DOI: 10.13382/j.jemi.B2306987

面向遥感图像的结构化图像描述网络*

李国燕[1] 田明达[1] 董春华[2] 郝志鹏[1]
(1. 天津城建大学计算机与信息工程学院 天津 300384;2. 天津城建大学地质与测绘学院 天津 300384)

摘 要:为了解决标准注意力方法只能生成粗粒度的注意力区域,既无法获取遥感对象之间的地理关系,也不能充分利用遥感图像语义内容的问题,提出了一种面向遥感图像的结构化图像描述网络(geo-object relational segmentation for remote sensing image captioning,GRSRC)。首先,针对遥感图像特征高度结构化的特点,提出基于结构化遥感图像语义分割的特征提取方法,通过增强编码器特征提取能力实现更准确的表达;同时,引入注意力机制对分割区域进行加权,使模型能够更加关注重要的语义信息;其次,针对遥感图像空间对象位置关系较为明确的特点,在注意力机制中融合地理空间关系,使生成的描述更加准确且具有空间一致性;最后,在 RSICD、UCM、Sydney 3 个公开的遥感数据集上进行实验评估,在 UCM 数据集上,BLEU-1 达到了 84.06、METEOR 达到了 44.35、ROUGE_L 达到了 77.01,相较于所对比的经典模型,分别提升了 2.32%,1.15%和 1.88%。实验结果说明模型能够更充分利用遥感图像语义内容,表明了该方法在遥感图像描述任务中具有较好的性能。

关键词:遥感影像;图像描述;空间关系;语义分割;注意力机制
中图分类号:TP753 **文献标识码**:A **国家标准学科分类代码**:520.20

Structured image description network for remote sensing images

Li Guoyan[1] Tian Mingda[1] Dong Chunhua[2] Hao Zhipeng[1]
(1. School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300384, China;
2. School of Geology and Surveying, Tianjin Chengjian University, Tianjin 300384, China)

Abstract: To address the limitation of standard attention mechanisms that can only generate coarse-grained attention regions, failing to capture the geographical relationships between remote sensing objects and underutilize the semantic content of remote sensing images, a structured image description network named GRSRC (geo-object relational segmentation for remote sensing image captioning) is proposed. Firstly, considering the highly structured nature of remote sensing image features, a feature extraction method based on structured semantic segmentation of remote sensing images is introduced, enhancing the encoder's feature extraction capability for more accurate representation. Simultaneously, an attention mechanism is incorporated to weight the segmented regions, enabling the model to focus more on crucial semantic information. Secondly, taking advantage of the well-defined spatial relationships among objects in remote sensing images, geographical spatial relations are integrated into the attention mechanism, ensuring more accurate and spatially consistent descriptions. Finally, experimental evaluations are conducted on three publicly available remote sensing datasets, RSICD, UCM, and Sydney. On the UCM dataset, BLEU-1 achieved 84.06, METEOR reached 44.35, and ROUGE_L attained 77.01, demonstrating improvements of 2.32%, 1.15%, and 1.88%, respectively, compared to classical models. The experimental results indicate that the model can better leverage the semantic content of remote sensing images, demonstrating its superior performance in remote sensing image captioning tasks.

Keywords: remote sensing; image caption; geo-relation; semantic segmentation; attention mechanism

收稿日期:2023-10-22 Received Date:2023-10-22
*基金项目:国家自然科学基金(52178295)项目资助

0　引　言

遥感图像描述是近年来飞速发展的一项重要的计算机视觉任务。如在监测和管理自然资源方面,可以了解不同地区的土地利用、植被覆盖、水资源等自然资源的状况,从而实现对这些资源的监测和管理;或在规划管理城市和交通方面,通过遥感图像描述,可以了解城市的布局、建筑密度、交通路线和道路状况等,从而进行城市和交通的规划管理[1-2]。

目前的遥感图像视觉领域,许多研究学者关注于遥感图像分类[3-4]、目标识别[5-7]、图像分割[8-9]等任务。对于这些任务,目标是对图像中的特定对象或区域进行识别、定位和分割等操作,输出结果通常是标注的图像或图像上的位置、类别等信息。但与上述任务不同,遥感图像描述的目标是从图像中提取区域特征、空间特征等信息,并获取描述语句中的语法语义特征,对图像内容进行简洁、自然的文字描述,输出结果是生成一段能够传达图像主要特征的自然语言。与其他任务相比,遥感图像描述是一项相对较新的遥感图像视觉任务,具有独特的任务目标和输出形式。

在图像描述领域中使用深度学习技术的历史可以追溯到 2014 年,COCO 数据集[10]的发布为评估图像描述模型提供了一个大规模的基准。这导致了几个编码器-解码器模型的发展,如 Show and Tell[1]、Show, Attend and Tell[11] 和 Deep Visual-Semantic Alignments for Generating Image Descriptions[12]等。这些算法大多数选择通过卷积神经网络提取图像特征,再通过循环神经网络生成自然语言描述,而在此之后,研究人员们也开始着手将这些模型应用在遥感图像中。

