第 41 卷 第 22 期/2021 年 11 月/光学学报 研究论文

光学学报

基于注意力和强化学习的遥感图像描述方法

农元君，王俊杰*

中国海洋大学工程学院，山东 青岛 266100

摘要 针对当前遥感目标检测方法只能识别出遥感目标的类别及位置，无法生成与遥感图像内容相关文本描述的问题，提出了一种基于注意力和强化学习的遥感图像描述方法。首先，采用卷积神经网络构建编码器，提取遥感图像的特征。其次，利用长短期记忆网络搭建解码器，学习图像特征与文本语义特征间的映射关系。然后，引入注意力机制，增强模型对显著性特征的关注，减少无关背景特征的干扰。最后，采用强化学习策略，根据离散且不可微的评价指标直接对模型进行优化，消除暴露偏差及优化方向不一致的缺陷。在公开遥感图像描述数据集中的实验结果表明，本方法的检测精度较高，对密集小目标、雾气积聚、背景特征与目标特征相似等复杂环境下的遥感图像具有良好的描述性能。

关键词 遥感；图像描述；强化学习；注意力机制；编码-解码

中图分类号 TP753；TP183 **文献标志码** A **doi**：10.3788/AOS202141.2228001

Remote Sensing Image Caption Method Based on Attention and Reinforcement Learning

Nong Yuanjun, Wang Junjie*

College of Engineering, Ocean University of China, Qingdao, Shandong 266100, *China*

Abstract The current remote sensing object detection methods, only identifying the category and location of remote sensing objects, cannot generate text caption related to the contents of remote sensing images. A remote sensing image caption method based on attention and reinforcement learning is proposed in this paper to solve this problem. First, the convolution neural network is used to construct an encoder and thereby extract remote sensing image features. Secondly, a decoder is built through the long short-term memory network to learn the mapping relationships of the image features with text semantic features. Thirdly, the attention mechanism is introduced to enhance the attention of the model on salient features and reduce the interference of irrelevant background features. Finally, the reinforcement learning strategy is adopted to optimize the model directly according to the discrete and non-differentiable evaluation indexes and thus to eliminate the defects of exposure bias and inconsistent optimization directions. Experimental results of public data sets of remote sensing image caption show that the method achieves high detection accuracy and has good caption performance for remote sensing images in complex environments such as dense small targets, fog accumulation, and similar background and object features.

Key words remote sensing; image caption; reinforcement learning; attention mechanism; encode-decode

OCIS codes 280.4788; 100.3008; 110.2960

1 引 言

随着遥感技术的快速发展，高分辨率遥感图像包含的信息日益丰富，极大推动了遥感领域的应用研究。遥感图像中蕴涵机场、港口、船舰等重要信息，通过提取遥感图像特征并学习图像特征与文本

收稿日期：2021-04-30；**修回日期**：2021-05-20；**录用日期**：2021-06-03

基金项目：山东省重点研发计划(2019GHY112081)

通信作者：*wjj@ouc.edu.cn

语义特征间的映射关系，实现对遥感图像内容的解译与描述，在国防安全、土地监测、城市规划及抗灾减灾等军用和民用领域具有广泛的应用价值。如在国防安全中，通过提取和捕捉遥感图像中机场与船舰等重要信息，生成与遥感图像内容相关且语义通顺的文本描述，可为军事安全管理人员提供军事情报，辅助其快速做出决策并展开任务部署；在民用领域，生成的遥感图像文本描述可准确提供关于灾害评估、农田利用情况、植被覆盖及城市变迁等重要信息，为相关管理人员提供决策支持。因此，对遥感图像进行描述具有重要意义。

已有研究大多基于深度学习的目标检测算法对遥感图像中的目标进行识别和检测[1-5]，并取得了较高的检测精度。但目标检测方法无法生成与遥感图像内容相关的文本描述，存在低层视觉特征与高层语义特征间的语义鸿沟，无法实现对遥感图像的感知和理解，具有一定的局限性。与目标检测不同，图像描述方法结合了计算机视觉和自然语言处理，通过提取遥感图像中的目标区域特征、空间特征、环境特征及场景特征等丰富的图像特征信息，并捕捉和挖掘文本语句内部单词之间的句法语义特征，学习图像特征和文本语义特征之间的联系及映射关系，实现对遥感图像的解译，生成与遥感图像内容相关且语义通顺的文本描述。

