

# 基于知识增强的输电线路场景图像描述

于平平<sup>1</sup>, 周晶晶<sup>1</sup>, 张立岩<sup>1</sup>, 苏鹤<sup>2</sup>

(1. 河北科技大学信息科学与工程学院, 河北石家庄 050018; 2. 河北工业大学电气工程学院, 天津 300131)

**摘要:** 针对输电线路场景中缺乏对潜在风险的刻画以及缺少领域知识支持等问题, 提出一种知识图谱增强的跨模态图像描述模型。以电力塔、树木、车辆、绝缘子等4种常见的场景目标物为基础, 通过人工标注和图像增强方法构建图像—英文描述数据集。在编码阶段, 基于该数据集构建特定领域知识图谱, 以增强 CLIP 跨模态模型的视觉特征提取能力。为增强模态间的交互能力, 设计轻量化 LLM-Adapter 模块, 通过门控交叉注意力机制动态融合视觉与文本特征。将获取到的语言感知视觉特征作为前缀输入到大语言模型中, 生成最终描述语句。实验结果表明, 在自制数据集和 MS COCO 2014 公开数据集上, 所提模型的 BLEU-4、CIDEr、SPICE 等评估指标相较对照模型均有所提升, 其中 BLEU-4 值达到 0.45, 生成的描述文本更连贯、内容更丰富。

**关键词:** 图像描述; 跨模态; 门控交叉注意力; 大语言模型; 知识图谱

DOI: 10.11907/rjdk.241742

开放科学(资源服务)标识码(OSID):

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)009-0206-07



## Image Description of Transmission Line Scenes Based on Knowledge Enhancement

YU Pingping<sup>1</sup>, ZHOU Jingjing<sup>1</sup>, ZHANG Liyan<sup>1</sup>, SU He<sup>2</sup>

(1. College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China;

2. College of Electrical Engineering, Hebei University of Technology, Tianjin 300131, China)

**Abstract:** A knowledge graph enhanced cross modal image description model is proposed to address the lack of characterization of potential risks and domain knowledge support in transmission line scenarios. Based on four common scene objects including power towers, trees, vehicles, and insulators, an image English description dataset is constructed through manual annotation and image enhancement methods. During the encoding phase, a domain specific knowledge graph is constructed based on this dataset to enhance the visual feature extraction capability of the CLIP cross modal model. To enhance the interaction ability between modalities, a lightweight LLM Adapter module is designed to dynamically fuse visual and textual features through a gated cross attention mechanism. Input the obtained language perception visual features as prefixes into a large language model to generate the final descriptive statement. The experimental results show that the proposed model has improved evaluation indicators such as BLEU-4, CIDEr, and SPICE compared to the control model on both the self-made dataset and the MS COCO 2014 public dataset. The BLEU-4 value reached 0.45, and the generated descriptive text was more coherent and rich in content.

**Key Words:** image description; cross-modal; gated cross-attention; large language model; knowledge graph

## 0 引言

输电线路是电力系统中的重要基础设施, 其安全稳定

性直接影响着整个电网的可靠性。然而, 输电线路沿线环境复杂多变, 树木侵扰、车辆靠近、绝缘子缺陷等问题时有发生, 这些因素常常导致线路故障和安全隐患<sup>[1]</sup>。传统的人工巡视方式效率低下, 难以对沿线各种潜在风险进行全

收稿日期: 2024-08-27

扫描二维码阅读全文:

基金项目: 河北省教育厅青年基金项目(QN2024052)

**作者简介:** 于平平(1984-), 女, 博士, 河北科技大学信息科学与工程学院副教授, 研究方向为数字图像处理、计算机视觉和人工智能等; 周晶晶(1999-), 女, 河北科技大学信息科学与工程学院硕士研究生, 研究方向为图像处理、计算机视觉、自然语言处理; 张立岩(1970-), 女, 硕士, 河北科技大学信息科学与工程学院副教授, 研究方向为 Web 应用程序开发技术和软件开发; 苏鹤(1993-), 男, 河北工业大学电气工程学院博士研究生, 研究方向为电工装备可靠性理论及应用。本文通讯作者: 张立岩。



面监控与准确描述。近年来,随着计算机视觉和自然语言处理领域的迅猛发展,图像描述技术备受关注。将该技术应用于输电线路场景中,可根据输电线路沿线环境图像自动生成针对电力塔、树木、车辆、绝缘子等目标物的文本描述,为输电线路的安全评估提供数据支持。