首个遥感图像描述模型由 Qu 等[13]提出,该模型利用深度神经网络进行高分辨率遥感图像的描述。Lu 等[14]将数种非深度学习方法与深度学习方法进行比较,研究了基于编码器-解码器的模型以及基于注意力的模型,并发布了一个名为 RSICD 的大型遥感图像描述数据集,为之后的研究打下了基础。Zhang 等[15]提出了一种多源交互式楼梯式注意力机制,将之前的语义向量作为查询,对感兴趣的区域特征应用注意力机制,以获得下一个词向量,从而提高生成句子的准确性和连贯性。Hitesh 等[16]使用辅助解码器进行多标签场景分类,辅助解码器通过多标签场景分类任务进行训练,这是首个利用多标签分类改进遥感图像描述的工作。Ren 等[17]为了提高遥感图像描述的准确性和多样性,提出了一种新颖的遮蔽引导 Transformer 网络,利用多头注意力机制提取特征并捕捉遥感图像中对象之间的关系。Liu 等[18]为了充分利用提取的多尺度信息生成准确和详细的句子,提出了多层聚合 Transformer 作为解码器。MLAT 通过自注意机制和聚合策略,能够充分利用不同 Transformer 编码层的特征。Xu 等[11]提出了软硬注意力机制,软注意力机制可以动态地调整注意力权重,提高生成质量,而硬注意力机制则更加简化和高效。农元君等[19]提出了一种基于强化学习的编码器-解码器模型,对密集小目标、雾气下的复杂环境描述效果较好。Zhao 等[20]提出一种细粒度、结构化的注意力方法,利用高分辨率遥感图像的结构化特征,生成分割掩码,以更好地进行遥感图像描述。Chen 等[21]提出一种基于地理空间关系的高分辨率遥感图像描述方法,基于图像中物体目标的拓扑关系、方向关系、距离关系来表达地理空间关系,收集制作数据进行训练,进一步强化了描述中空间关系的表达。

综上所述,现有的大多数遥感图像描述方法在生成遥感图像描述时无法充分利用遥感图像的特有结构以及遥感对象之间的空间关系。基于以上问题,本文提出一种基于语义分割和地理空间关系的遥感图像描述方法 GRSRC。基于 DeepLabV3 网络实现遥感图像空间对象的分割,获得掩码图,提取出地理对象空间关系,使模型对区域化较明显的遥感图像有更强的特征提取能力;同时构建 GORSA (geo-object relational segmentation attention)模块,利用掩码图和空间关系实现显著区域的方位注意机制,使模型更能捕捉显著区域以及方位特征,提高遥感图像描述的能力。

1　GRSRC 模型

本文提出的 GRSRC 模型结构如图 1 所示,模型主要由编码器、解码器和空间关系分割注意力模块 GORSA 3 个部分构成。编码器将图像映射为特征图;解码器将图像特征转化为描述语句;GORSA 模块生成掩码图,依据掩码图生成空间关系,并生成注意力权重。

编码器模块:ResNet(residual neural network)[22]是一种非常流行和有效的深度卷积神经网络结构。为有效提取遥感图像的特征,捕捉更丰富的信息,模型选择使用 ResNet-101 作为编码器,编码后的特征输入到空间关系分割注意力模块中。

空间关系分割注意力模块(GORSA):DeepLabV3[23]是一种经典的语义分割模型,用于将图像中的每个像素分配给特定的语义类别。空间关系分割注意力模块中采用 DeepLabV3 网络对输入图像进行分割,利用分割网络获取特征区域可以引导模型在训练过程中更加注意分割的区域。遥感图像相比于自然图像拥有更多的高聚合度区域,因此能够获得更高精度的提取图像特征,从而提高图像描述任务的性能,使得模型具有更加细致的注意机制。此外,在分割过程中可以获取到每个分割区域对应

的标签，通过利用获取到的类标签作为外部变量，在训练过程中相应增加类标签的可能性，可以提高训练精度。在地理空间关系的获取方面，本文利用分割后的区域以及标签，通过获取各区域质心，获取各个地理对象之间的空间关系。

解码器模块：GRSRC 使用长短期记忆网络作为模型的解码器，获取语句特征及图像特征之间的关系、语句特征和分割区域特征之间的关系，并融入对象之间的空间关系，实现对遥感图像的描述，生成通顺且合理的语句。下面详细介绍各个模块的细节。

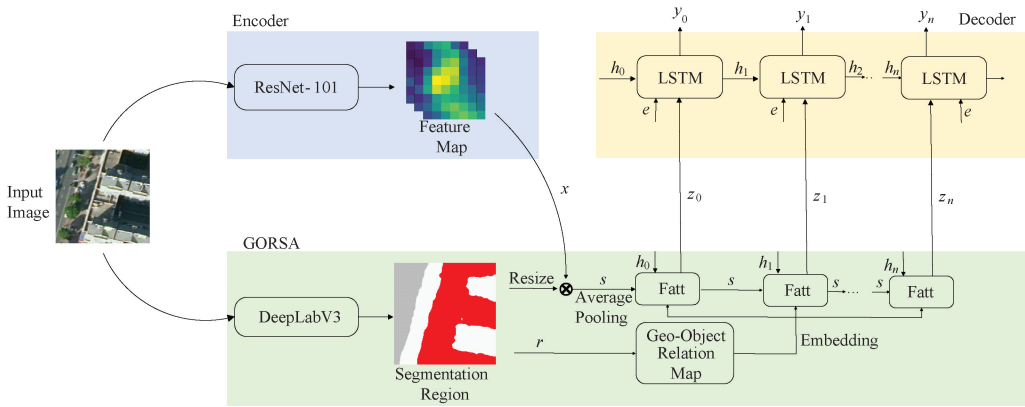

图 1 整体网络结构概述

Fig. 1 Overview of the overall network structure

1.1 编码器

考虑到 ResNet 具有深层次的网络结构，能够有效地提取遥感图像中的丰富、高级的特征，且已经在大规模的图像数据集上进行了预训练，因此模型采用 101 层的深度残差网络（ResNet-101）作为模型的编码器。由于模型并不需要利用 ResNet-101 的识别分类功能，因此在训练过程中选择将最后的平均池化层（Flatten）和全连接层去掉，直接获得特征图作为模型中的特征表示。编码时输入 ResNet-101 提取图像特征，并在调整尺寸后与分割图进行点积操作。特征 x_t 可表示为：

$$x_t = WR(I) \tag{1}$$

其中，W 为全连接层，R 为 ResNet-101 的特征提取操作。

1.2 解码器

采用 LSTM（long short-term memory）作为模型的解码器，LSTM 是 Hochreiter 等[17]提出的 RNN（recurrent neural network）的变体，用于解决 RNN 模型训练中梯度消失和爆炸的问题。本文改进 LSTM 结构如图 2 所示。相较于传统的 LSTM，该改进 LSTM 引入了上下文向量 $\hat{\boldsymbol{z}}_t$ 以作为输入的一部分代替原始的输入 x_t，以此来获取原始输入序列以外的信息，例如类别标签和空间位置等。这样的设计有助于模型更全面地理解数据，并捕捉到可能对模型预测有益的信息，使得模型能够更好地感知输入数据的上下文关系，提高模型的表达能力和泛化能力。