目前关于图像描述的研究大多聚焦于自然场景，面向遥感场景的图像描述研究较少。赵佳琦等[6]提出了一种基于多尺度与注意力特征增强的遥感图像描述生成方法，实现了对遥感图像的描述。Shi 等[7]基于深度学习和卷积神经网络(CNN)提出了一种遥感图像描述方法。Qu 等[8]提出了一种深度多模态神经网络模型，可用于高分辨率遥感图像的文本描述。Lu 等[9]构建了一个公开的遥感图像描述数据集 RSICD，并采用多模态方法和注意力方法生成关于遥感图像内容的描述。上述研究虽然实现了对遥感图像的描述，但容易受到遥感图像背景复杂、噪声信息多、目标占比小的影响，导致生成的遥感图像描述结果准确性较低，无法满足复杂环境下的遥感图像描述需求。如背景颜色与遥感目标颜色相似会导致遥感目标难以辨别，云层、大气颗粒及雾会给遥感图像特征的提取带来极大的困难。

针对上述问题，本文提出了一种基于强化学习和注意力机制的遥感图像描述方法。采用基于 CNN 的编码器对遥感图像进行特征提取；利用长短期记忆(LSTM)网络搭建解码器，以捕捉语句内部单词之间的语义特征；引入注意力机制，增强网络对显著性特征的关注与捕获；采用强化学习策略[10]，根据 BLUE(Bilingual evaluation understudy)[11]等指标直接对模型进行优化，以提升训练效果。实验结果表明，本方法在公开遥感图像描述数据集上取得了较高的精度，且具有良好的遥感图像描述性能，在密集小目标、雾气积聚遮挡、背景特征与目标特征相似等复杂环境下可准确生成与遥感图像内容相关的文本描述。

2 基于注意力和强化学习的遥感图像描述模型

为了实现遥感图像描述，受文献[10]的启发，提出了一种基于注意力机制和强化学习的遥感图像描述模型，其网络架构如图 1 所示。其中，MLP 表示多层感知机。该模型由编码器和解码器组成，首先，采用具有强健空间感知能力的 CNN 构建编码器(Encoder)，以提取出遥感图像中的目标区域特征信息、空间信息、环境信息及场景信息等丰富的图像特征。然后，采用具有良好文本信息处理能力的 LSTM 网络搭建解码器(Decoder)，以挖掘文本语句内部单词之间的句法特征及单词位置编码。最后，通过动态选择图像特征并学习图像特征和文本语义特征之间的映射关系，实现对遥感图像的解译，生成

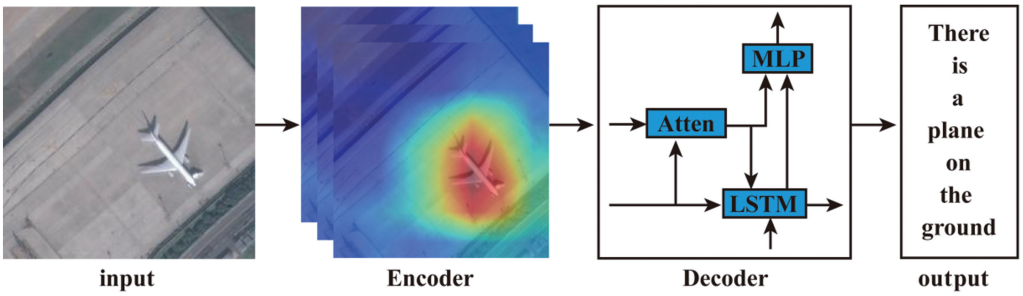

图 1 遥感图像描述模型的原理

Fig. 1 Principle of the remote sensing image caption model

与遥感图像内容相关且语义通顺的文本描述。为了增强模型对显著性特征的关注，提升描述结果的准确性，在解码器中引入注意力机制(Atten)，通过计算输入信息的重要性分布，协助模型在大量信息中快速捕捉到当前最需要的信息，减少关键信息的丢失。此外，为了消除暴露偏差及优化方向不一致的缺陷，在训练阶段引入强化学习策略[10]，根据BLUE等评价指标直接对模型进行优化，提升模型的训练效果。