目前,基于深度学习的图像描述模型成为主流<sup>[2]</sup>。其中,编码—解码架构是一种典型的依赖深度学习的图像描述方法。例如,Vinyals等<sup>[3]</sup>提出利用卷积神经网络(Convolutional Neural Networks, CNN)作为编码器提取图像特征,利用长短时记忆网络(Long Short-Term Memory, LSTM)作为解码器还原原始图像信息的CNN-LSTM模型。Rennie等<sup>[4]</sup>在CNN-LSTM模型的基础上加入双重注意力机制和自我评分训练提取图像的底层视觉特征,但生成的文本内容未考虑语言序列的层级结构,表达的语言信息不够丰富<sup>[5]</sup>。Mokady等<sup>[6]</sup>提出一种基于视觉—语言预训练的CLIPCap模型,通过使用对比学习的预训练模型实现视觉与文本的对齐。为提高文本生成质量,Peng等<sup>[7]</sup>加入文本提示以增强模型对图像、文本等模态的感知与理解。

尽管现有图像描述模型在图像描述任务中取得了显

著效果,但在特定领域的知识整合和跨模态间交互方面仍存在局限性。为解决上述问题,本文提出一种基于知识图谱增强的跨模态图像描述模型。针对特定场景领域知识不足的问题,利用输电线路知识图谱增强视觉编码提高模型的理解能力;同时设计LLM-Adapter模块,增强视觉—语言特征的跨模态交互能力。将融合后的语言感知视觉特征作为前缀输入到大型语言模型(Large Language Model, LLM)的解码器中,最终生成准确、流畅的输电线路场景描述文本。

## 1 模型构建

基于知识图谱增强的跨模态输电线路场景图像描述模型结构见图1。该模型采用编码—解码框架,在图像编码阶段采用知识图谱增强视觉编码模块提取图像视觉特征,通过LLM-Adapter模块实现视觉—语言特征模态间的交互融合。随后将融合后的语言感知视觉特征向量作为前缀输入到LLM的解码器中,通过自回归的方式逐步生成准确、流畅的描述文本。

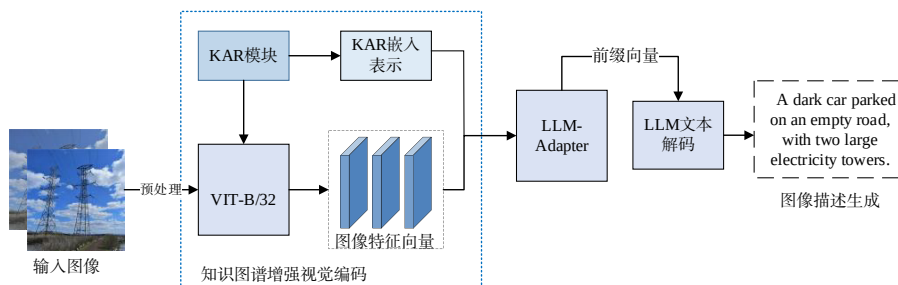


Fig. 1 Model structure

图1 模型结构

### 1.1 知识图谱增强视觉编码

#### 1.1.1 CLIP 编码器

早期图像描述模型主要采用CNN类架构,如VGGNet、ResNet等,作为视觉特征提取器。尽管这些模型在局部特征提取方面表现出色,但是其主要针对单一模态的图像数据进行处理,缺乏对文本信息的理解能力,导致视觉特征与语义文本之间的映射效果不佳。为此,本文选用CLIP(Contrastive Language-Image Pre-training)视觉—文本联合预训练模型进行视觉特征提取<sup>[8]</sup>。结构如图2所示。

在特征提取过程中,ViT-B/32将输入图像平铺为 $32 \times 32$ 的图像块(Patch)序列 $X$ ,并通过线性投影层将Patch转化成固定长度的特征向量序列。每个Patch的特征向量序列与其对应的视觉位置编码相结合,作为Transformer编码器的输入。该Transformer编码器通过多个自注意力机制动态评估各个Patch的重要性,从而有效捕捉图像的全局上下文特征。最终经过编码器的处理生成视觉向量 $V$ ,为生成准确的图像描述提供了重要特征信息。该过程用公式描述为:

$$V = \text{CLIP}_{\text{ViT}}(X) \quad (1)$$

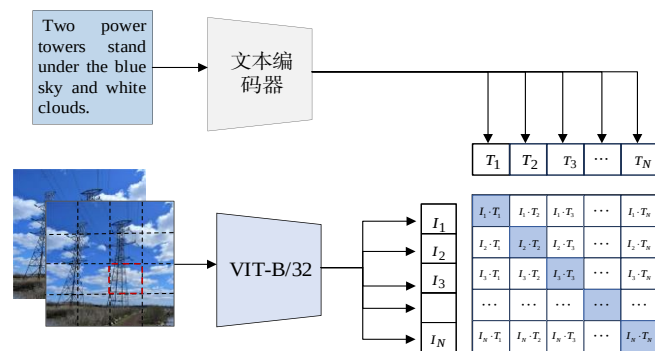


Fig. 2 CLIP structure

图2 CLIP结构

#### 1.1.2 KAR

CLIP模型虽然能够建立图像与文本之间的关联,但其主要依赖于大规模数据的统计相关性,缺乏对输电线路领域知识的深入理解,不能为图像描述提供足够的信息<sup>[9]</sup>。为解决这一问题,本文提出KAR(Knowledge-Augmented Representation)模块,实现结构化知识与视觉数据的融合,提升模型的语义理解能力和描述生成质量。该模块的核心在于利用输电线路领域知识图谱增强视觉表示<sup>[10]</sup>。

