodal contrastive learning for scene graph(MCL-SG). This method begins by extracting features from both image and text inputs, obtaining image and text features. Subsequently, a Transformer Encoder is employed to encode and fuse feature vectors, enabling a synergistic integration of information from diverse sources. Notably, MCL-SG incorporates a self-supervised contrastive learning strategy, calculating the similarity between image and text features. Training is accomplished by minimizing the dissimilarity between positive and negative samples, eliminating the need for extensive manual annotation. In this study, experiments are conducted using the VG(Visual Genome) dataset, a substantial public dataset for scene graph generation. Experiments are structured into three distinct hierarchical subtasks: SGDet, SGCls, and PredCls and the results demonstrate that, in the mean Recall@100 metric, MCL-SG achieves a 9.8% improvement in scene graph detection, a significant 14.0% enhancement in scene graph classification, and an 8.9% boost in relationship classification, thus proving the effectiveness of MCL-SG.

Keywords　Scene graph generation, Transformer model, Multimodal, Contrastive learning, Object detection

1　引言

场景图[1]由形式为〈主语-谓语-宾语〉的三元组组成,将图像内容抽象为图结构。图结构中的节点表示实体,节点间的连接(边)表示实体间的关系,具体细节如图 1 所示。场景图旨在促进对图像中复杂场景的理解,具有广泛的应用潜力,例如图像检索[1-2]、视觉问答[3-4]以及图像理解和推理[5-6]等。目前,场景图生成方法主要为基于目标检测[7-11]的两步式场景图生成方法[12-13],其具体流程为:首先使用目标检测算法检测图像中的实体(节点),然后推理实体之间的关系(边)。这种场景图生成具有方法过程简明、模块间耦合度低的优点;但由于其受限于图像的视觉特征,导致生成的关系类别不够准确。

近年来,自监督学习方法由于其无需外部监督信息的特点受到了极大的关注。对比学习[14-16]是一种有效的自监督学习方法,它的核心思想是在特征空间中将正样本和负样本进行对比,从而学习样本的特征表示,使得样本与正样本的特征表示尽可能接近,样本与负样本的表示尽可能不同。文献[17]是对比学习的一个重要里程碑,它引入了跨模态学习的概念,将图像和文本映射到统一的表示空间。这种跨模态对比学习的方法使得模型能够在处理图像和文本时具有更强的通用性和泛化能力,同时也解决了人工标注问题。

基金项目:国家重点研发计划(2019YFD1100901)
This work was supported by the National Key Research and Development Program of China(2019YFD1100901).
通信作者:赖腾(15183874170@163.com)

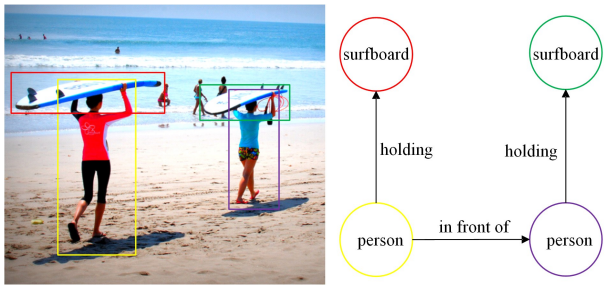

图 1 场景图结构示例

Fig. 1 Example of scene graph structure

针对现有场景图生成方法存在的问题，结合多模态对比学习，提出了一种基于多模态对比学习的场景图生成方法 MCL-SG。具体而言，该方法首先通过目标检测器[7-11]提取图像中的视觉特征、位置特征和标签特征，通过基于文献[18]的词法分析器[19]提取文本中的文本特征；然后，通过构建嵌入器获得特征嵌入，利用 Transformer 编码器[20]对特征进行编码和融合，得到视觉三元组和文本三元组；最后，采用多模态对比学习方法，计算视觉三元组和文本三元组之间的余弦相似度，通过最小化正样本和负样本之间的相似度差异来完成训练。在大型场景图生成公开数据集 Visual Genome[21]上进行实验，本文提出的场景图生成方法的性能有明显提高。

