北京大学学报(自然科学版) 第 58 卷 第 1 期 2022 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 58, No. 1 (Jan. 2022)
doi: 10.13209/j.0479-8023.2021.110

引入图像信息的多模态复述生成模型

马超 万璋 张玉洁† 徐金安 陈钰枫

北京交通大学计算机与信息技术学院, 北京 100044; † 通信作者, E-mail: yjzhang@bjtu.edu.cn

摘要 在商品描述、新闻评论等多模态场景下, 已有复述生成模型只能围绕文本信息生成复述。为了解决其因无法利用图像信息而导致的语义丢失问题, 提出多模态复述生成模型(multi-modality paraphrase generation model, MPG)来引入图像信息, 并用其生成复述。在 MPG 中, 为了引入与原句对应的图像信息, 首先根据原句构建抽象场景图, 并将与原句相关联的图像区域特征转换为场景图的结点特征。进一步地, 为了利用构建好的场景图来生成语义一致的复述句, 使用关系图卷积神经网络和基于图的注意力机制对图结点特征进行编码和解码。在评测阶段, 提出句对相似度计算方法, 从 MSCOCO 数据集中筛选出描述图像中相同物体的句对, 并将其作为复述测试集进行评测。实验结果显示, 所提出的 MPG 模型生成的复述拥有更好的语义忠实度, 表明在多模态场景下图像信息的引入对提高复述生成质量的有效性。

关键词 复述生成; 多模态; 抽象场景图; 注意力机制

Multi-modality Paraphrase Generation Model Integrating Image Information

MA Chao, WAN Zhang, ZHANG Yujie†, XU Jin'an, CHEN Yufeng

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: yjzhang@bjtu.edu.cn

Abstract In multi-modality scenarios such as commodity descriptions and news comments, existing paraphrase generation models can not utilize information from image and therefore result in the loss of semantics in the generated paraphrases. In order to solve this problem, this paper first propose the Multi-modality Paraphrase Generation (MPG) model to integrate image information for paraphrase generation. In MPG, in order to integrate the image information corresponding to the original sentence, the authors first construct an abstract scene graph and transform the image features into node features of the scene graph. Furthermore, the constructed scene graph was utilized to generate paraphrase, by using the relational graph convolutional neural network for encoder and graph-based attention mechanism for decoder. In the evaluation stage, a sentence pair similarity calculation method was proposed to select sentence pairs describing same objects from the MSCOCO data set, and then evaluation experiments were conducted. Experimental results show that the proposed MPG model achieve better semantic fidelity, which indicates that the integration of image information is effective in improving the quality of the paraphrase generation in multi-modality scenarios.

Key words paraphrase generation; multi-modality; abstract scene graph; attention mechansim

复述生成指在同一种语言内, 为给定的句子生成语义相同但表达形式不同的句子(即复述句), 例如给定句子"how far is Earth from Moon"生成复述句"what is the distance between Moon and Earth"。复述生成技术广泛地应用于自动问答[1]、机器翻译[2]和文本摘要[3]等任务中。在自动问答任务中, 利用复述句可以扩展检索文本, 提高检索系统的性能。在机器翻译和文本摘要任务中, 利用复述生成的多个参考译文和文摘可以使自动评测系统的性能得到有效的提升。目前的复述方法[4–5]只能围绕文

国家自然科学基金(61876198, 61976015, 61976016)资助
收稿日期: 2021–06–09; 修回日期: 2021–08–17

本形式的信息生成复述, 随着互联网的发展和社交媒体的传播, 我们在生活中遇到的大量信息不仅限于文本形式, 而是同时包含图像和文本的多模态形式, 并且这两种形式的信息相互关联。例如, 电商平台上商品的图片常伴随对商品的描述, 新闻网站和社交媒体上的图片也往往附带相关的描述和评论。这种包含文本和图像的多模态形式信息比单调的文本形式信息更能够吸引用户, 也更有助于用户理解。在这样广泛存在的多模态形式信息中, 图像信息包含丰富的语义, 而现有的复述生成方法只依赖于文本信息, 其生成的复述句难以包含图像中丰富的语义信息。

具体而言, 文本复述模型无法利用图像信息会导致两个方面的问题。在语义保持度方面, 传统模型在对文本信息生成复述的过程中, 大量的客观图像信息没有得到利用, 导致模型生成的复述句在语义上与原句存在偏离, 语义保持度较低。在复述的多样性方面, 文本形式的原句可能仅仅从某个观察角度, 对图像中相关物体进行描述, 而在其他观察角度上对图像中相同物体的描述应该属于相同语义的不同表达。例如, 空间中多个物体之间的位置关系可以通过不同的观察角度来描述, 但是它们位置上的逻辑关系并没有改变, 可以认为它们互为复述。然而, 传统的文本复述模型很难学习到这种相关物体之间的空间位置关系等图像特有的信息, 无法对相关的图像信息生成不同观察角度上的多样性表达。