本文从输电线路数据集中抽取9种实体类型、相应属性以及7种关系类型,构建输电线路领域的知识图谱。其中,实体类型包括电力塔、绝缘子等;属性包括高度、颜色、状态等;关系类型包括悬挂、遮挡、靠近、上空等。利用NetworkX库构建图结构,节点表示实体,边表示关系。具体如图3所示。

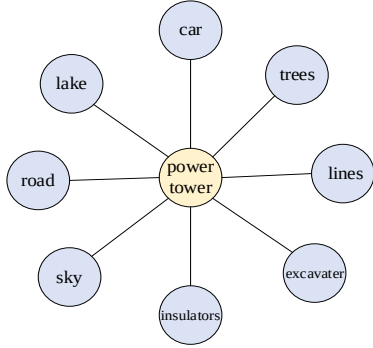


Fig. 3 Example of knowledge graph

图3 知识图谱示例

为了将知识图谱信息编码到模型中,本文采用图卷积神经网络(Graph Convolutional Networks, GCN)进行编码<sup>[11]</sup>。表示为:

$$K = GCN(G) \quad (2)$$

式中: $G$ 为图结构数据; $K$ 为知识图谱的向量表示。

通过全连接层,将 $K$ 扩展至与视觉特征向量 $V$ 相同的维度,并进行特征融合。表示为:

$$X = \alpha * FC(K) + (1 - \alpha) * V \quad (3)$$

式中: $FC$ 为全连接层; $V$ 为视觉特征; $\alpha$ 为融合权重参数。得到的 $X$ 张量不仅继承了CLIP模型强大的视觉—文本跨模态理解能力,而且具有KAR模块结构化的语义知识,进一步提升了模型的推理能力与语义理解深度。

## 1.2 LLM-Adapter模块

为了将CLIP的视觉—语言对齐能力与LLM的纯文本生成能力深度融合,使LLM适应下游任务,同时节约微调LLM的训练成本,本文设计轻量化LLM-Adapter模块。该模块主要由多层感知机(Multi-Layer Perceptron, MLP)映射网络和门控交叉注意力(Gated Cross Attention, GCA)模块组成<sup>[12]</sup>。

### 1.2.1 映射网络

为了将视觉特征空间调整到与文本解码器相匹配的维度空间,本文选取MLP作为映射网络。该映射网络将512维的图像特征向量 $V \in \mathbb{R}^{512}$ 转换为768维向量 $V' \in \mathbb{R}^{768}$ 。映射过程可表示为:

$$V' = f(W_2 * \text{ReLU}(W_1 * V + b_1) + b_2) \quad (4)$$

式中: $W_1 \in \mathbb{R}^{m \times 512}$ 和 $W_2 \in \mathbb{R}^{768 \times m}$ 为权重矩阵; $b_1, b_2$ 为偏置向量; $m$ 为隐藏层维度; $\text{ReLU}$ 为激活函数; $f$ 为输出层的归一化函数。该映射网络实现了视觉与文本两个模态间的有效对齐。

### 1.2.2 GCA

为了增强映射过程中视觉与文本两模态间的交互能力,本文引入GCA模块,结构如图4所示。该模块在传统交叉注意力机制的基础上通过额外引入门控因子 $g$ ,使模型能够有选择性地调整注意力的分配,灵活地关注视觉与文本之间的关键信息,从而提高跨模态交互效果。

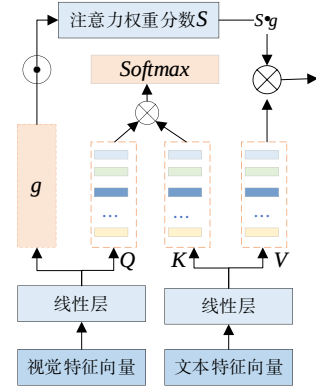


Fig. 4 GCA module structure

图4 GCA模块架构

文本特征 $T = \{t_1, t_2, \dots, t_k\}$ 在线性层被映射为GCA机制的键值向量 $K$ 和值向量 $V$ ;经过知识图谱增强的视觉特征向量 $X$ 则被映射为查询向量 $Q$ ,同时获得门控因子 $g$ 。注意力权重分数的计算公式为:

$$S = \text{softmax}(QK^T / \sqrt{d_k}) \quad (5)$$

式中: $S$ 为注意力权重矩阵。

采用门控因子 $g$ 对注意力分数进行自适应调整。计算公式为:

$$S' = S \cdot g \quad (6)$$

最终得到文本与知识图谱联合增强的视觉特征表示 $a$ 。表示为:

$$a = S' \cdot V \quad (7)$$

$a$ 将作为前缀输入到LLM中,逐步生成文本描述。通过引入门控因子,模型能够聚焦于真正相关的视觉信息,同时抑制与视觉特征和文本描述不相关或冗余的信息,生成更准确、更相关的图像描述。