2 相关工作

2.1 场景图生成

场景图是图像内容的图结构表示，其中，节点表示图像中的实体，边表示实体间的关系。场景图生成(Scene Graph Generation，SGG)旨在从输入图像中提取这种图结构。随着大数据(特别是大规模密集注释图像场景图数据集，如 Visual Genome 数据集[21])的出现，基于深度学习的全监督场景图生成方法大量涌现，如迭代信息传递[22]、循环网络[23]、树形结构编码[24]、图卷积和剪枝[13]以及偶然推理[25]等。但这些方法依赖于大量的人工标注和标注的质量。

针对上述问题，学术界提出了一系列的弱监督场景图生成方法，例如，文献[26]提出了从非本地化场景图中学习的方法，开发了一种消息传递机制来更新被检测实体的特征，并逐步细化实体和关系的标签；文献[27]提出了使用一阶匹配的弱监督场景图生成方法。但这类方法受限于图像特征，没有考虑文本语义信息，导致关系识别的准确率较低。

为了解决这一问题，文献[28]引入文本等多模态数据，探索了以语言结构为监督的场景图生成方法。与本文方法类似，文献[29]提出利用“图像-句子”对来生成场景图，但这一方法仍然需要人工处理语言监督信息。

2.2 对比学习

对比学习是一种自监督学习方法，它在特征空间中将正样本和负样本进行对比，从而学习样本的特征表示，使得样本与正样本的特征表示尽可能接近，样本与负样本的特征表示尽可能不同。文献[30]被提出后，许多研究将对比学习推向了一个新的高度。文献[15]用动量编码器构建了一个存储库，提供一致的负样本，对 Vision Transformer[31]产生了很好的效果；文献[14]通过强大的数据增强和可学习的非线性投影，增强了负样本的表示；文献[17]引入跨模态学习的概念，将图像和文本映射到统一的表示空间，使得处理图像和文本时具有更强的通用性和泛化能力。受上述文献启发，本文将对比学习与场景图生成相结合，提出了基于多模态对比学习的场景图生成方法 MCL-SG，提高了关系识别的准确率，提升了场景图生成的性能，且无需人工标注。

3 MCL-SG

本文提出了基于多模态对比学习的场景图生成方法 MCL-SG，由特征提取、特征嵌入、特征编码和对比学习 4 部分组成，具体细节如图 2 所示。首先，对图像和文本描述进行特征提取，得到关于图像的视觉特征$\boldsymbol{x}_n^r$、位置特征$\boldsymbol{x}_n^p$和标签特征$\boldsymbol{x}_n^o$，以及关于文本描述的文本特征$\boldsymbol{y}_d^e$。然后，通过构建视觉嵌入器、标签嵌入器和文本嵌入器，嵌入相应的特征向量。最后，利用 Transformer 编码器对特征向量进行编码和融合，得到视觉三元组和文本三元组。

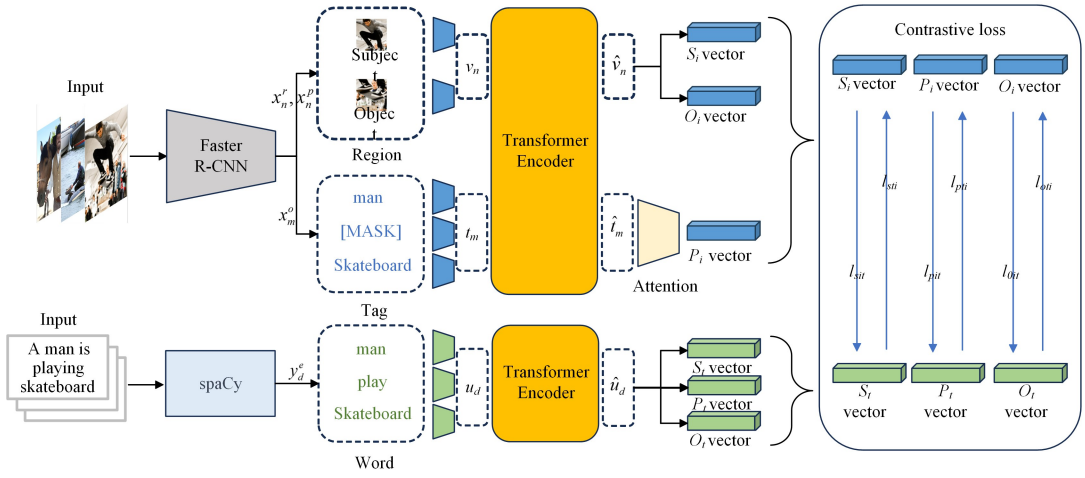

图 2 MCL-SG 结构

Fig. 2 MCL-SG structure

视觉三元组的构建函数如式(1)所示：

$$\boldsymbol{s}_i, \boldsymbol{p}_i, \boldsymbol{o}_i = e \circ t(\boldsymbol{x}_k^r, \boldsymbol{x}_l^r, \boldsymbol{x}_k^p, \boldsymbol{x}_l^p; \boldsymbol{x}_k^o, \boldsymbol{x}_l^o) \tag{1}$$