在解码阶段，LSTM 生成句子可用式(2)~(3)表达：

$$y_t = \mathrm{LSTM}(\boldsymbol{e}_t, \boldsymbol{h}_{t-1}, \hat{\boldsymbol{z}}_t), t \in \{1, \cdots, N\} \tag{2}$$

$$\hat{\boldsymbol{z}}_t = [x_t, z_t] \tag{3}$$

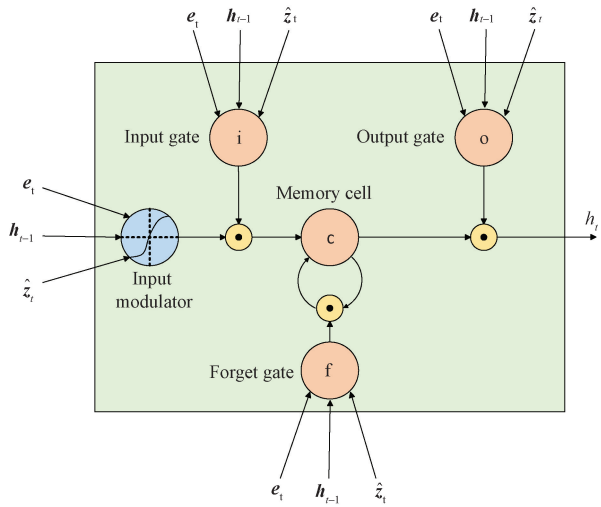

图 2 模型中使用的 LSTM 结构

Fig. 2 Structure diagram of the LSTM used in the GRSRC

其中，$\boldsymbol{y}_t$ 代表 t 时刻模型的预测单词向量，$\boldsymbol{e}_t$ 表示嵌入的单词向量，$\boldsymbol{h}_{t-1}$ 表示 LSTM 在 t 时刻的先前隐藏状态向量，$\boldsymbol{z}_t$ 代表 t 时刻的上下文向量，N 代表单词表的大小。其中 LSTM 在 t 时刻的内部状态可表示如下：

$$i_t = \sigma(W_{ih}h_{t-1} + W_{iz}\hat{\boldsymbol{z}}_t + \boldsymbol{b}_i) \tag{4}$$

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fz}\hat{\boldsymbol{z}}_t + \boldsymbol{b}_f) \tag{5}$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{oz}\hat{\boldsymbol{z}}_t + \boldsymbol{b}_o) \tag{6}$$

$$\tilde{c}_t = \tanh(W_{ch}h_{t-1} + W_{cz}\hat{\boldsymbol{z}}_t + \boldsymbol{b}_c) \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{8}$$

其中，门 i_t、f_t、o_t、h_t、c_t 分别表示 LSTM 的输入门、遗忘门、输出门、隐藏状态和存储器。$\boldsymbol{W}_{ix}$、$\boldsymbol{W}_{fx}$、$\boldsymbol{W}_{ox}$、$\boldsymbol{W}_{cx}$ 是可训练的权重矩阵，$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$ 是可训练的偏置矩阵，σ 是 Sigmoid 激活函数，$\odot$ 是点积操作，tanh 是双曲正切非线性激活函数。最后，为了获取当前时刻单词概率，使用 *SoftMax* 激活函数获取单词的概率分布，选取最高概率 p_t 作为输出，作为当前时刻输出的单词。

$$p_t = SoftMax(y_t) = \frac{\exp(y_t)}{\sum_{i=1}^{N} \exp(y_t^{(i)})} \tag{9}$$

1.3　空间关系分割注意力(GORSA)

根据遥感图像高度聚集的特点，本文提出了一种基于空间关系分割的注意力机制，首次引入 DeepLabV3 网络进行待描述图片的分割，以获取多个高聚集度的特征区域。相比于感兴趣区域池 ROI(region of interest)或选择性搜索(selective search)，本方法不但可以获取分割区域的类别，而且可以获取各区域之间的空间关系，利用遥感图像的结构特点提高描述中生成词汇的准确度。

空间分割注意力模块分为 3 部分，第 1 部分将分割后的特征进行聚集，第 2 部分通过分割图获取地理空间关系，第 3 部分以聚集特征生成上下文向量。下面详细介绍 3 部分的细节。

1)在获取特征分割图阶段，先利用 DeepLabV3 网络获取分割区域，再与编码器所生成的特征图进行点积操作以获取特征分割图，最后进行平均池化获得分割特征 s。分割机制结构如图 3 所示。

$X \in \mathbb{R}^{h \times w \times c}$ 表示编码器产生的图像特征，R_i 代表分割后的 i 类分割图，*Mean* 代表平均池化操作，对于单元 i，产生的分割特征表示 s 可以表示如下：

$$s = \{s_1, s_2, \cdots, s_i\} \tag{10}$$

$$s_i = Mean\left(\sum_{(x,y) \in R_i} X(x,y) \odot \mathrm{R}_i\right) \tag{11}$$

其中，R_i 是分割区域，其被从输入图像的大小调整为特征图的大小。在沿着区域 R_i 内的像素 (x,y) 之间执行求和。

2)在通过分割图获取地理空间关系阶段，地理空间关系表现为地理对象之间的空间分布依赖性，这种类型的依赖关系通常可以用方向关系和距离关系来表示，GRSRC 中获取空间关系的结构如图 4 所示。