## 1.3 LLM文本解码

在图像描述任务中,编码器负责提取图像视觉特征,解码器则将这些特征转化为语义丰富的文本描述。传统的LSTM网络虽然能够捕捉长期依赖关系,但是在生成文本时仍存在一些局限性,例如生成的描述往往缺乏复杂的语法结构和丰富的语义层次,导致输出文本比较单调或简单。为克服这一局限性,本文采用自回归训练的LLM模型GPT-2作为解码器。尽管CLIP和GPT-2模型均具有强大的文本表征能力,但是由于二者在预训练阶段并非联合优化,其潜在空间独立,可能会影响跨模态任务的性能。由于微调LLM会大大增加可训练参数的数量,本文选择在训练过程中冻结CLIP和GPT-2的权重参数,通过训练LLM-

Adapter对语言模型进行微调<sup>[13]</sup>。在训练过程中采用交叉熵损失函数,使用Adam优化器,初始学习率设置为1e-4, batch size为20,训练50个epoch。

2 实验方法与结果分析

2.1 实验参数设置

使用PyTorch 1.11深度学习框架对模型进行训练,采用Adam优化器进行参数调整,预设学习率为0.000 02,采用warm-up策略动态调整学习率。在文本生成阶段,采用beam search解码算法在每个时间步保留5个最可能的候选序列,以生成全局最优的输出序列。同时,采用Top-p方法采样生成准确且多样化的描述文本。具体超参数设置如表1所示。

Table 1 Experimental hyperparameter settings  
表 1 实验超参数设置

| 超参数设置                      | 取值       |
|----------------------------|----------|
| 迭代轮次(epoch)                | 500      |
| 批大小(batch size)            | 40       |
| 学习率(learning rate)         | 0.000 02 |
| 预热步数                       | 5 000    |
| 前缀长度(prefix_size)          | 512      |
| temperature                | 1.0      |
| 损失函数                       | 交叉熵损失函数  |
| 优化器(optimizer)             | Adam     |
| Probability value of Top-p | 0.8      |

2.2 数据集

通过收集图像数据自行构建输电线路图像描述数据集,涵盖车辆、绝缘子、树木和电力塔等典型场景元素。为增强模型的泛化能力和鲁棒性,减少过拟合风险,采用随机裁剪、翻转、旋转等数据增强技术将原始收集的2 000张图像数据扩充至5 000张。最终数据集按照8:1:1的比例分为训练集、验证集和测试集。

参考MS COCO数据集格式,首先通过Python脚本自动读取每张图片的ID;然后进行人工标注,为每张图片生成5句不超过20词的描述语句<sup>[13]</sup>。在标注过程中对数据进行语法检查与人工审核,确保其一致性和准确性。输电线路数据集描述示例如表2所示。其中,图像a、c、e分别代表车辆靠近、绝缘子损坏、树木侵扰3种典型的输电线路潜在异常情况;图像b、d则为输电线路正常场景。

2.3 图像描述评价指标

本文主要采用BLEU-4、METEOR、ROUGE、CIDEr和SPICE 5种图像描述评价指标对模型进行评估。其中,BLEU-4通过计算生成文本与参考文本之间n-gram的精准率来评价生成文本的准确性<sup>[14]</sup>;METEOR在BLEU的基础上引入召回率,使生成文本在保证精确度的同时更加符合人类语言的表达习惯<sup>[15]</sup>;ROUGE通过对生成文本与参考文本的相同部分进行比较来计算句子中单词的重复率和排列顺序的相似性<sup>[16]</sup>;CIDEr同样通过比较生成文本与