为了解决上述问题, 我们考虑引入图像信息为商品描述、新闻描述生成更加多样化的表达, 不仅能在此类场景中为用户提供更加丰富、全面的描述, 还可以应用于扩展商品或新闻的检索文本, 提高搜索引擎的效率。在本研究中, 我们首次提出基于图像和文本的多模态复述生成任务: 给定原句以及与其相关联的图像, 利用与原句相对应的图像信息来生成原句的复述句。针对该任务, 我们进一步提出引入图像信息的多模态复述生成模型(multi-modality paraphrase generation model, MPG)。图像中包含的语义信息非常丰富, 远多于句子级的描述表达的信息。句子级的描述往往只对应原始图像的部分区域所蕴含的语义信息, 其余的图像区域对于生成原句的复述是冗余信息。因此, 我们在多模态范围内将复述更具体地定义为在描述一幅图像中相同物体的情况下, 对同一语义的不同形式的表达。在 MPG 中, 为了保证复述句与原句对于图像所关注的物体具有一致性, 我们引入抽象场景图[6], 根据原句对应的相关物体, 选择性地获取图像信息。抽象场景图是由实体、属性和关系 3 种抽象结点组成的有向图, 其中不包含语义标签, 只包含一个抽象的结构, 与图像中相关物体的区域相对应。我们将原句在抽象场景图中对应的图像信息作为原句在图像中反映的语义信息。进一步地, 为了利用原句对应的图像信息生成与原句语义一致的复述句, 我们使用关系图卷积神经网络和基于图的注意力机制, 对结点特征进行编码和解码。最后, 为了更准确地衡量模型的性能, 我们围绕定义中对“图像内相同物体”的这一限定, 提出相似度计算方法, 最大程度地对 MSCOCO 数据集中描述相同物体的句对进行筛选, 并制作相关的测试集进行评测。

1 相关研究

传统的文本意义上的复述定义指在同一种语言内相同语义的不同形式的表达。在多模态领域也有相关工作对复述的概念进行定义。Chu 等[7]首先提出使用图像信息来关联不同的视觉概念词, 他们将一幅图像中对同一视觉概念的不同短语表达称为基于视觉定位的复述(visually grounded paraphrases), 并提出基于图像注意力机制的相似度计算模型, 进一步利用计算得到的相似度进行聚类, 最后得到基于视觉定位的复述。因此, 该项工作对多模态复述的定义主要是在短语级别上进行的, 而非句子级别。Liu 等[8]提出一种利用视觉复述对的两阶段解码模型, 为同一幅图像生成多样化的图像描述。该项工作将多模态复述定义为描述相同图像的不同句子。但是, 由于图像中的信息非常丰富, 对同一幅图像的不同描述可能会由于关注的物体不同而具有完全不同的语义, 这样的定义也有别于传统的文本意义上复述的定义中对相同语义的限定。

本文在多模态领域将复述定义为在描述一幅图像中相同物体的情况下, 对相同语义不同形式的表达。这样, 我们不仅在多模态领域对复述在句子级别进行定义, 也是对传统文本意义上复述的定义在语义不变性上的延续。根据我们的定义, 本文进一步引入抽象场景图来获取图像信息, 并使用图卷积神经网络和基于图的注意力机制, 对相关物体包含的图像信息进行编码和解码, 从而在对应图像内相关物体相同的前提下, 生成句子级的复述。

在与复述任务类似的机器翻译任务中，过去几年陆续出现利用图像信息来提高翻译质量的相关工作。根据跨模态的学习方法，可将基于多模态的机器翻译工作分为两大类：一类是在训练和推理过程中，显式地将视觉特征和文本特征从一个模态转换到另一个模态[9]；另一类是在训练过程中，隐式地对齐文本模态和视觉模态，生成基于视觉的文本特征。相关的研究表明，在Multi30K数据集上，大部分多模态机器翻译模型的表现都十分出色[10−11]。在机器翻译领域引入多模态信息后的大量成功实践，深刻揭示了将多模态信息应用在自然语言处理任务上的有效性和必要性。本文借鉴多模态机器翻译的相关工作，首次在复述任务上对图像信息的利用开展研究。