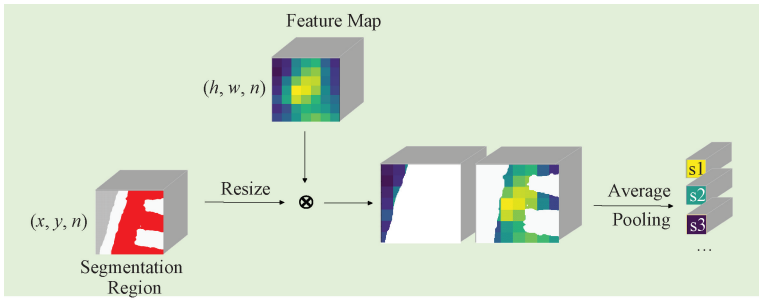

图 3　分割机制结构

Fig. 3　Diagram of the segmentation mechanism structure

为了获取各个对象之间的空间关系，考虑到 DeepLabV3 分割后区域的不规则性，选择将分割区域的质心位置作为空间对象的绝对位置，进而由对象位置判断对象之间的空间关系。表 1 中显示了地理空间关系的表示方法，采用方向关系和距离关系来表示。其中判断对象间方向关系的公式可表示如下：

$$\mathrm{d}x = x_2 - x_1 \tag{12}$$

$$\mathrm{d}y = y_2 - y_1 \tag{13}$$

$$\frac{\arctan\dfrac{\mathrm{d}x}{\mathrm{d}y}}{\pi} \times 180, \mathrm{d}x > 0 \ \&\ \mathrm{d}y > 0 \tag{14}$$

$$\frac{\arctan\dfrac{\mathrm{d}x}{\mathrm{d}y}}{\pi} \times 180 + 180(\mathrm{d}x < 0 \ \&\ \mathrm{d}y < 0) \mid (\mathrm{d}x > 0 \ \&\ \mathrm{d}y < 0) \tag{15}$$

$$\frac{\arctan\dfrac{\mathrm{d}x}{\mathrm{d}y}}{\pi} \times 180 + 360, \mathrm{d}x < 0 \ \&\ \mathrm{d}y > 0 \tag{16}$$

其中，$\mathrm{d}x$、$\mathrm{d}y$ 为两个空间对象的坐标位置之差，由 θ 判断方向关系的条件如表 1 所示，将当前质心位置周围 360°平均分为 8 份，作为 8 个方向的判断条件。

判断对象间距离关系的公式可表示如下：

$$\phi = \sqrt{w^2 \times h^2} \tag{17}$$

$$L = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{18}$$

$$\boldsymbol{r} = e(\phi) + e(L) \tag{19}$$

其中，w 和 h 分别代表图像的宽度和高度，ϕ 代表图像的对角线长度，L 代表两个空间对象之间的距离，根据对象间的距离，由表 1 的条件进行关系判断。

获取到对象之间的关系后将其转换为词嵌入向量 $\boldsymbol{r}$，

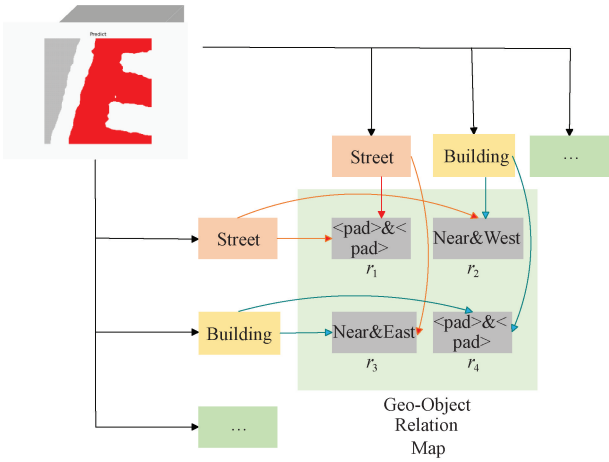

图 4　地理空间关系机制结构

Fig. 4　The structure of the geospatial relationship mechanism

未在对应分割区域计算权重 $\widetilde{\alpha}_t^{(i)}$ 时 $\boldsymbol{r}$ 向量设为全空，在对应分割区域计算权重 $\widetilde{\alpha}_t^{(i)}$ 时则连接在 s_i 和 h_{t-1} 后参与计算。

表 1　地理对象关系和判断条件

Table 1　Geographical object relations and judgment conditions

地理关系	关系词	关系判断
方向关系	North	337.5<=θ or θ<22.5
	Northwest	22.5<=θ<67.5
	West	67.5<=θ<112.5
	Southwest	112.5<=θ<157.5
	South	157.5<=θ<202.5
	Southeast	202.5<=θ<247.5
	East	247.5<=θ<292.5
	Northeast	292.5<=θ<337.5
距离关系	Close	$L <= \frac{\phi}{4}$
	Near	$L <= \frac{\phi}{2}$
	Far	$L > \frac{\phi}{2}$

3）在获取上下文向量阶段，本文方法中的上下文向量 $\boldsymbol{z}_t$ 是时间 t 图像的相应分割区域单元的动态表示。给定特征表示 s_i 和先前的隐藏状态 h_{t-1}，计算注意力权重值 $\widetilde{\alpha}_t^{(i)}$，它表示区域单元和生成的词向量 y_t 之间的相关性，计算权重值的结构如图 5 所示，公式可表示如下：

$$\widetilde{\alpha}_t^{(i)} = F_{att}(s_i, h_{t-1}, r) \tag{20}$$

其中，$Fatt$ 表示多层感知器，该感知器经过训练以生成注意力权重。为了构建网络 F_{att}，首先通过全连接层将 s_i 和 h_{t-1} 的尺寸调整相同，再连接关系表中对应地理关系的词嵌入向量。然后，将变换后的向量相加，以融合来自区域单元和上下文的信息，并将融合向量进一步送到另一个全连接层，以产生注意力权重 $\widetilde{\alpha}_t^{(i)}$，公式可表示如下：

$$F_{att}(s_i, h_{t-1}, \boldsymbol{r}) = f_4(RELU(f_1(s_i) + f_2(h_{t-1}) + f_3(r))) \tag{21}$$