2.1 编码器

遥感图像中蕴涵丰富的视觉特征信息(目标区域特征、类别特征、属性特征等)以及更深层次的语义信息(目标交互关系、场景信息、环境信息、空间信息等)，这些视觉特征和语义特征是实现场景理解和生成遥感图像描述的基础。通过提取和捕捉遥感图像中丰富的视觉特征和语义特征，可实现视觉与语义特征的深度融合和传递，增强模型对遥感图像的感知与理解。遥感图像的成像距离较远，导致图像中的遥感目标占比较小，不易于识别和捕捉。此外，遥感图像环境背景复杂、噪声信息多，如背景颜色与遥感目标颜色相似导致遥感目标难以辨别，云层、大气颗粒及雾等情况也给遥感图像特征的提取和捕捉带来极大的困难和挑战。CNN具有强大的空间感知能力，可从遥感图像中提取丰富的视觉特征和语义特征，因此，本方法采用CNN中的残差网络(ResNet-101)构建编码器，对遥感图像进行特征提取。对于输入的遥感图像 F，采用CNN进行编码和特征提取，并利用线性映射层 W_1 将提取的特征映射至词嵌入空间中，以获取嵌入量 x_t，可表示为

$$x_t = W_1 X_{\mathrm{CNN}}(F), \tag{1}$$

式中，X_{CNN} 为CNN的特征提取操作，t 为对应时刻。

2.2 解码器

遥感图像描述除了需要提取和学习遥感图像特征外，还需捕捉和挖掘文本语句内部单词之间的句法特征及单词位置编码，通过动态选择图像特征学习图像特征和文本语义特征之间的联系及映射关系，消除图像与文本跨模态数据间的壁垒，实现对遥感图像的解译，生成与遥感图像内容相关且语义通顺的文本描述。LSTM是一种循环神经网络，对文本信息具有良好的学习和处理能力。LSTM每个时间状态的神经元不仅接收当前状态的输入，还接收上一时刻状态的输入，从而利用上下文信息使网络充分学习到语句的上下文历史信息，解决语句单词之间的长期依赖问题，减少信息的丢失，提升遥感图像描述的准确性。因此，本方法采用LSTM网络搭建解码器，捕捉语句内部单词之间的句法特征及单词位置编码，并学习图像特征和文本语义特征之间的映射关系。t 时刻LSTM的输入门 i_t、遗忘门 f_t、输出门 o_t、细胞状态 c_t、隐藏状态 h_t 可表示为

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \tag{2}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \tag{3}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o), \tag{4}$$

$$c_t = i_t \otimes \varphi(W^*_{zx}x_t + W^*_{zh}h_{t-1} + b^*_z) + f_t \otimes c_{t-1}, \tag{5}$$

$$h_t = o_t \otimes \tanh(c_t), \tag{6}$$

式中，W 与 b 分别为权重与偏置，σ 为Sigmoid激活函数，$\otimes$ 为矩阵元素相乘法，φ 为Maxout非线性激活函数。

对于 t 时刻下的隐藏状态 h_t，将其与权重 W_s 相乘，并输入Softmax激活函数以获取所有单词的概率分布，选取概率值最大的单词 w_t 作为输出，可表示为

$$w_t = \mathrm{Softmax}(W_s h_t)。 \tag{7}$$

用交叉熵作为模型的损失函数，可表示为

$$L(\theta) = -\sum_{t=1}^{T} \log[p_\theta(w_t^* \mid w_1^*, \cdots, w_{t-1}^*)], \tag{8}$$

式中，$p_\theta(w_t^* \mid w_1^*, \cdots, w_{t-1}^*)$ 由(7)式决定。通过在训练中最小化该交叉熵损失函数，求解出最优的模型参数 θ，完成模型的训练和学习。

2.3 注意力机制

遥感图像具有背景复杂、噪声信息多、目标占比小且容易受云层、大气颗粒及雾影响的特点，给遥感图像描述带来极大的困难。为了增强模型对显著性信息的关注，减少噪声信息的干扰，提升描述结果的准确性，本方法引入注意力机制。注意力机制是一种受人类视觉注意力机制启发而提出的权重参数分配机制，通过计算输入信息的重要性分布，在不同时刻重点关注显著性强的特征，忽略其他无关背景的特征并降低云层等噪声的影响，协助模型在大量信息中快速捕捉到当前最需要的信息，增强模型对遥感图像的感知和理解，使生成的描述更加准确。注意力机制的可视化效果如图2所示，注意力机制通过重点关注遥感图像中显著性强的飞机及船舰区域特征(深色区域)，忽略其他无关背景特征的干扰，增强模型对遥感图像的感知与理解，减少关键信息的丢失。

对于从遥感图像中 N 个不同位置提取出的特

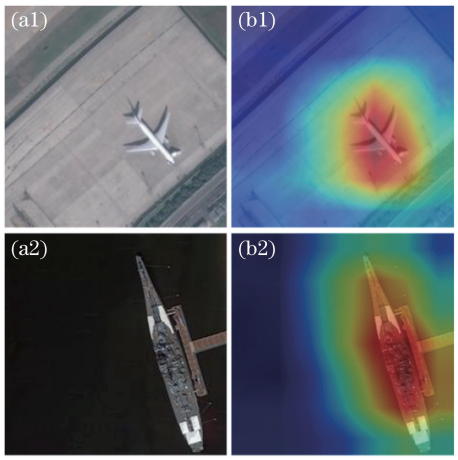

图 2 注意力机制的可视化。(a)原始图像;(b)注意力图

Fig. 2 Visualization of the attention mechanism. (a) Original image; (b) attention image