2 引入图像信息的多模态复述生成模型

本文围绕有效地获取图像信息，并利用其生成复述的问题，提出引入图像信息的多模态复述生成模型。模型输入一幅图像以及一句长度为 n 的图像描述 $C=\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$，生成长度为 m 的复述句 $Y=\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$。为了保证原句 C 与复述句 Y 描述的图像物体的一致性，我们利用抽象场景图，根据原句，从图像中获取并限制生成复述所需的图像特征。进一步地，为了利用图像特征生成与原句语义一致但表达形式不同的复述句，我们采用关系图卷积神经网络，对结点上下文信息进行编码。最后，基于两层LSTM网络，使用基于图的注意力机制，对编码后的抽象场景图进行解码，生成复述句。模型的整体框架如图1所示。

2.1 抽象场景图的构建

首先，输入原句 $C=\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$ 和一幅与原句相关联的图像 I。为了保证原句与复述句在图像中相关物体的一致性，我们使用抽象场景图，根据原句对应的相关物体，选择性地引入图像信息。为此，我们使用Stanford scene graph parser[12]场景图生成器，生成与原句 C 相关的场景图 G'。场景图(scene graph)包含实体结点 o、属性结点 a 和关系结点 r 这3种类型的结点以及结点之间的有向边。实体结点对应图中的某个物体；属性结点是对实体结点对应物体属性(如颜色、大小和状态等)的具体描述，并依附于实体结点，同一个实体结点可以依附有多个属性结点；关系结点是对两个实体结点对应物体之间关系的描述。场景图中每个结点都有相应的语义标签和结点类型标签。

具体地，使用Faster-RCNN[13]作为目标检测器，对图像 I 中的物体进行检测，得到一组候选物体的

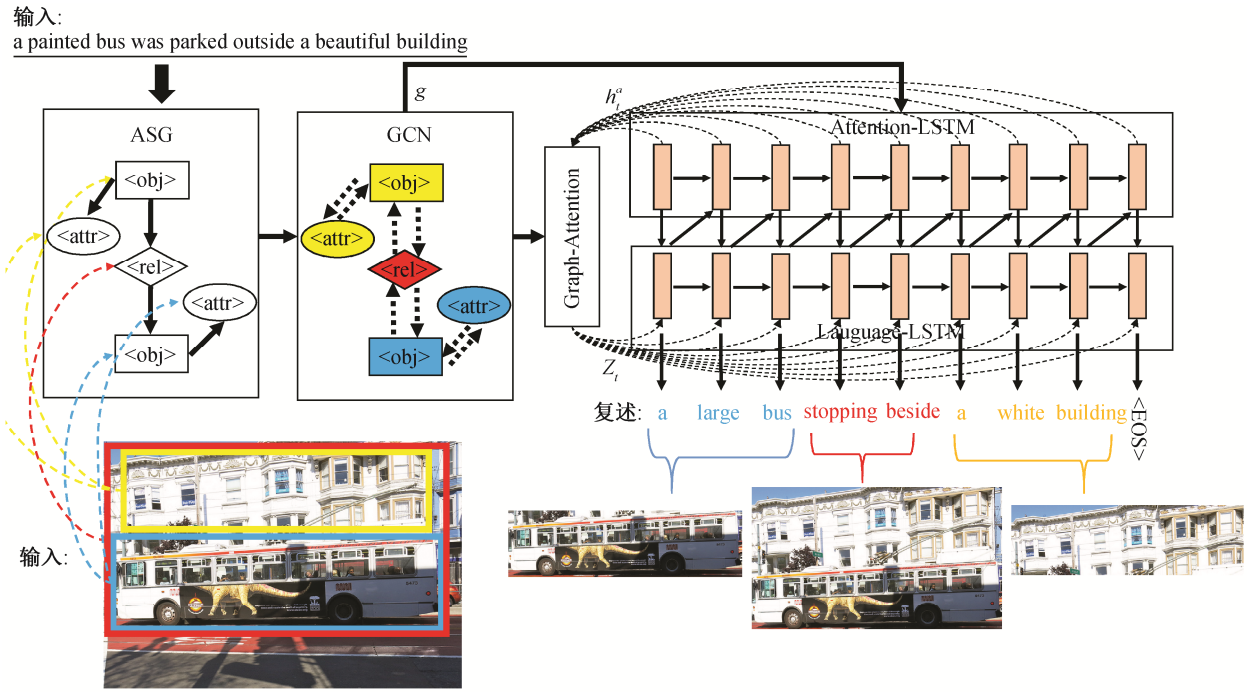

图 1　本文提出的引入图像信息的多模态复述生成模型的整体框架

Fig. 1　The proposed framework of multi-modality paraphrase generation model integrating image information