Table 2 Example description of transmission line dataset  
表 2 输电线路数据集描述示例

| 序号 | 图像                                                                                   | 文本描述                                                                                    |
|----|--------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| a  |    | 1. There is a moving vehicle on the road.                                               |
|    |                                                                                      | 2. A picture of a moving vehicle.                                                       |
|    |                                                                                      | 3. There is a white car on the road.                                                    |
|    |                                                                                      | 4. A picture with trees and vehicle.                                                    |
|    |                                                                                      | 5. Above the vehicle, there are power lines.                                            |
| b  |    | 1. There are normal insulators on the power tower.                                      |
|    |                                                                                      | 2. The two insulators across the tower are normal.                                      |
|    |                                                                                      | 3. There is an electricity tower with two insulator strings hanging on the green grass. |
|    |                                                                                      | 4. Standing in a field is a power tower adorned with insulators.                        |
|    |                                                                                      | 5. There stands a power tower with insulators.                                          |
| c  |   | 1. Damaged grey insulators are vertically strung on a power tower.                      |
|    |                                                                                      | 2. This gray insulator string is damaged.                                               |
|    |                                                                                      | 3. This hanging gray string of insulators is damaged.                                   |
|    |                                                                                      | 4. The suspended string of gray insulators has been damaged.                            |
|    |                                                                                      | 5. This hanging gray disc shaped insulator string is broken.                            |
| d  |  | 1. There are numerous trees located beneath the power tower.                            |
|    |                                                                                      | 2. A multitude of trees can be found underneath the power tower.                        |
|    |                                                                                      | 3. The power tower is surrounded by a large number of trees.                            |
|    |                                                                                      | 4. There is a profusion of trees beneath the power tower.                               |
|    |                                                                                      | 5. A plethora of trees exists under the power tower.                                    |
| e  |  | 1. The branches of many trees overshadow the power towers.                              |
|    |                                                                                      | 2. Overgrowth has resulted in numerous tree branches obscuring transmission towers.     |
|    |                                                                                      | 3. The branches of a large number of trees shade that covers the transmission towers.   |
|    |                                                                                      | 4. Lots of tree branches overshadowing the transmission towers.                         |
|    |                                                                                      | 5. Numerous tree branches shading transmission towers.                                  |

参考文本之间的相似部分来实现对单词重复率和排列顺序相似度的评估<sup>[17]</sup>;SPICE通过将生成文本和参考文本解析成句法依赖树,并进一步映射成场景图来衡量句子相似

性<sup>[18]</sup>。以上指标的取值范围均为 $[0,1]$ ,数值越大表示模型性能越优秀。

## 2.4 实验结果与分析

### 2.4.1 比较实验

在自制输电线路数据集上对本文模型与 CNN-NIC<sup>[3]</sup>、CLIPCap<sup>[6]</sup>、CNet-NIC<sup>[11]</sup>、Transformer<sup>[19]</sup>、M<sup>2</sup>Transformer<sup>[20]</sup>、Up-Down<sup>[21]</sup>、RFNet<sup>[22]</sup>、RSTNet<sup>[23]</sup>、Oscar<sup>[24]</sup>、GCN-LSTM<sup>[25]</sup>等几种主流模型进行比较。其中,CNN-NIC是Google提出的经典图像描述模型,通过目标检测模型提取图像特征并输入到LSTM中生成图像描述;CLIPCap是采用对比预训练的图像描述模型;CNet-NIC模型是基于ConceptNet知识库增强的图像描述模型;Transformer是基于自注意力机制与交叉自注意力机制的图像描述模型;M<sup>2</sup>Transformer是采用双向Transformer编码器的图像描述模型;Up-Down和RFNet是通过注意力机制来关注空间区域的模型;RSTNet是基于视觉关系推理的神经网络图像描述模型;Oscar模型是将检测到的对象标签作为锚点的图像描述模型;GCN-LSTM模型是基于GCN与LSTM的图像描述模型。各模型在自制输电线路数据集上的性能指标比较结果见表3。可以看出,本文模型的5项评估指标均优于所有对照模型。其中,BLEU-4值达到0.450,较CLIPCap和GCN-LSTM分别提升了0.185和0.140,表明模型生成的文本与实际文本之间的差异很小;METEOR和ROUGE指标分别达到0.293和0.495,略优于CNet-NIC模型;CIDEr得分为0.973,略高于Oscar模型。以上结果表明,通过引入输电线路领域的知识图谱和GCA模块,本文

模型实现了目标信息与知识的有效对应,并生成了包含常识性背景知识的图像描述结果。

**Table 3 Comparison results of performance indicators of various models on self-made transmission line dataset**

**表3 各模型在自制输电线路数据集上的性能指标比较结果**

| 模型                         | BLEU-4       | CIDEr        | METEOR       | SPICE        | ROGUE        |
|----------------------------|--------------|--------------|--------------|--------------|--------------|
| CNN-NIC                    | 0.230        | 0.780        | 0.231        | 0.180        | 0.470        |
| CLIPCap                    | 0.265        | 0.932        | 0.246        | 0.200        | —            |
| Transformer                | 0.232        | 0.821        | 0.254        | 0.170        | 0.480        |
| M <sup>2</sup> Transformer | 0.246        | 0.928        | 0.271        | 0.236        | 0.474        |
| Up-Down                    | 0.250        | 0.917        | 0.230        | 0.230        | 0.466        |
| RFNet                      | 0.320        | 0.955        | 0.272        | 0.209        | 0.470        |
| RSTNet                     | 0.283        | 0.901        | 0.270        | 0.232        | 0.446        |
| Oscar                      | 0.420        | 0.972        | 0.285        | 0.245        | —            |
| GCN-LSTM                   | 0.310        | 0.955        | 0.270        | 0.220        | 0.445        |
| CNet-NIC                   | 0.340        | 0.915        | 0.234        | 0.230        | 0.450        |
| <b>本文模型</b>                | <b>0.450</b> | <b>0.973</b> | <b>0.293</b> | <b>0.251</b> | <b>0.495</b> |