其中，$\boldsymbol{s}_i$，$\boldsymbol{p}_i$，$\boldsymbol{o}_i$分别代表构成三元组的主语、谓语和宾语的视觉特征表示；$\boldsymbol{x}_k^r$，$\boldsymbol{x}_l^r$是从图像中识别出的主语和宾语的视觉特征，由目标检测模型提取；$\boldsymbol{x}_k^p$，$\boldsymbol{x}_l^p$表示这些主语和宾语的位置特征，包括它们在图像中的相对位置和尺寸；$\boldsymbol{x}_k^o$，$\boldsymbol{x}_l^o$是标签

特征，是从目标检测模型中获得的主语和宾语的分类标签。

文本三元组的构建函数如式(2)所示：

$$\boldsymbol{s}_t,\boldsymbol{p}_t,\boldsymbol{o}_t=e^{\circ}t(\boldsymbol{y}_k^e,\boldsymbol{y}_l^e,\boldsymbol{y}_p^e) \tag{2}$$

其中，$\boldsymbol{s}_t$，$\boldsymbol{p}_t$，$\boldsymbol{o}_t$分别代表构成三元组的主语、谓语和宾语的文本特征表示；$\boldsymbol{y}_k^e$，$\boldsymbol{y}_l^e$，$\boldsymbol{y}_p^e$是从文本中提取出的词嵌入，分别对应主语词嵌入、谓语词嵌入和宾语词嵌入。这些词嵌入由 spaCy 词法分析器生成，能够捕捉每个词在语句中的语义。

3.1 输入处理

针对图像数据，使用 Faster R-CNN 目标检测器[7]来提取图像中每个区域的视觉特征$\boldsymbol{x}_n^r$、位置特征$\boldsymbol{x}_n^p$和标签特征$\boldsymbol{x}_n^o$。针对文本描述数据，首先进行标准的预处理，然后使用词法分析器[19]提取句子中的主谓宾结构，最后生成词嵌入$\boldsymbol{y}_d^e$。

3.2 嵌入器

3.2.1 视觉嵌入器

视觉嵌入器将图像区域的视觉特征$\boldsymbol{x}_n^r$和位置特征$\boldsymbol{x}_n^p$转换为嵌入$\boldsymbol{v}_n$，计算式如式(3)所示：

$$\boldsymbol{v}_n=LN(LN(\boldsymbol{W}_r\boldsymbol{x}_n^r)+LN(\boldsymbol{W}_p\boldsymbol{x}_n^p)+\boldsymbol{e}_n^t) \tag{3}$$

其中，n索引所有区域特征，包括主语区域(k)和宾语区域(l)；$\boldsymbol{W}_r$和$\boldsymbol{W}_p$是两个可学习参数矩阵；e_n^t为一个区域的类型嵌入。

3.2.2 标签嵌入器

标签嵌入器接受来自主语区域(k)和宾语区域(l)的区域标签的词嵌入$\boldsymbol{x}_k^o$和$\boldsymbol{x}_l^o$，以及一个特殊词“MASK”的词嵌入$\boldsymbol{x}_p^o$，其表示缺失的谓语。该嵌入器将输入的词嵌入和位置嵌入转换为标签嵌入$\boldsymbol{t}_m$，计算式如式(4)所示：

$$\boldsymbol{t}_m=LN(\boldsymbol{W}_e\boldsymbol{x}_m^o+\boldsymbol{e}_m^p) \tag{4}$$

其中，m索引k，l和p，$\boldsymbol{W}_e$是一个可学习参数矩阵，$\boldsymbol{e}_m^p$表示当前令牌的位置嵌入。

3.2.3 文本嵌入器

文本嵌入器接受来自文本描述的主语词嵌入$\boldsymbol{y}_k^e$、谓语词嵌入$\boldsymbol{y}_l^e$和宾语词嵌入$\boldsymbol{y}_p^e$。该嵌入器将输入的词嵌入和位置嵌入转换为文本嵌入$\boldsymbol{u}_d$，计算式如式(5)所示：

$$\boldsymbol{u}_d=LN(\boldsymbol{W}_f\boldsymbol{y}_d^e+\boldsymbol{e}_d^p) \tag{5}$$