其中，f_1、f_2、f_3 和 f_4 是 4 个全连接层，ReLU 表示线性单元激活函数，$\boldsymbol{r}$ 表示地理关系的词嵌入向量矩阵。然后，使用 $SoftMax$ 层对时间步长 t 处的 N 个区域的注意力权重进行归一化，以产生最终的注意力向量 α_t，公式可表示如下：

$$\boldsymbol{\alpha}_t = SoftMax([\widetilde{\alpha}_t^{(1)}, \cdots, \widetilde{\alpha}_t^{(N)}]) \tag{22}$$

得到了注意力加权向量后，上下文向量 $\boldsymbol{z}_t$ 就可以表示为区域特征 s_i 和它们的注意力权重 $\alpha_t^{(i)}$ 的线性组合，公式可表示如下：

$$\boldsymbol{z}_t = \sum_{i=1}^{N} \alpha_t^{(i)} s_i \tag{23}$$

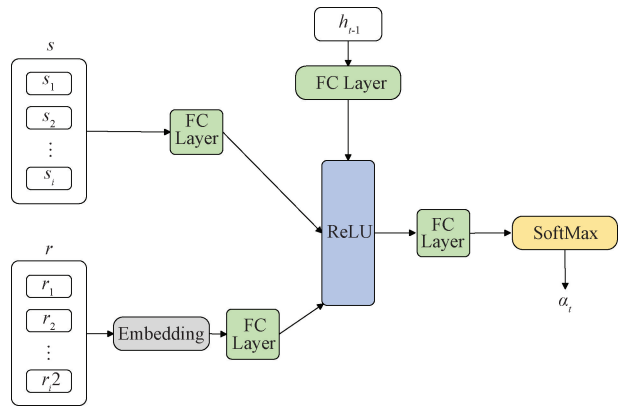

图 5　上下文向量机制结构

Fig. 5　Context vector mechanism structure diagram

2　实验结果与分析

本章主要通过实验来验证 GRSRC 模型的有效性和鲁棒性。首先对遥感图像描述中常用的评价指标及数据集进行介绍；然后介绍实验仿真参数的设置；最后在不同场景下验证 GRSRC 模型的性能，并与其他算法进行对比试验[12-13,19,24]。

2.1　实验数据与平台

GRSRC 模型在训练中需要进行大量的运算，因此模型在搭载有 GPU 的服务器中进行实验。服务器的配置：操作系统为 Ubuntu22.04，CPU 为 Intel Core i9-10900X 3.70GHz×20，GPU 为 GeForce RTX 3070Ti，内存为 32 G，同时采用 CUDA11.6 进行加速处理。

为了验证本方法的有效性，采用公开的遥感图像描述数据 RSICD[13]、UCM[12]、Sydney[12] 对本方法进行训练

和验证。

RSICD 是遥感图像描述任务中使用最广泛的数据集。它包含从 AID 数据集[25]和其他平台收集的 10 921 幅遥感图像,其中,训练集、验证集和测试集的图像数量分别为 8 734、1 094 和 1 093 张。

UCM 标题数据集基于 UC Merced 土地使用数据集[26]构建,它包含来自 21 种场景的 2 100 幅遥感图像,其中,训练集、验证集和测试集的图像数量分别为 1 680、210 和 210 张,图像尺寸为 256×256。Sydney 数据集基于悉尼陆地数据集[27]构建。它总共包含 613 张从澳大利亚悉尼的谷歌地球图像中收集的遥感图像,其中,训练集、验证集和测试集的图像数量分别为 490、62 和 61 张,图像尺寸为 500×500。

以上每个数据集中每张图像都提供了 5 个反映图像内容的描述语句。数据集中的部分图像及标注语句如图 6 所示。

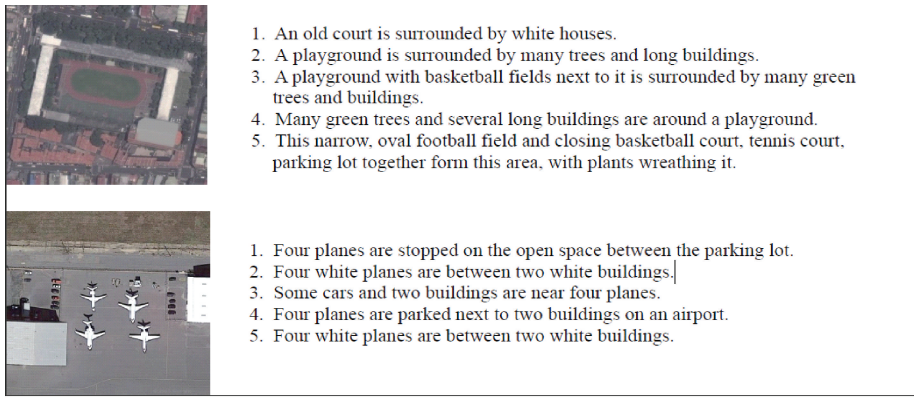

图 6　RSICD 数据集中的部分图像和标注语句

Fig. 6　Partial images and annotation captions in the RSICD dataset

2.2　实验参数

实验基于 Pytorch 深度学习框架进行。编码器的初始学习率设置为 1×10^{-4},解码器的初始学习率设置为 3×10^{-4}。训练次数(Epoch)设置为 200 次,BatchSize 设置为 32,采用 Adam 作为优化器。训练过程中使用学习率衰减,若准确率提升,则将学习率衰减为当前学习率的 0.9 倍,以避免网络波动。LSTM 隐藏层、词嵌入以及上下文向量的维度都设置为 512,空间关系图设置为 8×8,集束搜索的大小设置为 5。