征 I_i(i 表示某个位置),注意力机制通过计算其重要性分布,筛选出显著性强的特征 I_t,可表示为

$$I_t=\sum_{i=1}^{N}\alpha_t^i I_i, \tag{9}$$

$$\alpha_t=\mathrm{Softmax}(a_t+b_\alpha), \tag{10}$$

$$a_t^i=W\tanh(W_{aI}I_i+W_{ah}h_{t-1}+b_a), \tag{11}$$

式中,h 为隐藏状态,W 与 b 分别为权重与偏置。

在解码阶段,注意力机制将筛选出的显著性特征 I_t 传递至 LSTM 的细胞状态 c 进行处理和利用,以减少关键信息的丢失,增强网络对遥感图像内容的感知和理解,可表示为

$$c_t=i_t\otimes\varphi(W_{zx}^{*}x_t+W_{zI}^{*}I_t+W_{zh}^{*}h_{t-1}+b_z^{*})+f_t\otimes c_{t-1}。 \tag{12}$$

2.4 强化学习策略

传统图像描述方法在训练阶段采用反向传播算法在给定前一个真实单词的情况下,最大化下一个真实单词出现的概率;在测试阶段,则基于模型先前生成的单词预测下一个单词出现的概率。这种方式会造成训练阶段和测试阶段不匹配,引起暴露偏差现象,导致测试阶段容易产生误差并不断累积,降低生成描述文本的质量。此外,模型在训练阶段采用交叉熵损失函数进行优化,而在测试阶段采用 BLUE 等离散且不可微的指标评价生成文本的质量,该方式会出现优化方向不一致的缺陷,导致网络无法直接利用 BLUE 等评价指标进行优化训练,当交叉熵损失函数最小时不一定会产生最佳评价结果。

为了消除暴露偏差及优化方向不一致的缺陷,本方法引入强化学习策略[10]。强化学习策略中的梯度算法能对不可微的离散变量进行端到端训练,并根据 BLUE 等指标直接对模型进行优化,提升模型的训练效果。强化学习策略将解码端的 LSTM 视为智能体,与图像和文本特征等外部环境进行交互,并定义学习策略 p 指导模型对下一个单词进行预测。每次预测完成后,智能体 LSTM 对其细胞状态、隐藏状态及注意力权重等参数进行更新。生成文本描述后,强化学习策略采用 BLUE 等指标衡量文本描述与人工标注参考语句间的契合度及相似度,赋予 LSTM 一个期望奖励,并以最小化该负期望奖励作为目标,对模型进行优化,可表示为

$$L(\theta)=-E_{w^s\sim p_\theta}[r(w^s)], \tag{13}$$

式中,θ 为模型参数,w^s 为各个时刻单词组成的序列,$r()$为奖励函数,$E()$为期望函数。实际应用中,$L(\theta)$一般通过策略 p_θ 进行单采样获得,可表示为

$$L(\theta)\approx-r(w^s),w^s\sim p_\theta。 \tag{14}$$

强化学习采用策略梯度算法计算 $L(\theta)$的梯度,可表示为

$$\nabla_\theta L(\theta)=-E_{w^s\sim p_\theta}[r(w^s)\nabla_\theta\log p_\theta(w^s)]。 \tag{15}$$

实际应用中,为了便于求解,采用蒙特卡罗单采样进行近似估计,可表示为

$$\nabla_\theta L(\theta)\approx-r(w^s)\nabla_\theta\log p_\theta(w^s)。 \tag{16}$$

采样的随机性和上下文归一化的缺乏,导致采用强化学习策略计算梯度时会造成较大的方差,引起训练过程不稳定。为了减小方差,引入一个基准因子 b,对期望奖励函数进行约束和校正,可表示为

$$\nabla_\theta L(\theta)=-E_{w^s\sim p_\theta}[(r(w^s)-b)\nabla_\theta\log p_\theta(w^s)], \tag{17}$$

为了保持对梯度的无偏估计,基准因子 b 可以为任意不依赖于 w^s 的函数。采用蒙特卡罗单采样进行近似估计时,梯度$\nabla_\theta L(\theta)$可表示为

$$\nabla_\theta L(\theta)\approx-[r(w^s)-b]\nabla_\theta\log p_\theta(w^s)。 \tag{18}$$