语义标签以及候选物体在图像中相应的区域坐标和区域特征。进一步地，计算 G'中实体结点的语义标签与实体检测得到的候选物体的语义标签之间的语义相似度，利用语义相似度将实体结点和候选物体进行匹配，进而通过已知候选物体的区域坐标确定 G'中每个实体结点对应的区域坐标。最后，将 G'中所有结点的语义标签全部移除，只保留图形布局和结点的部分信息(只包含结点类型和区域坐标)，得到抽象场景图 G，它与原句描述的图像相关区域相对应。我们将抽象场景图(abstract scene graph)记为 $G=(V, E)$(V 和 E 分别为结点集和边集)，其中结点之间的边都是单向的，其结构如图 1 中左侧部分所示。

2.2 基于关系图卷积神经网络的编码器

为了利用原句对应的图像信息生成复述句，首先对抽象场景图中的所有结点进行编码。通过前面构造的抽象场景图，得到图中每个实体结点的区域坐标。然后，对抽象场景图 G 中第 i 个结点的特征 $\boldsymbol{v}_i$ 进行初始化。当结点 i 是实体结点时，通过该实体结点在图中对应物体的区域坐标得到相应的区域图像特征，并将其作为该实体结点的特征 $\boldsymbol{v}_i$。当结点 i 是属性结点时，通过该属性结点找到其依附的实体结点，并将实体结点的区域图像特征作为该属性结点的特征 $\boldsymbol{v}_i$。当结点 i 是关系结点时，通过该关系结点所连接的两个实体结点找到这两个实体结点对应的区域坐标，然后以区域坐标的并集作为该关系结点的区域坐标，得到相应的区域特征，并将其作为该关系结点的特征 $\boldsymbol{v}_i$。由于实体结点和属性结点的特征是从相同物体对应的图像区域提取，仅依赖图像区域特征不能很好地区分结点的类型。为了解决这个问题，进一步对结点进行类型编码，得到编码后的结点特征 $\boldsymbol{x}_i^{(0)}$。

$$\boldsymbol{x}_i^{(0)}=\begin{cases}\boldsymbol{v}_i\odot\boldsymbol{W}_r[0], & i\in o;\\ \boldsymbol{v}_i\odot(\boldsymbol{W}_r[1]+\mathbf{pos}[i]), & i\in a;\\ \boldsymbol{v}_i\odot\boldsymbol{W}_r[2], & i\in r。\end{cases}\tag{1}$$

其中，$\boldsymbol{W}_r\in\mathbb{R}^{3\times d}$ 为可学习的参数矩阵，d 为特征维数，$\boldsymbol{W}_r[k]$表示 $\boldsymbol{W}_r$ 第 k 行。$\mathbf{pos}[i]$是位置向量，用来区分连接同一实体结点的不同属性节点的顺序。通过类型编码，得到抽象场景图 G 中每个结点的特征表示 $X=\{\boldsymbol{x}_1^{(0)},...,\boldsymbol{x}_{|V|}^{(0)}\}$。

虽然抽象场景图的边是单向的，但是所有结点之间都存在上下文信息的交互。为了更好地编码结点之间的上下文信息，我们将抽象场景图 G 扩展为双向的多关系图 $\mathcal{G}=\{\mathcal{V},\mathcal{E},\mathcal{R}\}$。进一步地，使用关系图卷积神经网络[14]对$\mathcal{G}$进行上下文编码:

$$\boldsymbol{x}_i^{(l+1)}=\sigma\left(\boldsymbol{W}_0^{(l)}\boldsymbol{x}_i^{(l)}+\sum\nolimits_{s\in\mathcal{R}}\sum\nolimits_{j\in N_i^s}\frac{1}{|N_i^s|}\boldsymbol{W}_s^{(l)}\boldsymbol{x}_j^{(l)}\right),\tag{2}$$

式中，N_i^s 表示关系 $s\in\mathcal{R}$ 下第 i 个节点的相邻结点，σ 为 ReLU 激活函数，$\boldsymbol{W}_*^{(l)}$ 为关系图卷积神经网络第 l 层需要学习的参数。利用一个层，可以为每个结点编码直接相邻结点的上下文信息；堆叠多个层，可以在图中编码更广泛的上下文信息。我们堆叠 L 层，最后使用第 L 层的输出作为最终的结点特征表示 $X=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{|V|}\}$。通过取 X 的平均值，得到一个全局图结点嵌入 $\hat{\boldsymbol{g}}=\frac{1}{|V|}\sum_i\boldsymbol{x}_i$。我们将全局图结点嵌入与使用 Faster-RCNN 得到的全局图像特征融合为全局编码特征 $\boldsymbol{g}$。

2.3 基于图的注意力机制解码器

为了利用编码后的抽象场景图来生成与原句对应的复述句，我们使用基于图的注意力机制解码器对抽象场景图进行解码。在解码器使用两层 LSTM 网络结构，分别为注意力层和复述层。首先将全局编码特征 $\boldsymbol{g}$ 输入注意力层中，得到 t 时刻注意力层的隐层状态 $\boldsymbol{h}_t^a$:

$$\boldsymbol{h}_t^a=\mathrm{LSTM}([\boldsymbol{g};\boldsymbol{\omega}_{t-1};\boldsymbol{h}_{t-1}^l],\boldsymbol{h}_{t-1}^a;\boldsymbol{\theta}^a),\tag{3}$$