2.3　评价指标

使用 3 种不同的指标来评估生成描述的准确性,包括 BLEU、ROUGE-L 和 METEOR,这些都在图像描述领域中广泛使用。

BLEU 通过使用 n-gram 来测量生成的描述和真实描述之间的准确度,常见的指标有 BLEU-1、BLEU-2、BLEU-3、BLEU-4 这 4 种。BLEU 的分数范围为 0~1,越接近 1 表示生成的结果与参考结果越相似。

由于 BLEU 没有考虑召回率,为了解决这一弱点,METEOR 被引入来优化 BLEU 中逐字匹配计算准确度的问题。它不仅考虑了 n-gram 的匹配,还使用了外部词对齐和词干处理等技术来对翻译进行比较。METEOR 的分数范围为 0~1,越接近 1 表示生成的结果与参考结果越相似。

ROUGE-L 是 ROUGE 的修改版本,它使用生成的描述和真实描述之间的最长公共子序列(LCS)来计算具有召回偏差的 F 测度。ROUGE-L 的分数范围为 0~1,越接近 1 表示生成的结果与参考结果越相似。

2.4　实验结果

为了说明提出的 GRSRC 模型的编码解码结构、区域分割注意、地理关系注意的有效性,设计不同分组进行消融实验以验证不同模块对于图像描述的性能影响。以 ResNet101 作为编码器,LSTM 作为解码器的结构为基础,分别进行分割机制的实验和空间机制的实验。所有实验均在公开数据集 RSICD 上进行,通过定量计算来评价不同模块带来的图形模式性能影响,实验结果如表 2 所示。

表 2　不同模块对于描述准确率的影响(RSICD)

Table 2　Influence of different modules on description accuracy (RSICD)

模块	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
Encode-Decode	66.27	51.13	51.29	44.19	31.44	66.25
Encode-Decode + Segmentation	69.10	52.98	53.82	46.62	33.49	68.12
Encode-Decode + Segmentation + Geo-Relation	69.49	53.67	54.19	47.13	34.43	68.61

从实验结果中可以得出:引入分割机制后,模型在BLEU-1、BLEU-2、BLEU-3、BLEU-4、ROUGE-L 和METEOR 评价指标上的得分提升了 2.83%、1.85%、2.53%、2.43%、2.05%、1.87%、5.19%,说明该模块通过引入分割机制,完成了更精确的注意特征获取,有效提升了遥感图像描述能力。分割机制中的分割效果实验如图 7 所示。引入了空间关系机制后模型的指标得分提升了 0.39%、0.69%、0.37%、0.51%、0.94%、0.49%、3.38%,说明该模块通过引入空间机制,获取了更精确的语义信息,有效提升了遥感图像描述能力。对于空间机制中质心的获取实验结果如图 8 所示。这表明了根据语义分割以及地理空间关系的遥感图像描述方法可以提升模型的训练效果。

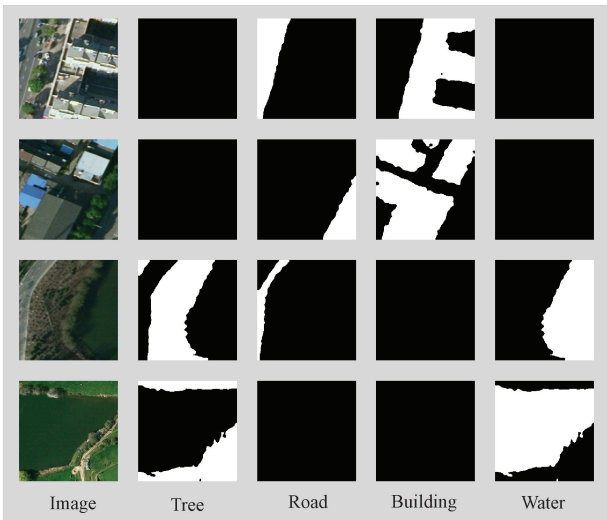

图 7　经过分割后的区域和一些对应类别

Fig. 7　Segmented areas and corresponding categories

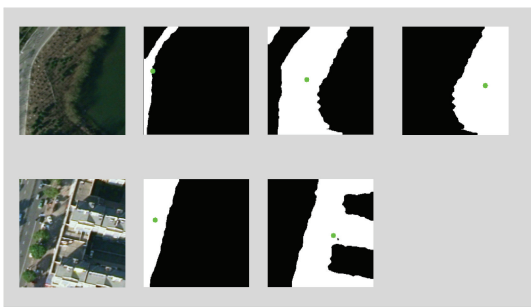

图 8　获取各个分割区域的质心位置

Fig. 8　Get the centroid position of each segmented region

为验证 GRSRC 模型的有效性与可行性,本节将GRSRC 模型与遥感图像描述领域一些具有代表性的模型进行了比较,比较方法包括 mRNN、mLSTM、mGRU、DFEN、Soft attention and Hard attention、SA。这些模型的细节描述如下。

mRNN、mLSTM、mGRU:3 种方法都使用 VGG-16 作为编码器,但使用不同的 RNN(初始 RNN、LSTM 和GRU)作为解码器。

Soft attention and Hard attention:两个模型的基本框架都是 VGG-16+LSTM。软注意力根据隐藏状态对不同的图像特征部分进行加权,而硬注意力则对不同的特征部分进行采样并通过强化学习进行优化。