采用链式求导法则,得到最终的梯度表达式为

$$\nabla_\theta L(\theta)=\sum_{t=1}^{T}\frac{\partial L(\theta)}{\partial s_t}\frac{\partial s_t}{\partial\theta}, \tag{19}$$

式中,s_t 为 Softmax 函数的输入。采用蒙特卡罗单采样进行近似估计时,(19)式中的$\frac{\partial L(\theta)}{\partial s_t}$可表示为

$$\frac{\partial L(\theta)}{\partial s_t}\approx[r(w^s)-b][p_\theta(w_t\mid h_t)-l_{w_t^s}], \tag{20}$$

式中,$l_{w_t^s}$ 为词的 one-hot 向量表示,w_t 与 h_t 分别为

t 时刻的单词及内部向量表示。

3 实验结果与分析

为了实现遥感图像描述，在搭载图形处理器(GPU)的工作站上，采用公开遥感图像描述数据集对本方法进行训练和学习。训练完成后采用图像描述中常用的 BLUE 等评价指标对该模型进行评估，并与当前图像描述性能较好的其他方法进行对比，以验证本方法的有效性，实验流程如图 3 所示。

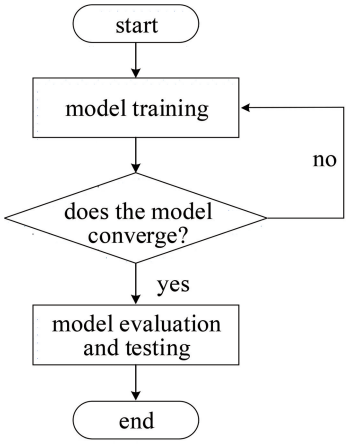

图 3 实验流程图

Fig. 3 Flow chart of the experiment

3.1 实验数据与平台

为了验证本方法的有效性，采用公开的遥感图像描述数据集 RSICD[9] 对本方法进行训练和验证。该数据集包含 10921 张遥感图像，其中，训练集、验证集和测试集的图像数量分别为 8734 张、1094 张和 1093 张，每张图像提供了 5 个反映图像内容的描述语句。RSICD 数据集数据量丰富、遥感目标类别多且覆盖多种环境背景，可用于遥感图像描述模型的训练和测试。RSICD 数据集中的部分图像及标注语句如图 4 所示。

遥感图像描述模型在训练中需要进行大量的运算，消耗的计算资源较大，因此，在搭载有 GPU 的工作站中进行实验。工作站的配置：操作系统为 Ubuntu16.04，CPU 为 Intel Core i7-8700，GPU 为 GeForce GTX 2060，内存为 16 G，同时采用 CUDA 10.1 与 CUDNN 7.6.3 进行加速处理。

1. Three planes are parked next to a road on an airport.
2. An airport near a road has three planes parked.
3. Three planes are parked on the airport beside a road.
4. Three planes are parked next to a road on an airport.
5. What stand beside this straight road are several airplanes with red tail wings.

1. An old court is surrounded by white houses.
2. A playground is surrounded by many trees and long buildings.
3. Many green trees and several long buildings are around a playgrond.
4. This narrow, oval football field and closing basketball court, tennis court, parking lot together form this area, with plants wreathing it.
5. A playground with basketball fields next to it is surrounded by many green trees and buildings.

(a) (b)

图 4 RSICD 数据集中的部分图像及标注语句。(a)图像；(b)标注语句

Fig. 4 Some images and annotation captions in RSICD dataset. (a) Image; (b) annotation captions

3.2 实验参数设置

实验基于 Pytorch 深度学习框架进行，将初始学习率设置为 5×10^{-5}，训练次数(Epoch)设置为 50 次，batch_size 设置为 10，采用 Adam 作为优化器。训练中采用学习率衰减的方式，每迭代 3 次将学习率衰减为当前学习率的 0.8 倍，以防止网络出现较大的波动。为了加快网络收敛，采用预训练好的 ResNet-101 权重初始化网络参数，LSTM 隐含层、词嵌入以及注意力机制的维度均设置为 512。为了提高生成描述文本的质量，训练中采用集束搜索(Beam search)策略，同时将集束搜索的大小设置为 2。

3.3 评价指标

为了评价本方法的描述性能，采用图像描述中常用的 BLUE[11]、ROUGE_L(Recall-oriented understudy for gisting evaluation)[12]、METEOR(Metric for evaluation of translation with explicit ordering)[13] 及 CIDEr(Consensus-based image description evaluation)[14] 等作为评价指标。其中，BLUE、ROUGE_L 和 METEOR 指标的取值范围为[0,1]，CIDEr 的取值范围为[0,5]，评价指标得分