式中，$\boldsymbol{\omega}_{t-1}$ 为上一时刻的词向量，$\boldsymbol{h}_{t-1}^l$ 为上一时刻复述层的隐层状态，$\boldsymbol{h}_{t-1}^a$ 为上一时刻注意力层的隐层状态，[;]表示向量的连接，$\boldsymbol{\theta}^a$ 为可学习的参数。

在 t 时刻，得到注意力层输出的隐层状态 $\boldsymbol{h}_t^a$，通过基于关系图卷积神经网络的编码器，得到抽象场景图中的全部结点表示 $X=\{\boldsymbol{x}_1,\ ...,\ \boldsymbol{x}_{|V|}\}$。为了提取结点的特征，我们基于图注意力机制，计算 $\boldsymbol{h}_t^a$ 与结点向量集 X 中结点的相似度，并将它作为 t 时刻的注意力权重 $\boldsymbol{\alpha}_t$:

$$\hat{\boldsymbol{\alpha}}_{t,i}=\boldsymbol{\omega}_c^{\mathrm{T}}\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_i+\boldsymbol{W}_{hc}\boldsymbol{h}_t^a),\tag{4}$$

$$\boldsymbol{\alpha}_t=\mathrm{softmax}(\hat{\boldsymbol{\alpha}}_t),\tag{5}$$

式中，$\boldsymbol{W}_{xc}$，$\boldsymbol{W}_{hc}$ 和 $\boldsymbol{\omega}_c$ 是参数。我们通过注意力权重 $\boldsymbol{\alpha}_t$ 计算，得到 t 时刻的上下文语义向量 $\boldsymbol{Z}_t=\sum_{i=1}^{|V|}\boldsymbol{\alpha}_{t,i}\boldsymbol{x}_i=\boldsymbol{\alpha}_t^{\mathrm{T}}\boldsymbol{X}$。

将 t 时刻的上下文语义向量 $\boldsymbol{Z}_t$ 以及 t 时刻注意

力层的隐层状态 $\boldsymbol{h}_t^a$ 输入到复述层，得到 t 时刻复述层的隐层状态 $\boldsymbol{h}_t^l$：

$$\boldsymbol{h}_t^l = \mathrm{LSTM}([\boldsymbol{Z}_t;\boldsymbol{h}_t^a],\boldsymbol{h}_{t-1}^l;\boldsymbol{\theta}^l)。\tag{6}$$

最终，通过 t 时刻复述层的隐层状态 $\boldsymbol{h}_t^l$，利用 softmax 分类器预测 t 时刻生成单词的概率来生成下一个单词:

$$p(\boldsymbol{y}_t|y<t)=\mathrm{softmax}(\boldsymbol{W}_p\boldsymbol{h}_t^l+\boldsymbol{b}_p)。\tag{7}$$

式(6)和(7)中，$\boldsymbol{\theta}^l$，$\boldsymbol{W}_p$ 和 $\boldsymbol{b}_p$ 都是可学习的参数。

2.4 模型训练和实施细节

根据输入的原句 $C=\{\boldsymbol{c}_1,\ \ldots,\ \boldsymbol{c}_n\}$ 和图像 I，生成长度为 m 的复述句 $Y=\{\boldsymbol{y}_1,\ \ldots,\ \boldsymbol{y}_m\}$，我们采用交叉熵损失训练整个模型。目标函数为

$$L=-\log\sum_{i=1}^{m}p\left(\boldsymbol{y}_i|\boldsymbol{y}_1\ldots\boldsymbol{y}_{i-1},C,I\right)。\tag{8}$$

我们使用在 VisualGenome 数据集上预训练的 Faster-RCNN，为抽象场景图的结点提取对应的图像特征，并使用在 ImageNet[15] 上预训练的 ResNet 152 来抽取全局的图像特征。在基于关系图的卷积网络编码器中，设置图卷积层数 $L=2$，并设置特征维度为 512 维。解码器中的隐藏层大小设置为 512 维，词向量维度设置为 512 维。设置 Batch 的大小为 64，并在解码器使用 dropout 正则化技术，设置 drop 率为 0.3。在训练过程中，学习率设置为 0.0001。在测试过程中，使用 beam search 算法，大小设置为 5。

3 评测实验与结果分析

为了验证本文所提方法的有效性，保证原句与复述句所描述相关物体的一致性，我们设计相似度计算方法，对 MSCOCO 数据集进行筛选，重新制作测试集，并进行相关的评测实验。