由于图像描述任务具有特殊性,定量评价指标数据并不能完全反映生成文本的真实质量,还需要通过人工手段对部分案例进行评估,以考察模型生成文本的准确性。选取本文模型与CLIPCap等代表性模型进行比较,部分实验结果如表4所示。可以看出,本文模型能够更好地识别图像中的实体、属性关系等。例如,在例3中,Transformer模型将“绝缘子”错误地描述成了“火车”;Up-Down模型描述出的绝缘子数量存在问题;CLIPCap模型虽然准确识别出“绝缘子”,但是错误地引入图像中并不存在的“网”这一元素,导致描述出现偏差。相比之下,本文模型不仅准确识别出场景中的电力塔和绝缘子,而且正确表达了二者之间的空间位置关系。

**Table 4 Comparison of image text generation results between the proposed model and other models**

**表4 本文模型与其他模型图像文本生成结果比较**

| 序号 | 图像                                                                                  | 文本生成结果                                                                                                                                                                                                                                                                                                     |
|----|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 例1 |  | <p><b>Transformer:</b> In the sky, there are towering power structures.</p> <p><b>Up-Down:</b> There are power towers in the sky.</p> <p><b>CLIPCap:</b> The towers hanging many insulators.</p> <p><b>Ours:</b> A power tower adorned with insulators stands tall in the sky.</p>                         |
| 例2 |  | <p><b>Transformer:</b> A power tower stands tall nearby.</p> <p><b>Up-Down:</b> A power tower stands in the blue sky.</p> <p><b>CLIPCap:</b> In the sky, a towering power tower stands proudly.</p> <p><b>Ours:</b> A red car is parked next to a power tower in the distant skyline.</p>                  |
| 例3 |  | <p><b>Transformer:</b> This is a train on tower.</p> <p><b>Up-Down:</b> There are two insulators on the tower.</p> <p><b>CLIPCap:</b> There is an insulator and net is the tower.</p> <p><b>Ours:</b> The normal insulator horizontally on the tower.</p>                                                  |
| 例4 |  | <p><b>Transformer:</b> There are many trees under sky.</p> <p><b>Up-Down:</b> There's a lot of power lines in the sky.</p> <p><b>CLIPCap:</b> This is a truck under the transmission line.</p> <p><b>Ours:</b> Above the construction vehicle, there are numerous power transmission lines in the sky.</p> |

2.4.2 消融实验

(1) KAR 模块的影响。为验证 KAR 模块的有效性,在输电线路数据集上进行定量分析实验。消融实验结果如图 5 所示。其中, KAR 表示添加 KAR 模块的实验结果, -KAR 表示无 KAR 模块的实验结果。可以看出,添加 KAR 模块后,模型在所有评估指标上均有所提高。

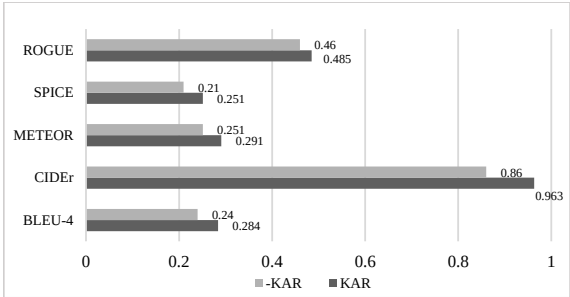


Fig. 5 Ablation experimental results of the KAR module  
图 5 KAR 模块消融实验结果

选取两张具有代表性的图像对有无 KAR 模块时模型生成的文本进行定性分析。结果如图 6 所示。其中 a 为添加 KAR 模块后生成的描述, b 为未添加 KAR 模块时生成的描述。可以看出, a 描述不仅准确涵盖了绝缘子、电力塔和工程车等目标物,而且包含更多细节信息,如绝缘子串的损坏情况、工程车与电力塔的空间位置关系等。而 b 描述则忽略了与场景相关的事物,过于关注不相关的内容,导致描述内容单薄、信息量不足。实验结果证实了 KAR 模块在提高图像描述质量方面具有一定作用。