越高，表明生成的评价描述语句与人工标注的参考语句越相似，图像描述效果越好。BLUE 指标通过对比待评价语句与参考语句之间 n 个连续字符的匹配情况进行评价，根据 n 的取值，可将 BLUE 指标分为 BLUE-1、BLUE-2、BLUE-3 和 BLUE-4；ROUGE_L 指标基于召回率进行评价，通过计算待评价语句和参考语句之间的最长公共子序列度量两者在语句上的相似性；METEOR 通过计算待评价语句与参考语句之间的精度与召回率的调和平均值进行评价；CIDEr 指标从语法性、显著性和重要性方面出发，通过计算待评价语句与参考语句之间的余弦距离衡量两者的相似性。

3.4 实验结果

为了验证本方法的有效性，以 BLUE 等为评价指标，将训练好的模型在 RSICD 遥感图像描述数据集中进行测试，并与当前描述性能较好的其他图像描述方法进行对比，结果如表 1 所示。可以发现，相比其他方法，本方法取得了较高的精度，其 BLUE-1、BLUE-2、BLUE-3、BLUE-4、METEOR、ROUGE_L 及 CIDEr 指标的得分分别为 0.703、0.562、0.467、0.385、0.354、0.642 及 2.051，均优于 NIC[15]、Attend[16] 及 Adaptive[17] 方法。

表 1 不同方法的图像描述检测结果

Table 1 Image caption detection results of different methods

Method	BLUE-1	BLUE-2	BLUE-3	BLUE-4	METEOR	ROUGE_L	CIDEr
NIC[15]	0.658	0.519	0.436	0.357	0.322	0.605	1.926
Attend[16]	0.676	0.533	0.442	0.364	0.330	0.617	1.965
Adaptive[17]	0.685	0.546	0.451	0.372	0.338	0.629	1.984
Ours	0.703	0.562	0.467	0.385	0.354	0.642	2.051

为了探究编码-解码架构（Encode-Decode）、注意力机制及强化学习（Reinforce）的有效性，设置不同的分组实验，以验证不同模块对描述性能的影响，实验结果如表 2 所示。可以发现，以编码-解码为基础架构的模型取得了较好的检测精度，其 BLUE-1、BLUE-4、METEOR、ROUGE_L 及 CIDEr 指标得分分别为 0.667、0.361、0.325、0.612 及 1.958，原因是编码-解码模型通过提取遥感图像特征，并捕捉语句内部单词之间的句法特征，实现对遥感图像的解译与描述。引入注意力机制后，模型的指标得分分别提升了 0.014、0.008、0.011、0.013 及 0.019。原因是注意力机制可协助模型在不同时刻重点关注显著性强的特征，忽略其他无关背景的特征并降低云层等噪声的影响，提升描述结果的准确性。引入强化学习策略后，模型的指标得分分别提升了 0.022、0.016、0.018、0.017 及 0.074，这表明强化学习策略可根据 BLUE 等指标直接对模型进行优化，消除暴露偏差以及优化方向不一致的缺陷，提升模型的训练效果。

表 2 不同模块对遥感图像描述的影响

Table 2 Influence of different modules on remote sensing image caption

Method	BLUE-1	BLUE-4	METEOR	ROUGE_L	CIDEr
Encode-Decode	0.667	0.361	0.325	0.612	1.958
Encode-Decode＋Atten	0.681	0.369	0.336	0.625	1.977
Encode-Decode＋Atten＋Reinforce	0.703	0.385	0.354	0.642	2.051

3.5 图像描述结果的可视化

图 5 为本方法与其他方法在遥感图像描述测试集中的部分描述结果。可以发现，本方法在密集小目标[图 5(d1)]、雾气覆盖遮挡[图 5(d2)]、目标颜色特征与背景环境颜色特征较为相似[图 5(d3)]的复杂环境下，可准确生成与遥感图像内容相关的文本描述，这表明本方法具有较强的鲁棒性和适应性，对复杂环境下的遥感图像具有良好的描述性能。此外，在图 5(d1)中，本模型检测出汽车的属性为白色和黑色，表明模型充分提取和学习了遥感目标的视觉特征，对遥感目标的细节属性信息具有良好的表征性能。而 NIC[15]、Attend[16] 及 Adaptive[17] 方法虽然也生成了相应的文本描述，但存在描述不够准确的缺陷。如与图 5(d2)的描述相比，Adaptive 方法在图 5(c2)中未能识别和描述出操场（playground）周围的建筑（buildings），存在一定的漏检现象。