3.1 实验数据

在 MSCOCO 数据集中，训练集包含超过 110K 幅图像，验证集 5K 幅和测试集 5K 幅，每幅图像最多包含 5 个人工标注的图像描述。传统的文本复述工作使用复述句对(包括原句和参考句)进行训练，生成复述句，复述句的语义来源于原句。与之不同，本文的多模态复述生成模型在训练过程中不使用原句的语义信息，而是利用原句提取抽象场景图，进而利用与句子相关联的图像特征作为语义信息。因此，为了在训练中最大程度地保证原句与参考句描述相关物体的一致性，得到最优的训练效果，同时也能在有限的数据集上扩大训练数据的规模，最大化模型的泛化能力，我们在训练过程中将原句同时作为参考句输入解码器中，计算交叉熵对模型进行训练，而不采取传统的文本复述模型中使用复述句对的训练方式。

在测试过程中，与传统的文本复述工作相同，我们使用原句与参考句构成的复述句对进行测试，计算通过原句生成的复述句与参考句的相关指标(如 BLEU)来衡量模型的性能。以往的文本复述工作在 MSCOCO 数据集上对复述句对的选取是随机的，即在同一幅图像的 5 条图像描述中，任意选择 4 条描述，构建为两对复述句对用于测试。在这些工作中，将同一幅图像的不同描述视为复述关系。但是，由于图像中的语义信息过于丰富，这样随机选取的复述句对的语义相似度会很低。本文对多模态复述的定义是描述图像中相同物体不同形式的表达。我们重新制作测试集数据，根据多模态复述的定义，在 MSCOCO 数据集上对原句和参考句进行重新筛选，从而在现有数据的基础上最大程度地确保它们描述图像中相关物体的一致性。为此，我们设计了制作描述图像中相同物体数据集的方法。

以往文本复述工作在 MSCOCO 数据集上对复述句对的选取方式有可能导致语义差异过大，并且不能保证原句与复述句是对相同的实体进行描述。与以往的方法不同，我们首先定义相似度评价函数，用于筛选描述图像中相同实体的复述句对。分别计算句对之间的 TF-IDF 相似度[16]、BLEU-1 值[17]和 Word2Vec 相似度[18]，对它们进行加权求和，得到一个综合评价后的相似度得分:

$$\begin{aligned}&\mathrm{Similarity}(S_1,S_2)=\alpha\times\mathrm{Score}_{\text{TF-IDF}}(S_1,S_2)+\\&\beta\times\mathrm{Score}_{\text{Word2Vec}}(S_1,S_2)+\gamma\times\mathrm{Score}_{\text{BLEU}}(S_1,S_2),\end{aligned}\tag{9}$$

其中权重设置为 α=0.4，β=0.4，γ=0.2。然后，我们对一张图片的 5 条描述计算两两之间的相似度并降序排序，从高到低依次选取相似度高于阈值 0.65 且不包含已被选择的句对，作为筛选后的复述句对。

经过对 MSCOCO 测试集的筛选，得到 4192 对复述句对，作为筛选后的测试集，对本文模型进行评测。图 2 是筛选前后的一些示例，可以看到用随机选取的方式选择出的复述句对会在语义上产生偏离，所描述的视觉概念也不相同。例如，左侧例子

输入句:
many kites are in the sky at the beach
参考句:
people out at the beach flying kites
重新筛选后
输入句:
some people are flying kites at the beach
参考句:
people out at the beach flying kites

输入句:
a man instructing a group of kids on a soccer field
参考句:
a group of young children standing around a field
重新筛选后
输入句:
a man instructing a group of kids on a soccer field
参考句:
a soccer coach is instructing the children on the field

图 2 测试集构建示例

Fig. 2 Examples of test set construction

中输入句着重围绕“kites”, “sky”和“beach”这 3 个实体以及它们之间的关系进行描述, 但是参考句着重描述的是“people”, “beach”和“kites”这 3 个实体以及它们之间的关系。经过重新筛选, 输入句和参考句都围绕“people”, “kites”, “beach”实体及其关系描述。可以看到, 筛选之后的复述句对描述的视觉概念进一步重合, 语义保持度更高, 符合我们对多模态复述的关于相同物体不同表达的定义, 体现出对数据集重新筛选的有效性和必要性。

3.2 结果分析

在测试阶段, 使用 BLUE, ROUGE[19], BERTScore[20]和 Sentence-BERT[21]作为评价指标, 比较结果如表 1 所示。

在本文的测试集上, 我们对比评测 4 种代表性的公开发表的文本复述生成模型, 描述如下: 1) Pegasus 模型[22]和 T5 模型①是在大规模复述语料上进行预训练的复述生成模型; 2) Residual-LSTM 模型[23]是一种基于编码解码框架的复述生成模型, 其编码器和解码器在 LSTM 神经网络的基础上加入残差连接, 我们对该模型进行复现, 将其用于评测; 3) Transformer 模型[24]使用全注意力结构在编码器和解码器中替代 LSTM 神经网络, 在复述等端到端任