DFEN:以 ResNet 作为编码器,双层 LSTM 作为解码器,并加入了特征增强模块。

SA:以 ResNet 作为编码器,LSTM 作为解码器,并加入了细粒度结构化注意力机制。

表 3 展示了 RSICD 数据集上不同网络模型在相同评价标准下的实验数据结果。由表 3 可知 GRSRC 模型在BLEU-1 评价指标上得分达到 79.49,相比 DFEN 的76.60 提升了 2.89%,在 BLEU-2 评价指标上得分达到63.67,相比 DFEN 的 63.60 提升了 0.7%,在 ROUGE_L 评价指标上达到 68.61,相比 DFEN 的 68.50 提升了1.1%。在 METEOR 评价指标中与最优的网络模型相差1.87%,而在 BLEU-3、BLEU-4 指标中略显不足。

表 3　基于 RSICD 数据集的准确率

Table 3　Accuracy based on RSICD dataset　(%)

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
mRNN	45.58	28.25	18.09	12.13	15.69	31.26
mLSTM	50.57	32.42	23.19	17.46	17.84	35.02
mGRU	42.56	29.99	22.91	17.98	19.41	37.97
Soft-Attention	67.53	53.08	43.33	36.17	32.55	61.09
Hard-Attention	66.69	51.82	41.64	34.07	32.01	60.84
DFEN	76.60	63.60	**53.80**	**46.30**	**37.30**	68.50
SA	70.16	56.14	46.48	39.34	32.91	57.06
GRSRC	**79.49**	**63.67**	44.19	37.13	35.43	**68.61**

表 4 展示了在 UCM 数据集上的实验数据结果,可以看出 GRSRC 表现出了较为良好的性能,在 BLEU-1、METEOR、ROUGE_L 指标上相较于表现较为优秀的 DFEN 模型高 2.32%,1.15% 和 1.88%,而在 BLEU-2、BLEU-3 和 BLEU-4 中,相较于性能较好的 SA 模型低2.72%,5.49%,8.19%。

表 4　基于 UCM 数据集的准确率

Table 4　Accuracy based on UCM dataset　(%)

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
mRNN	60.10	50.70	32.80	20.80	19.30	
mLSTM	63.50	53.20	37.50	21.30	20.30	
mGRU	42.56	29.99	22.91	17.98	19.41	37.97
Soft-Attention	74.54	65.45	58.55	52.50	38.86	72.37
Hard-Attention	81.57	73.12	67.02	61.82	42.63	76.98
DFEN	85.10	78.40	72.80	67.70	45.90	80.50
SA	85.38	**80.35**	**75.72**	**71.49**	46.32	81.41
GRSRC	**87.42**	77.63	70.23	63.30	**47.05**	**82.38**

表 5 展示了在 Sydney 数据集上的实验数据结果，可以看出 GRSRC 依然保持了较好的性能，在 BLEU-1、BLEU-2、METEOR、ROUGE_L 指标上相较于表现较为优秀的 SA 模型高 6.11%，1.75%，4.81%，4.02%，而在 BLEU-3 和 BLEU-4 中，相较于性能较好的 SA 模型低 2.05%、4.29%。

表 5　基于 Sydney 数据集的准确率

Table 5　Accuracy based on Sydney dataset　(%)

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
mRNN	51.30	37.50	20.40	19.30	18.50	
mLSTM	54.60	39.50	22.30	21.20	20.50	
mGRU	69.64	60.92	52.39	44.21	31.12	59.17
Soft-Attention	73.22	66.74	62.23	58.20	39.42	71.27
Hard-Attention	75.91	66.10	58.89	52.58	38.98	71.89
DFEN	79.80	69.70	61.40	54.20	37.30	72.30
SA	77.95	70.19	**63.92**	**58.61**	39.54	72.99
GRSRC	**84.06**	**71.94**	61.87	54.32	**44.35**	**77.01**

通过上述表格的对比结果，可以表明：

1）GRSRC 模型在 RSICD、UCM 和 Sydney 3 个遥感图像描述通用数据集的实验结果较稳定，说明了模型在遥感图像描述领域具有一定的泛用性和鲁棒性。

2）GRSRC 模型在 BLEU-1、BLEU-2、METEOR、ROUGE_L 指标上表现相对较好，在其他指标上表现略显不足。其中，在 BLEU-1、BLEU-2 上表现较好而在更高阶的 BLEU-n 上表现较差说明模型更倾向于生成准确的单词短语，而对句子结构、词语的组合不够好。这可能是由于采用 LSTM 结构导致的，相比于 Transformer 的全局注意机制，LSTM 更注重于词汇间的时序关系，而非利用全局注意力优化句子结构。

3）在 METEOR、ROUGE_L 指标上模型的表现较好，也说明模型能够在某种程度上理解图像并产生与参考描述相似的词语，说明分割的标签以及地理对象间的空间关系起到了提高准确率的作用。

图 9 中展示了一些在 RSICD 数据集上的描述结果，可以看到本文的模型更倾向于产生区域信息的描述和方向信息的描述。

图 9　基于 RSICD 数据集的描述结果

Fig. 9　Captioning results of the GRSRC method on the RSICD dataset

3　结　论

针对现有遥感图像描述模型只能生成粗粒度的注意力区域，无法获取遥感对象之间的地理关系，不能充分利用遥感图像语义内容的问题，引入 GORSA 模块增强主干网络对结构性特征的提取能力，提升了模型对地理空间对象关系的理解。通过在 RSICD、UCM、Sydney 3 个数据集上的各项对比试验表明，GRSRC 模型在 BLEU、ROUGE-L 和 METEOR 评价指标上有一定提升，例如在 UCM 数据集上，BLEU-1 达到了 84.06、METEOR 达到了 44.35、ROUGE_L 达到了 77.01。说明 GRSRC 模型描述较为准确且能够凸显遥感图像对象以及空间关系，一定程度上解决了遥感对象空间关系利用不充分、遥感图像语义内容利用不准确的问题，进一步增强了遥感图像描述模型在城市管理、交通管控等方面的实用性。

参考文献

[1] ORIOL V, ALEXANDER T, SAMY B, et al. Show and tell: A neural image caption generator [C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 3156-3164.