其中，d索引k，l和p，$\boldsymbol{W}_f$是可学习参数矩阵，$\boldsymbol{e}_d^p$表示当前令牌的位置嵌入。

3.3 编码器

将视觉嵌入$\boldsymbol{v}_n$、标签嵌入$\boldsymbol{t}_m$和文本嵌入$\boldsymbol{u}_d$进一步送入多层 Transformer 编码器[20]，该编码器使用多头自注意力机制、多层感知机(MLP)和层归一化，为每个输入$\boldsymbol{v}_n$，$\boldsymbol{t}_m$和$\boldsymbol{u}_d$输出带上下文信息的嵌入$\hat{\boldsymbol{v}}_n$，$\hat{\boldsymbol{t}}_m$和$\hat{\boldsymbol{u}}_d$。这个 Transformer 编码器可以认为是在所有输入令牌之间传递消息。然后对所有的嵌入输出进行特征融合，得到视觉三元组($\boldsymbol{s}_i$，$\boldsymbol{p}_i$，$\boldsymbol{o}_i$)和文本三元组($\boldsymbol{s}_t$，$\boldsymbol{p}_t$，$\boldsymbol{o}_t$)，计算式如式(6)－式(10)所示：

$$\boldsymbol{s}_i=\hat{\boldsymbol{v}}_k+\boldsymbol{W}_v x_k^o \tag{6}$$

$$\boldsymbol{o}_i=\hat{\boldsymbol{v}}_l+\boldsymbol{W}_q\boldsymbol{x}_l^o \tag{7}$$

$$\boldsymbol{s}_t=\hat{\boldsymbol{u}}_k \tag{8}$$

$$\boldsymbol{p}_t=\hat{\boldsymbol{u}}_p+\boldsymbol{W}_s\hat{\boldsymbol{u}}_k+\boldsymbol{W}_o\hat{\boldsymbol{u}}_l \tag{9}$$

$$\boldsymbol{o}_t=\hat{\boldsymbol{u}}_l \tag{10}$$

其中，$\boldsymbol{W}_v$，$\boldsymbol{W}_q$，$\boldsymbol{W}_s$和$\boldsymbol{W}_o$都是可学习参数矩阵。特别地，对图像的谓语特征融合加入注意力机制，计算式如式(11)－式(14)所示：

$$\boldsymbol{Q}=\boldsymbol{W}_{tq}\hat{\boldsymbol{t}}_p \tag{11}$$

$$\boldsymbol{K}_s=\boldsymbol{W}_{sk}\hat{\boldsymbol{t}}_k,\boldsymbol{K}_o=\boldsymbol{W}_{ok}\hat{\boldsymbol{t}}_l \tag{12}$$

$$\boldsymbol{V}_s=\boldsymbol{W}_{sv}\hat{\boldsymbol{v}}_k,\boldsymbol{V}_o=\boldsymbol{W}_{ov}\hat{\boldsymbol{v}}_l \tag{13}$$

$$\boldsymbol{p}_i=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}_s^{\mathrm{T}}}{\sqrt{d_k}}\right)V_s+\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}_o^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_o \tag{14}$$

其中，$\boldsymbol{W}_{tq}$，$\boldsymbol{W}_{sk}$，$\boldsymbol{W}_{ok}$，$\boldsymbol{W}_{sv}$和$\boldsymbol{W}_{ov}$为可学习参数矩阵，d_k是$\boldsymbol{K}$的维度，$\boldsymbol{Q}\boldsymbol{K}_s^{\mathrm{T}}$和$\boldsymbol{Q}\boldsymbol{K}_o^{\mathrm{T}}$除以$\sqrt{d_k}$可以达到稳定训练的效果，防止权重过大。