表 1 和已有基于神经网络的复述生成系统的性能比较结果

Table 1 Experiment results on paraphrase generation compared with previous models

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-1	Rouge-2	Rouge-L	BERTScore	Sentence-BERT
Pegasus	27.80	16.88	10.46	6.53	46.81	20.13	43.06	91.88	67.75
T5	45.26	30.14	20.64	14.47	51.41	23.16	46.38	92.40	71.20
Residual-LSTM	46.43	31.55	22.15	16.05	47.21	21.86	43.32	91.71	60.77
Transformer	47.68	32.73	23.04	16.96	48.10	22.67	44.38	91.80	61.31
MPGM (本文)	58.79↑	43.49↑	32.12↑	23.79↑	61.41↑	33.91↑	61.32↑	93.32↑	74.59↑

① https://github.com/ceshine/finetuning-t5

务中是目前表现最好的模型，我们复现该模型，将其用于评测。

从表 1 可以看出，本文模型在全部 9 个评测指标上均有显著提升；其中 BLEU-1 到 BLEU-4 的评分方面，本文模型的性能比相关指标最优的模型分别高出 11.11%, 10.76%, 9.08%和 6.83%。在 Rouge-1，Rouge-2 和 Rouge-L 上，本文模型的性能比相关指标最优的模型分别高出 10%, 10.75%和 14.94%。因此，本文模型生成的复述与人工给出的参考句更加相似，此外，BLEU 和 Rouge 等基于 N-gram 的分数更高，只能表明生成的复述和参考句重叠程度高，并不能完全体现复述句的质量。之前的研究表明，BLEU 不能完美地评价文本生成任务[25]。我们同样认为，BLEU 对于复述生成任务来说，不是一个完全合理的度量指标。复述在本质上是高度多样化的，但在测试集中只存在一个参考句用于评价生成句。例如，句子“我能做些什么来克服焦虑”与人工给出的参考句“我做什么来减少我的焦虑”，如果模型生

输入句:
a cat laying next to a stainless steel bowl

参考句:
cat laying down with a bowl in front of it

文本复述生成模型:
a cat is laying down

本文模型:
a black cat is sitting beside a metal bowl

输入句:
a man on a surf board rides the waves

参考句:
a man is in the water and riding a surfboard

文本复述生成模型:
a man is riding a surf board

本文模型:
a person on a surfboard riding a wave in a wet suit

输入句:
a woman throws a tennis ball up in the air

参考句:
a woman standing on a tennis court holding a tennis racquet

文本复述生成模型:
a woman throws a tennis ball

本文模型:
a woman is playing tennis during a tennis match

输入句:
there are sheep standing in the grass near a fence

参考句:
several sheep are foraging in an open field

文本复述生成模型:
there are animals in the grass

本文模型:
many sheep grazing in a field near a fence

图 3　结果对比示例

Fig. 3　Examples of comparison results

成相应的复述为“我如何克服焦虑”或“克服焦虑的最好方法是什么”，这样的复述句在BLEU等相关指标的分数就会很低，但却是正确且质量好的复述。因此，为了全面地评价模型的性能，我们还在针对语义的评价指标BERTScore和Sentence-BERT上进行比较。我们认为，在BLEU和Rouge等基于N-gram的评价指标以及针对语义的评价指标BERT-Score和Sentence-BERT上同时表现出色的句子才是正确且高质量的复述句。

在语义评价指标方面，本文模型的BERTScore和Sentence-BERT分数也比相关指标最优的模型分别高出0.92%和3.39%，表明本文模型在语义方面比其他模型表现得更加出色。因此，本文模型不仅产生与人工给出的参考句更加相似的复述句，而且这些复述句具有更高的语义保持度。

3.3 实例分析

图3展示MSCOCO数据集输入句子的不同模型的输出。选择T5模型的输出作为文本复述模型的生成示例，与本文模型进行对比。结果表明，一方面，文本复述模型生成的语句比输入句子短，丢失大量的信息，语义保留性差；另一方面，文本复述模型生成的句子中名词基本上不变，多样性较差。与文本复述模型相比，本文模型通过对图像信息的利用，不仅很好地保持了复述的语义，而且生成的复述句具有较好的多样性。