[2] CHEN X, ZITNICK C L. Learning a recurrent visual

representation for image caption generation [J]. arxiv preprint arxiv:1411.5654, 2014.

[3] LI Y, CHEN R, ZHANG Y, et al. Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network[J]. Remote Sensing, 2020, 12(23): 4003.

[4] 陈燕,杨艳,杨春兰,等. 基于阶段聚焦损失和并行增广策略的遥感图像场景分类[J]. 电子测量与仪器学报, 2023, 37(1):116-122.
CHEN Y, YANG Y, YANG CH L, et al. Remote sensing image scene classification via stage-based focal loss and parallel data augmentation strategy [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(1):116-122.

[5] YE J, HE J J, PENG X J, et al. Attention-driven dynamic graph convolutional network for multi-label image recognition [C]. Computer Vision – ECCV 2020: 16th European Conference, Springer, UK, Glasgow, 2020: 649-665.

[6] YE Y, REN X, ZHU B, et al. An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images[J]. Remote Sensing, 2022, 14(3): 516.

[7] QIAN X, LIN S, CHENG G, et al. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion[J]. Remote Sensing, 2020, 12(1): 143.

[8] MOU L C, ZHU X X. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images [EB/OL]. [2018-05-05]. http://arxiv.org/abs/ 1805.02091.

[9] PAN X, GAO L, MARINONI A, et al. Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network[J]. Remote Sensing, 2018, 10(5): 743.

[10] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]. Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[11] XU K, JIMMY B, RYAN K, et al. Show, attend and tell: Neural Image caption generation with visual attention [C]. International Conference on Machine Learning, 2015: 2048-2057.

[12] ANDREJ K, LI F F. Deep visual-semantic alignments for generating image descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 39: 664-676.

[13] QU B, LI X L, TAO D CH, et al. Deep semantic understanding of high resolution remote sensing image [C]. 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), IEEE, 2016: 1-5.

[14] 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, 2017, 56: 2183-2195.

[15] ZHANG X, LI Y, WANG X, et al. Multi-source interactive stair attention for remote sensing image captioning[J]. Remote Sensing, 2023, 15(3): 579.

[16] HITESH K, SUDIPAN S, BIPLAB B, et al. Exploring transformer and multilabel classification for remote sensing image captioning [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[17] REN Z, GOU S, GUO Z, et al. A mask-guided transformer network with topic token for remote sensing image captioning [J]. Remote Sensing, 2022, 14(12): 2939.

[18] LIU CH Y, ZHAO R, SHI ZH X. Remote-sensing image captioning based on multilayer aggregated transformer [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[19] 农元君,王俊杰. 基于注意力和强化学习的遥感图像描述方法[J]. 光学学报, 2021, 41(22):206-214.
NONG Y J, WANG J J. Remote sensing image caption method based on attention and reinforcement learning [J]. Acta Optica Sinica, 2021, 41(22):206-214.

[20] ZHAO R, SHI ZH W, ZOU ZH X. High-resolution remote sensing image captioning based on structured attention [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[21] CHEN J, HAN Y R, WAN L, et al. Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network [J]. International Journal of Remote Sensing, 2019, 40: 6482-6498.

[22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[23] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [J]. arXiv preprint arXiv:

1706. 05587, 2017.

[24] ZHAO W H, YANG W ZH, CHEN D, et al. DFEN: Dual feature enhancement network for remote sensing image caption [J]. Electronics, 2023, 12(7): 1547.

[25] XIA G S, HU J W, HU F, et al. AID: A benchmark data set for performance evaluation of aerial scene classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.

[26] YANG Y, NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010: 270-279.

[27] ZHANG F, DU B, ZHANG L P. Saliency-guided unsupervised feature learning for scene classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 53(4): 2175-2184.

作者简介

李国燕,2006 年于河北师范大学获得学士学位,2009 年于河北工业大学获得硕士学位,2013 年于河北工业大学获得博士学位,现为天津城建大学副教授,硕士生导师,主要研究方向为机器视觉、下一代网络技术。

E-mail: ligy@ tcu. edu. cn

Li Guoyan received her B. Sc. degree from Hebei Normal University in 2006, her M. Sc. degree from Hebei University of Technology in 2009 and her Ph. D. degree from Hebei University of Technology in 2013. She is currently an associate professor M. Sc. supervisor at Tianjin Chengjian University. Her main research interests include machine vision and next-generation network technologies.

田明达,2021 年于天津城建大学获得学士学位,现为天津城建大学计算机与信息工程学院硕士研究生,主要研究方向为遥感图像描述。

E-mail: 1210062788@ qq. com

Tian Mingda received his B. Sc. degree from Tianjin Urban Construction University in 2021. He is currently a M. Sc. candidate in the School of Computer and Information Engineering at Tianjin Urban Construction University. His main research interest includes remote sensing image description.

董春华,2009 年于天津城建大学获得硕士学位,现为天津城建大学讲师,主要研究方向为 GIS、遥感信息分析。

E-mail: dch@ tcu. edu. cn

Dong Chunhua received her M. Sc. degree from Tianjin Urban Construction University in 2009. She is currently a lecturer at Tianjin Urban Construction University. Her main research interests include GIS and remote sensing information analysis.

郝志鹏(通信作者),2002 年于天津职业技术师范学院获学士学位,2010 年于天津大学获硕士学位,现为天津城建大学讲师,主要研究方向为脑波信息分析处理。

E-mail: hzpqina@ 163. com

Hao Zhipeng (Corresponding author) received his B. Sc. degree from Tianjin Vocational and Technical Normal University in 2002 and his M. Sc. degree from Tianjin University in 2010. He is currently a lecturer at Tianjin Urban Construction University. His main research interests include brainwave information analysis and processing.