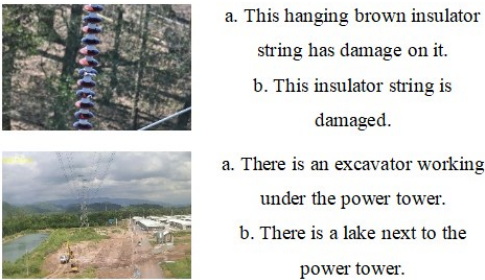


Fig. 6 Qualitative analysis results of KAR module  
图 6 KAR 模块定性分析结果

(2) GCA 模块的影响。图 7 为 GCA 模块的消融实验结果。其中, GCA 表示添加 GCA 模块的实验结果, -GCA 表示未添加 GCA 模块的实验结果。可以看出,添加 GCA 模块后,模型在所有评估指标上均有所提高。这是由于 GCA 模块能够增强模型的跨模态交互能力,有助于生成与人类描述更接近的句子。

为了更好地评估 GCA 模块对模型性能的影响,将文本生成过程中的注意力权重可视化。结果如图 8 所示。该示例描述了一辆工程车在电力塔和房屋附近作业的场景。从热力图可以看出,生成“truck”“driving”“power tower”和“house”这些与图像语义密切相关的词汇时,注意力权重分布较为集中;而对于语义不相关的“and”,注意力权重则分布较为分散,说明 GCA 模块能有效利用语义信息推断出

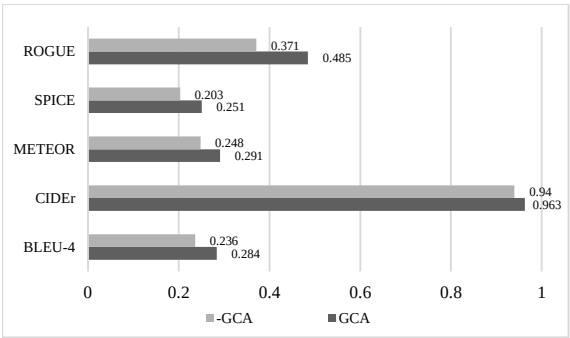


Fig. 7 Ablation experimental results of the GCA module  
图 7 GCA 模块消融实验结果

每个生成单词及其相关的注意力权重分布,从而提高生成文本的质量。

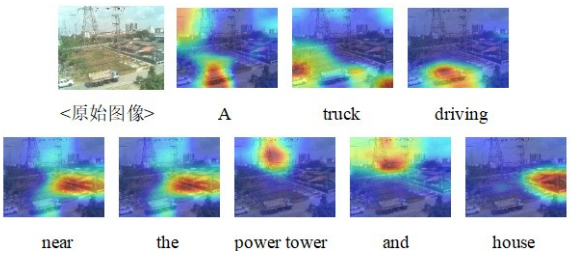


Fig. 8 Visualization of attention weights in text generation process  
图 8 文本生成过程中的注意力权重可视化

2.4.3 泛化性能实验

为验证本文模型的泛化性能,在 MS COCO caption 2014 Karpathy 公开数据集上与其他主流算法进行比较<sup>[26]</sup>。实验结果如表 5 所示。可以看出,本文模型的 BLEU-4 指标达到 0.430,相较次优模型 CNet-NIC 提高了 2.6%。在 CIDEr 指标上,本文模型略低于 CNet-NIC 模型,但仍然优于大多数对照模型。

Table 5 Comparison of the performance of different models on MS COCO dataset

| 表 5 在 MS COCO 数据集上不同模型性能比较 |        |        |       |       |
|----------------------------|--------|--------|-------|-------|
| 模型                         | BLEU-4 | METEOR | CIDEr | SPICE |
| NIC                        | 0.245  | 0.235  | —     | —     |
| CLIPCap                    | 0.320  | 0.270  | 1.08  | 0.200 |
| Up-Down                    | 0.363  | 0.277  | 1.20  | 0.216 |
| Oscar                      | 0.360  | 0.300  | 1.20  | 0.237 |
| CNet-NIC                   | 0.404  | 0.302  | 1.24  | —     |
| 本文模型                       | 0.430  | 0.305  | 1.21  | 0.219 |

表 6 为本文模型与其他模型在 MS COCO 数据集上的可视化结果。可以看出,本文模型生成的图像描述内容较为丰富,与对照模型中表现最佳的 CNet-NIC 模型生成的描述相似。

3 结语

本文提出一种基于知识图谱增强的跨模态图像描述模型,利用知识图谱增强视觉编码模块提取图像的视觉信

**Table 6 Visualization result of different models on the MS COCO dataset****表 6 不同模型在 MS COCO 数据集上的可视化结果**

| 图像                                                                                | 文本生成结果                                                                                                                                                                                                                    |
|-----------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | <b>CLIPCap:</b> A train on a road with trees and power lines<br><b>CNet-NIC:</b> Red train on a mountain with trees and power lines.<br><b>Ours:</b> A red train moving along a road surrounded by trees and power lines. |
|  | <b>CLIPCap:</b> Town road with traffic lights and cars.<br><b>CNet-NIC:</b> A street with traffic lights and buildings.<br><b>Ours:</b> A town street with traffic lights, parked car, and buildings..                    |