4 结语

本文首次提出基于多模态的复述生成任务，用于在多模态场景下，解决传统模型在生成复述过程中因无法利用客观图像信息而导致部分语义丢失或产生歧义的问题。本文探索图像信息在复述生成任务上的利用，进而提出引入图像信息的多模态复述生成模型(MPG)。在MPG中，我们利用抽象场景图关联句子相关物体的图像特征。同时，使用基于关系图卷积神经网络建模抽象场景图中结点的上下文信息，在解码过程中引入基于图的注意力机制，提取图像特征，生成复述句。实验结果显示，与传统的文本复述模型相比，本文模型显著地增强了复述句与原句的语义保持度，同时能够带来表达形式多样的复述句，提高了复述生成的质量，体现出引入图像信息用于复述生成的有效性和必要性。

未来的研究工作中，我们需要探索利用物体间的空间位置关系等文本信息很难学习到的、图像中特有的信息，进而对相关的图像信息生成不同观察角度的表达。同时，在本工作的基础上补充消融实验，进一步证明引入图像信息的必要性以及模型各部分建模的有效性。

参考文献

[1] Dong L, Mallinson J, Reddy S, et al. Learning to paraphrase for question answering // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, 2017: 875–886

[2] Zhou Z, Sperber M, Waibel A. Paraphrases as Foreign Languages in Multilingual Neural Machine Translation // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, 2019: 113–122

[3] Zhao S, Rui M, He D, et al. Integrating transformer and paraphrase rules for sentence simplification // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). Brussels, 2018: 3164–3173

[4] Lin Z, Li Z, Ding N, et al. Integrating linguistic knowledge to sentence paraphrase generation. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8368–8375

[5] Lin Z, Wan X. Pushing paraphrase away from original sentence: a multi-round paraphrase generation approach [C/OL] // Findings of the Association for Computational Linguistics: ACL/IJCNLP. 2021 [2021–08–10]. https://aclanthology.org/2021.findings-acl.135.pdf

[6] Chen S, Jin Q, Wang P, et al. Say as you wish: fine-grained control of image caption generation with abstract scene graphs // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 9959–9968

[7] Chu C, Otani M, Nakashima Y. iParaphrasing: extracting visually grounded paraphrases via an image // Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe, 2018: 3479–3492

[8] Liu L, Tang J, Wan X, et al. Generating diverse and descriptive image captions using visual paraphrases // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 4240–4249

[9] Yin Y, Meng F, Su J, et al. A novel graph-based multi-modal fusion encoder for neural machine translation [C/OL] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). 2020 [2020–10–23]. https://aclanthology.org/2020.acl-main.273.pdf

[10] Hirasawa T, Yamagishi H, Matsumura Y, et al. Multimodal machine translation with embedding prediction // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies: Student Research Workshop. Minneapolis, 2019: 86–91

[11] Lee J, Cho K, Weston J, et al. Emergent translation in multi-agent communication [C/OL] // 6th International Conference on Learning Representations (ICLR). Vancouver, 2018 [2020–10–15]. https://openreview.net/pdf?id=H1vEXaxA-

[12] Schuster S, Krishna R, Chang A, et al. Generating semantically precise scene graphs from textual descriptions for improved image retrieval // Proceedings of the Fourth Workshop on Vision and Language. Lisbon, 2015: 70–80

[13] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137–1149

[14] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks // The Semantic Web—15th International Conference. Heraklion, 2018: 593–607

[15] Jia D, Wei D, Socher R, et al. ImageNet: a large-scale hierarchical image database // Computer Society Conference on Computer Vision and Pattern Recognition. Miami, 2009: 248–255

[16] Salton G, Yu C T. On the construction of effective vocabularies for information retrieval. Acm Sigplan Notices, 1973, 10(1): 48–60

[17] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002: 311–318

[18] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [C/OL] // 1st International Conference on Learning Representations. Scottsdale, 2013 [2020–09–20]. https://arxiv.org/pdf/1301.3781.pdf

[19] Lin C Y. Rouge: a package for automatic evaluation of summaries // Text Summarization Branches Out. Barcelona, 2004: 74–81

[20] Zhang T, Kishore V, Wu F, et al. Bertscore: Evaluating text generation with BERT [C/OL] // 8th International Conference on Learning Representations. Addis Ababa, 2020 [2021–03–15]. https://openreview.net/forum?id=SkeHuCVFDr

[21] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, 2019: 3980–3990

[22] Zhang J, Zhao Y, Saleh M, et al. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization // Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 11328–11339

[23] Prakash A, Hasan S A, Lee K, et al. Neural paraphrase generation with stacked residual LSTM networks // 26th International Conference on Computational Linguistics. Osaka, 2016: 2923–2934

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, 2017: 5998–6008

[25] Sulem E, Abend O, Rappoport A. BLEU is not suitable for the evaluation of text simplification // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, 2018: 738–744