3.4 对比损失

训练过程中，从训练数据中抽取一个小批量的N个输入对(Image，Text)，并分别计算三元组的主语表示对($\boldsymbol{s}^v$，$\boldsymbol{s}^n$)、谓语表示对($\boldsymbol{p}^v$，$\boldsymbol{p}^n$)和宾语表示对($\boldsymbol{o}^v$，$\boldsymbol{o}^n$)。训练目标包含三组损失函数。第一组损失函数的第i对图像-文本对比损失如式(15)所示：

$$\ell_i^{(\mathrm{sit})}=-\log\frac{\exp(\langle\boldsymbol{s}_i^v,\boldsymbol{s}_i^n\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle\boldsymbol{s}_i^v,\boldsymbol{s}_k^n\rangle/\tau)} \tag{15}$$

其中，τ为可调参数。第i对文本-图像对比损失如式(16)所示：

$$\ell_i^{(\mathrm{sti})}=-\log\frac{\exp(\langle\boldsymbol{s}_i^n,\boldsymbol{s}_i^v\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle\boldsymbol{s}_i^n,\boldsymbol{s}_k^v\rangle/\tau)} \tag{16}$$

同理可以得到第二组损失函数($\ell_i^{(\mathrm{pit})}$，$\ell_i^{(\mathrm{pti})}$)和第三组损失函数($\ell_i^{(\mathrm{oit})}$，$\ell_i^{(\mathrm{oti})}$)。主语综合损失$\mathscr{L}_s$、谓语综合损失$\mathscr{L}_p$和宾语综合损失$\mathscr{L}_o$的计算式如式(17)－式(19)所示：

$$\mathscr{L}_s=\frac{1}{N}\sum_{i=1}^{N}(\ell_i^{(\mathrm{sit})}+\ell_i^{(\mathrm{sti})}) \tag{17}$$

$$\mathscr{L}_p=\frac{1}{N}\sum_{i=1}^{N}(\ell_i^{(\mathrm{pit})}+\ell_i^{(\mathrm{pti})}) \tag{18}$$

$$\mathscr{L}_o=\frac{1}{N}\sum_{i=1}^{N}(\ell_i^{(\mathrm{oit})}+\ell_i^{(\mathrm{oti})}) \tag{19}$$

最后复合损失函数如式(20)所示：

$$\mathscr{L}=\lambda_s\mathscr{L}_s+\lambda_p\mathscr{L}_p+\lambda_o\mathscr{L}_o \tag{20}$$

其中，λ_s，λ_p，λ_o为可调参数。

4 实验结果与分析

为了验证 MCL-SG 的有效性，本文在大型场景图生成公开数据集 Visual Genome(VG)[21]上进行了不同层级的监督实验，并在 SGDet，SGCls 和 PredCls 3 个层次的子任务上对该方法进行了性能评估。

4.1 数据集与实验设置

VG 数据集是一个常用的大型视觉关系数据集，包含了图像中的实体、关系以及区域文本描述。在其原始数据集中，108077 张图像共有超过 8 万个实体类别，实体标注非常杂乱。因此，在实验中使用最普遍的数据预处理和数据集划分方法[22]，保留了最常见的 150 个实体类别和 50 个关系类别。其中，训练集占 70%，测试集占 30%。此外，保留了 5000 张图像作为验证集，以检测本文方法的有效性。

场景图生成分为 3 个子任务：

1)场景图检测任务(Scene Graph Detection，SGDet)：需要同时定位实体并识别实体和关系类别。

2)场景图分类任务(Scene Graph Classification，SGCls)：

在只给定实体边界框的情况下，同时对实体类别和关系类别进行分类。

3)关系分类任务(Predicate Classification，PreCls)：在给定真实的实体边界框和实体类别的情况下对关系类别进行分类。

4.2 评价指标

在以往的工作中[12,22]，场景图生成的子任务通常采用Recall@K(R@K)来评估方法的性能，该指标表示每个图像测试样本的前 K 个置信度最高的关系三元组的召回率。但由于VG数据集的长尾效应，传统的R@K指标偏好于头部类别样本，不能证明不同方法对低频关系的有效性，无法反映来自长尾数据的影响。因此，与文献[32]的工作类似，本文在全监督实验中采用了mean Recall@K(mR@K)作为评价指标。mR@K可以更公平地评估模型的性能，将每个关系类别视为平等，不会因样本数量而给予头部关系更多的重视。该指标单独计算每个关系类别的召回率R@K，然后求其平均值。此外，为保证每对主宾实体只有一个关系预测，所有实验都在图约束条件下进行。