息,通过微调轻量化 LLM-Adapter 模块使 LLM 模型输出更加丰富的描述文本。实验结果表明,该模型在 BLEU-4、METEOR、ROUGE、CIDEr 和 SPICE 等 5 项评价指标上表现出色,生成的描述文本内容更加准确、丰富,但仍存在一些方面有待改进。首先,当前构建的数据集和知识图谱虽然提升了模型性能,但其规模和覆盖范围有限,未来将扩充数据集的规模和深度;其次,图像细粒度区域描述能力有待提升。未来将开展对密集图像描述的研究,并将优化后的模型部署到自建平台,实现技术成果的实际落地与应用。

#### 参考文献:

- [1] YANG R X, SHAO J, LUO Y, et al. Controllable image caption generation based on encoder-decoder for power construction scene[J]. Power System Technology, 2022, 46(7): 2572-2581.  
杨润霞,邵洁,罗岩,等. 基于编解码器的电力施工场景可控图像字幕生成[J]. 电网技术, 2022, 46(7): 2572-2581.
- [2] ZHAO Y Q, JIN Z, ZHANG F, et al. Deep-learning based image captioning: analysis and prospects[J]. Journal of Image and Graphics, 2023, 28(9): 2788-2816.  
赵永强,金芝,张峰,等. 深度学习图像描述方法分析与展望[J]. 中国图象图形学报, 2023, 28(9): 2788-2816.
- [3] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]//Boston: IEEE Conference on Computer Vision and Pattern Recognition, 2015.
- [4] RENNIE S J, MARCHERET E, MROUEH Y, et al. Selfcritical sequence training for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7008-7024.
- [5] YU Y W, SHI S C, WANG H J. Image caption based on Bert word vectors and ordered memory network[J]. Software Guide, 2023, 22(3): 125-133.  
俞艺文,施水才,王洪俊. 基于 Bert 词向量与有序记忆网络的图像描述[J]. 软件导刊, 2023, 22(3): 125-133.
- [6] MOKADY R, HERTZ A, BERMANO A H. Clipcap: clip prefix for image captioning[DB/OL]. <https://arxiv.org/abs/2111.09734>.
- [7] PENG Z, WANG W, DONG L, et al. Kosmos-2: grounding multimodal large language models to the world[DB/OL]. <https://arxiv.org/abs/2306.14824>.
- [8] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning, 2021: 8748-8763.
- [9] DUAN M S, XIAO L. Review of knowledge graph and its application in the field of food worm[J]. Software Guide, 2021, 20(8): 241-246.  
段梦诗,肖乐. 知识图谱技术综述及在粮虫领域的应用[J]. 软件导刊, 2021, 20(8): 241-246.
- [10] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494-514.
- [11] SANTISTEBAN S S, ATTITO S, AWAIS M, et al. Improved image captioning via knowledge graph-augmented models[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, 2024: 4290-4294.
- [12] ZHANG R, HAN J, ZHOU A, et al. LLaMA-Adapter: efficient fine-tuning of language models with zero-init attention[DB/OL]. <https://arxiv.org/abs/2303.16199>.
- [13] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[DB/OL]. <https://arxiv.org/abs/2101.00190>.
- [14] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.
- [15] DENKOWSKI M, LAVIE A. Meteor universal: language specific translation evaluation for any target language[C]//Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014: 376-380.
- [16] LIN C Y. Rouge: a package for automatic evaluation of summaries[C]//Text Summarization Branches Out, 2004: 74-81.
- [17] VEDANTAM R, LAWRENCE Z C, PARIKH D. Cider: consensus-based image description evaluation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4566-4575.
- [18] ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: semantic propositional image caption evaluation[C]//14th European Conference on Computer Vision, 2016: 382-398.
- [19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5998-6008.
- [20] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed memory transformer for image captioning[C]//Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10578-10587.
- [21] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6077-6086.
- [22] JIANG W, MA L, JIANG Y G, et al. Recurrent fusion network for image captioning[C]//Proceedings of the European Conference on Computer Vision, 2018: 499-515.
- [23] ZHANG X Y, SUN X S, LUO Y P, et al. Rstnet: captioning with adaptive attention on visual and non-visual words[C]//Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 15465-15474.
- [24] ZHAO S, LI L, PENG H, et al. Aligned visual semantic scene graph for image captioning[J]. Displays, 2022, 74: 102210.
- [25] YAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning[C]//Proceedings of the European Conference on Computer Vision, 2018: 684-699.
- [26] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]//Boston: 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.

(责任编辑:尹晨茹)