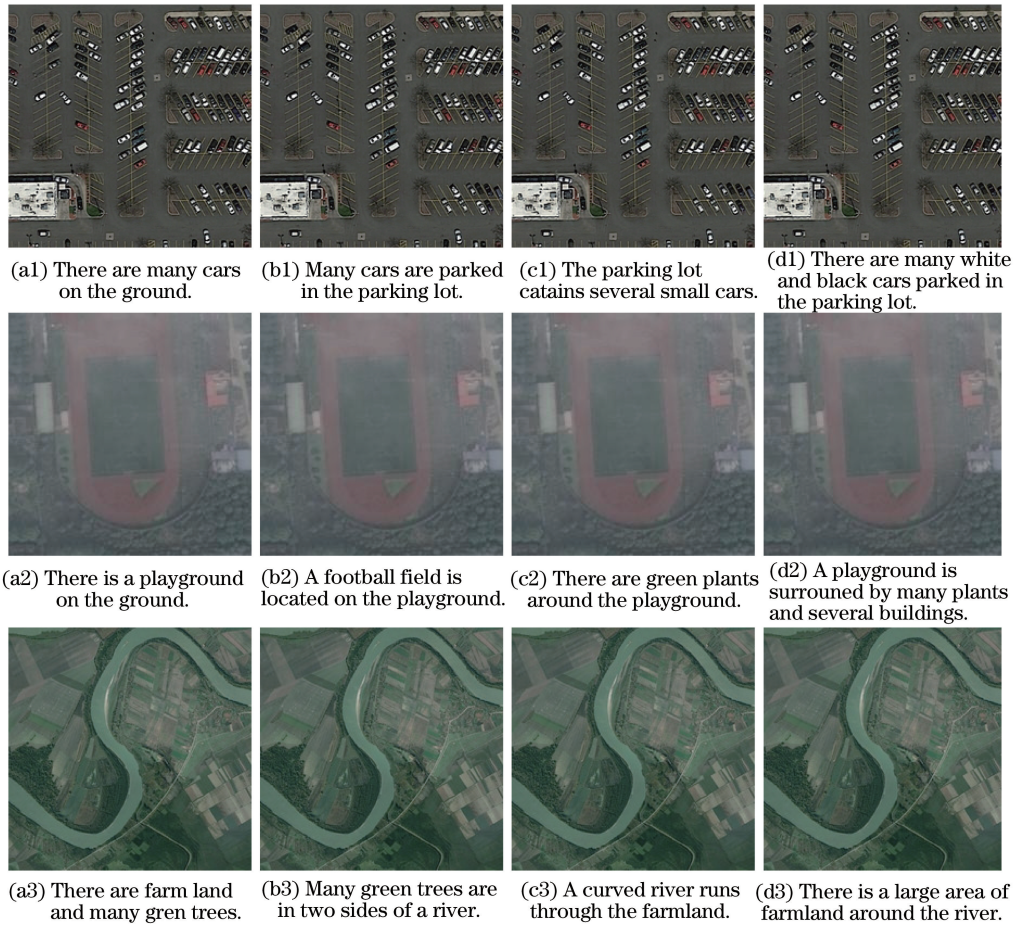

图 5 不同方法的遥感图像描述结果。(a) NIC;(b) Attend;(c) Adaptive;(d) Ours

Fig. 5 Remote sensing image caption results of different methods. (a) NIC; (b) Attend; (c) Adaptive; (d) Ours

3.6 注意力机制的可视化结果

为了验证注意力机制的有效性,图 6 展示了注意力机制在遥感图像描述文本生成过程中的可视化效果。可以发现,注意力机制通过对图像特征进行筛选,重点关注显著性强的目标区域特征,并摒弃其他冗余特征和噪声信息,增强模型对遥感图像内容的感知和理解,提升了描述结果的准确性。

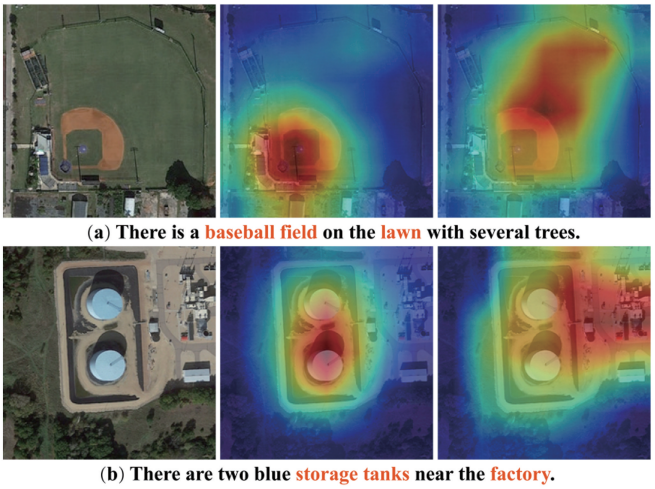

图 6 注意力机制在遥感图像描述中的可视化结果。(a)图像 1;(b)图像 2

Fig. 6 Visualization results of attention mechanism in remote sensing image caption. (a) Image 1; (b) image 2