4.3 实验细节

MCL-SG使用预训练的Faster R-CNN[7]作为目标检测器，保留每张图像的前36个实体，并从检测器中提取1536维区域特征；使用文献[18]中基于规则的词法分析器[19]，解析图像中的区域文本描述得到主谓宾三元组。编码器结构由文献[20]的Transformer编码器实现，它有2个自注意力层，每层有12个注意力头。所有实验均采用SGD优化器进行训练，epoch大小和批处理大小均采用原始方法设置，初始学习率为0.001。

4.4 弱监督实验分析

将MCL-SG与其他弱监督场景图生成方法进行对比分析，结果如表1所列。

表1 弱监督SGDet的实验结果

Table 1 Results of weakly supervised SGDet

Method	R@50	R@100
VSPNet[26]	4.7	5.4
LSWS[28]	7.3	8.7
SGNLS[29]	10.0	11.5
MCL-SG	**10.9**	**12.4**

VSPNet[26]将实体和关系表示为两种类型的节点，并将主语和宾语视为它们之间的两类语义边，然后用图对齐算法训练弱监督场景图生成模型；LSWS[28]设计了一个注意力矩阵来完成文本图结构与被检测实体之间的对齐，然后发送匹配标签来指导场景图生成的训练过程，并通过迭代实现优化；SGNLS[29]利用检测器提供的检测标签生成伪标签，然后送入全监督场景图生成模型。从实验结果可以看出，MCL-SG优于其他弱监督场景图生成方法，与SGNLS方法相比，R@50和R@100指标分别提升了9%和7.8%，证明了MCL-SG可以在弱监督下高质量的检测图像中的场景图。

4.5 全监督实验分析

为验证MCL-SG在全监督下的有效性，将Motifs[23]和VCTree[32]等主流的场景图生成方法与MCL-SG在3个子任务SGDet，SGCls和PredCls上进行对比实验，结果如表2—表4所列。

表2 全监督SGDet的实验结果

Table 2 Results of fully supervised SGDet

Method	mR@20	mR@50	mR@100
Motifs[23]	4.1	5.5	6.8
VCTree[32]	4.2	5.7	6.9
EBL[33]	5.7	7.7	9.1
SG[34]	6.4	8.3	9.2
MCL-SG	**7.1**	**8.8**	**10.1**

表3 全监督SGCls的实验结果

Table 3 Results of fully supervised SGCls

Method	mR@20	mR@50	mR@100
Motifs[23]	6.5	8.0	8.5
VCTree[32]	10.4	7.5	7.9
EBL[33]	6.2	12.5	13.5
SG[34]	8.9	11.2	12.1
MCL-SG	**10.6**	**12.9**	**13.8**

表4 全监督PredCls的实验结果

Table 4 Results of fully supervised PredCls

Method	mR@20	mR@50	mR@100
Motifs[23]	11.5	14.6	15.8
VCTree[32]	11.7	18.2	16.1
EBL[33]	14.2	14.9	19.7
SG[34]	14.5	18.5	20.2
MCL-SG	**16.0**	**20.2**	**22.0**

从实验结果可以看出，本文提出的MCL-SG在SGDet，SGCls和PredCls 3个子任务上的平均召回率mR@K均得到了一定的提高。与SG[34]相比，在SGDet子任务上mR@K指标提升9.8%，在SGCls子任务上mR@100指标提升14.0%，在PredCls子任务上mR@100指标提升8.9%，表明MCL-SG有效地提高了关系识别的准确率，从而提升了场景图生成的性能。

结束语 本文提出了一种基于多模态对比学习的场景图生成方法MCL-SG，通过融合图像和文本等多模态信息提升了特征的表现能力，提高了关系识别准确率；此外，其利用对比学习的自监督策略进行训练，无需人工标注。在VG数据集上进行不同层级的监督实验，将MCL-SG与多种主流的场景图生成方法进行了对比分析。结果表明，融合了图像和文本信息的MCL-SG有效地提高了关系识别的准确率，场景图生成的性能也得到了提升，证明了MCL-SG的有效性。

参 考 文 献

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

[2] WANG S, WANG R, YAO Z, et al. Cross- modal scene graph matching for relationship-aware image-text retrieval[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 1508-1517.

[3] GHOSH S, BURACHAS G, RAY A, et al. Generating natural language explanations for visual question answering using scene graphs and visual attention[J]. arXiv: 1902.05715, 2019.

[4] DAMODARAN V, CHAKRAVARTHY S, KUMAR A, et al. Understanding the role of scene graphs in visual question answering[J]. arXiv: 2101.05479, 2021.

[5] ADITYA S, YANG Y, BARAL C, et al. Image understanding using vision and reasoning through scene description graph[J]. Computer Vision and Image Understanding, 2018, 173: 33-45.

[6] ZHANG J, KALANTIDIS Y, ROHRBACH M, et al. Large-scale visual relationship understanding [C]// Proceedings of the

AAAI Conference on Artificial Intelligence. 2019：9185-9194.
[7] RENS，HE K，GIRSHICK R，et al. Faster R-CNN：Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analaysis and Machine Intelligence，2016，39(6)：1137-1149.
[8] YANG J，LU J，LEE S，et al. Visual curiosity：Learning to ask questions to learn visual recognition[J]. arXiv：1810. 00912，2018.
[9] JERBI A，HERZIG R，BERANT J，et al. Learning object detection from captions via textual scene attributes[J]. arXiv：2009. 14558，2020.
[10] YE K，ZHANG M，KOVASHKA A，et al. Cap2det：Learning to amplify weak caption supervision for object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：9686-9695.
[11] ZAREIAN A，ROSA K D，HU D H，et al. Open-vocabulary object detection using captions[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：14393-14402.
[12] LU C，KRISHNA R，BERNSTEIN M，et al. Visual relationship detection with language priors[C]// Computer Vision-ECCV 2016：14th European Conference，Amsterdam，The Netherlands，Part I 14. Springer International Publishing，2016：852-869.
[13] YANG J，LU J，LEE S，et al. Graph R-CNN for scene graph generation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：670-685.
[14] CHEN T，KORNBLITH S，NOROUZI M，et al. A simple framework for contrastive learning of visual representations[C]// International Conference on Machine Learning. PMLR，2020：1597-1607.
[15] HE K，FAN H，WU Y，et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：9729-9738.
[16] GRILL J B，STRUB F，ALTCHé F，et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in Neural Information Processing Systems，2020，33：21271-21284.
[17] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR，2021：8748-8763.
[18] SCHUSTER S，KRISHNA R，CHANG A，et al. Generating semantically precise scene graphs from textual descriptions for improved image retrieval[C]// Proceedings of the Fourth Workshop on Vision and Language. 2015：70-80.
[19] WU H，MAO J，ZHANG Y，et al. Unified visual-semantic embeddings：Bridging vision and language with structured meaning representations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：6609-6618.
[20] CHEN Y C，LI L，YU L，et al. Uniter：Universal image-text representation learning[C]// European Conference on Computer Vision. Cham：Springer International Publishing，2020：104-120.
[21] 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.
[22] 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.
[23] 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.
[24] WANG W，WANG R，SHAN S，et al. Sketching image gist：Human-mimetic hierarchical scene graph generation[C]// European conference on computer vision. Cham：Springer International Publishing，2020：222-239.
[25] 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.
[26] ZAREIAN A，KARAMAN S，CHANG S F. Weakly supervised visual semantic parsing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：3736-3745.
[27] SHI J，ZHONG Y，XU N，et al. A simple baseline for weakly-supervised scene graph generation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：16393-16402.
[28] YE K，KOVASHKA A. Linguistic structures as weak supervision for visual scene graph generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：8289-8299.
[29] ZHONG Y，SHI J，YANG J，et al. Learning to generate scene graph from natural language supervision[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：1823-1834.
[30] OORD A，LI Y，VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv：1807. 03748，2018.
[31] CHEN X，XIES，HE K. An empirical study of training self-supervised vision transformers[C]// CVF International Conference on Computer Vision(ICCV). 2021：9620-9629.
[32] 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.
[33] SUHAIL M，MITTAL A，SIDDIQUIE B，et al. Energy-based learning for scene graph generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：13936-13945.
[34] KHANDELWAL S，SUHAIL M，SIGAL L. Segmentation-grounded scene graph generation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：15879-15889.

ZHU Xudong，born in 1973，Ph.D，associate professor，master supervisor. His main research interests include privacy-preserving and scene graph generation.

LAI Teng，born in 1998，postgraduate. His main research interests include scene graph generation and deep learning.