4 结 论

为了实现对遥感图像的描述，通过采用 CNN 构建编码器、利用 LSTM 搭建解码器、引入注意力机制、采用强化学习策略，提出了一种遥感图像描述方法。为验证本方法的有效性，采用公开的遥感图像描述数据集进行了训练和验证。实验结果表明，本方法取得了较高的精度，对复杂环境背景下的遥感图像具有良好的图像描述性能，可实现对遥感图像的解译与描述。下一步还将对模型进行改进和优化，以进一步提升遥感图像的描述性能。

参 考 文 献

[1] Yao Q L, Hu X, Lei H. Object detection in remote sensing images using multiscale convolutional neural networks[J]. Acta Optica Sinica, 2019, 39(11): 1128002.
姚群力, 胡显, 雷宏. 基于多尺度卷积神经网络的遥感目标检测研究[J]. 光学学报, 2019, 39(11): 1128002.

[2] Zhu M M, Xu Y L, Ma S P, et al. Airport detection method with improved region-based convolutional neural network[J]. Acta Optica Sinica, 2018, 38(7): 0728001.
朱明明, 许悦雷, 马时平, 等. 改进区域卷积神经网络的机场检测方法[J]. 光学学报, 2018, 38(7): 0728001.

[3] Xu Z J, Ding Y. Ship object detection of remote sensing images based on adaptive rotation region proposal network[J]. Laser & Optoelectronics Progress, 2020, 57(24): 242805.
徐志京, 丁莹. 自适应旋转区域生成网络的遥感图像舰船目标检测[J]. 激光与光电子学进展, 2020, 57(24): 242805.

[4] Dong Y F, Zhang C T, Wang P, et al. Airplane detection of optical remote sensing images based on deep learning[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041007.
董永峰, 仉长涛, 汪鹏, 等. 基于深度学习的光学遥感图像飞机检测算法[J]. 激光与光电子学进展, 2020, 57(4): 041007.

[5] Zhang M, Wang S C, Yang D F. Air-to-ground target detection algorithm based on attention learning in key areas[J]. Laser & Optoelectronics Progress, 2020, 57(4): 041006.
张萌, 王仕成, 杨东方. 重点区域注意力学习的空对地目标检测算法[J]. 激光与光电子学进展, 2020, 57(4): 041006.

[6] Zhao J Q, Wang H Z, Zhou Y, et al. Remote sensing image description generation method based on attention and multi-scale feature enhancement[J]. Computer Science, 2021, 48(1): 190-196.
赵佳琦, 王瀚正, 周勇, 等. 基于多尺度与注意力特征增强的遥感图像描述生成方法[J]. 计算机科学, 2021, 48(1): 190-196.

[7] Shi Z W, Zou Z X. Can a machine generate humanlike language descriptions for a remote sensing image?[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6): 3623-3634.

[8] Qu B, Li X L, Tao D C, et al. Deep semantic understanding of high resolution remote sensing image[C]//2016 International Conference on Computer, Information and Telecommunication Systems (CITS), July 6-8, 2016, Kunming, China. New York: IEEE Press, 2016: 1-5.

[9] Lu X Q, Wang B Q, Zheng X T, et al. Exploring models and data for remote sensing image caption generation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(4): 2183-2195.

[10] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1179-1195.

[11] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-ACL '02, July 7-12, 2002. Philadelphia, Pennsylvania. Morristown, NJ, USA: Association for Computational Linguistics, 2001: 311-318.

[12] Lin C Y. ROUGE: A package for automatic evaluation of summaries[EB/OL]. (2004-07-15)[2021-04-10]. https://aclanthology.org/W04-1013/.

[13] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[EB/OL]. (2005-06-12)[2021-04-18]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=00D9354FBD891E7E5E554DD61609BFE8?doi=10.1.1.61.2290&rep=rep1&type=pdf.

[14] Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 4566-4575.

[15] Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator[C]//2015 IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3156-3164.

[16] Xu K, Ba J L, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [C]//ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. 2015: 2048-2057.

[17] Lu J S, Xiong C M, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3242-3250.