引文格式：王振宇,万刚,王孜妍. 基于CLIP模型和迁移学习的遥感图像描述方法[J]. 地理空间信息, 2025, 23(9): 1-5. doi:10. 3969/j.issn. 1672-4623.2025.09.001

基于CLIP模型和迁移学习的遥感图像描述方法

王振宇[1]，万　刚[1*]，王孜妍[1]
（1. 航天工程大学 航天信息学院，北京 101416）

摘　要：遥感图像描述是指根据遥感图像生成合理化的自然语言描述，是计算机视觉、自然语言处理之间的交叉领域任务。该任务针对不同类型的地物需要大量异质的图像和描述文本。为降低对数据、计算资源的需求，提出了一种遥感图像描述方法，利用微调后的视觉大模型CLIP对遥感图像进行特征编码，构建一个高效的遥感领域的视觉—语言编码器，再配合Transformer架构的解码器生成图像描述。在3个公开遥感图像描述数据集上的实验结果表明，该方法能明显提升生成图像描述语句的质量，具有一定的竞争力。

关键词：遥感图像描述；CLIP模型；迁移学习；Transformer架构
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2025）09-0001-05

Remote Sensing Image Caption Method Based on CLIP Model and Transfer Learning

WANG Zhenyu[1], WAN Gang[1], WANG Ziyan[1]
(1. School of Space Information, Space Engineering University, Beijing 101416, China)

Abstract: Remote sensing image caption means to generate rational natural language caption based on remote sensing image, which is a cross-domain task between computer vision and natural language processing. However, this task requires mass heterogeneous images and caption text for different types of ground objects. In order to reduce the requirements on data and computing resources, we proposed a remote sensing image caption method. We used the fine-tuned visual model CLIP as the encoder of remote sensing images, and successfully constructed an efficient vision language encoder in the field of remote sensing. Then, we used the decoder of Transformer architecture to generate image caption. Experimental results based on three public remote sensing image caption datasets show that the proposed method significantly improves the quality of generated image caption, and is quite competitive with other mainstream methods.

Key words: remote sensing image caption, CLIP model, transfer learning, Transformer architecture

遥感技术的迅速发展推动了遥感领域的应用研究，而遥感图像描述旨在让机器描述人类语境中遥感图像的内容，使计算机拥有“看图说话”的能力[1]。不同于遥感图像分割[2]、目标检测[3]和地物分类[4]等任务，遥感图像描述的目的是生成全面的句子。为生成包含丰富语义的描述句，遥感图像描述需要准确识别图像中包含的不同层次的地面元素，如机场、港口、舰船、车辆、桥梁、水体、植被等，分析其属性、隶属层级和空间依赖关系并生成可理解的自然语言，在军事侦察、抢险救灾、环境监测、城市规划等军民用领域具有广泛的应用前景。

传统遥感图像描述方法将任务划分为编码（图像特征提取）、解码（描述语句生成）两个阶段，通常在多样性较小的小样本数据集上进行，有限的数据集不足以使模型拟合[5]；且RNN网络难以并行化训练，训练时间较长，自身长程记忆能力有限也使其存在性能瓶颈[6]。针对上述不足，本文提出的方法的贡献在于：①通过迁移学习，将预训练的视觉大模型CLIP[7]作为编码器，有效降低模型训练耗费的计算资源和时间；②采用遥感图像分类数据集对模型编码器进行微调，成功构建了一个能识别遥感图像地理场景基础特征的编码器，进一步提升在小样本数据集学习情景上的效果；③采用Transformer[8]结构的解码器，配合CILP结构的编码器生成描述。本文通过RSICD等3个公开遥感图像描述数据集证明了本文方法的有效性，充分揭示了“大模型+迁移学习”方法在遥感图像处理领域的潜力。

早期图像描述因过多依赖前期的视觉处理过程，对于描述生成的模型优化有限，因此难以生成高质量的描述。近年来受机器翻译领域中编码器—解码器架

第一作者简介：王振宇（1997—），硕士在读，研究实习员，研究方向为遥感图像处理、多传感器数据融合与应用，E-mail：1061025786@qq.com。
通信作者：万刚（1972—），博士，教授，研究方向为虚拟地理环境和时空大数据，E-mail：casper_51@163.com。

构的启发，图像描述可通过端到端的学习方法将描述过程转化为图像到描述的“翻译”过程。Xu等[9]首先将注意力机制引入图像描述，再针对图像描述提出了各种注意力机制。注意力机制的应用可进一步提高图像描述生成文本的流畅性和相关性。虽然自然图像描述中的许多方法可应用于遥感图像描述研究，但需解决一些挑战：①自然图像通常有一个清晰的观察方向，而遥感影像没有固定的观察方向[5]，且拍摄过程中易受云层、光线、植被等背景影响；②遥感图像中的物体尺寸变化较大，导致同一地物在不同分辨率遥感图像中的尺寸和外观差异巨大，存在尺度模糊性。针对这些挑战，部分学者[5,10-12]基于编码器—解码器框架，对模型架构进行优化，并根据遥感图像的时空特点设计新的注意力机制，虽然大幅提升了遥感图像的描述性能，但大多仅致力于改进或增强遥感图像的特征编码性能，较少针对当前遥感图像描述任务的小样本学习情景进行研究。

OpenAI在2021年发布的CLIP模型[7]，是一个是以文本为监督信号进行训练的可迁移视觉模型。其训练数据集是从互联网收集的4亿图像—文本对。CLIP模型包括Image Encoder和Text Encoder两个子模块，分别用于提取图像和文本特征。通过对比学习图文特征，CLIP模型实现了在诸多任务上的零样本和少样本分类能力。对比零样本CLIP模型和有监督的ResNet50在27个数据集上的表现发现[7]，在16个数据集上CLIP可超过ResNet50，具有相当不错的性能。充分利用CLIP模型的跨视觉—语言检索能力，并结合迁移学习实现遥感图像和自然语言描述在语义空间的高效对齐是本文方法的初衷。

Transformer是一种用于seq2seq任务的编码器—解码器模型[8]，被广泛应用于自然语言处理、图像识别、语音识别等深度学习任务中。相较于LSTM，Transformer模型完全依靠注意力机制对输入输出序列的全局依赖关系进行建模，易于并行化计算；且无需按顺序处理每个单词，一次性处理整个序列并创建注意力矩阵，拥有更强的记忆能力。Liu等[13]首先提出基于双分支Transformer的方法对双时相的遥感图像进行变化描述，达到了SOTA的成绩。因此，本文方法采用Transformer架构的解码器。

1 研究方法

1.1 方法流程

本文的任务可定义为：给定图像描述数据集 $\{x^i, c^i\}_{i=1}^N$，x 为图像，c 为描述句，N 为数据集中图像—描述对数量。将 c 表示为词的序列，即 $c^i = c_1^i, \cdots, c_l^i$，为便于并行计算，将词序列填充至描述文本最大长度 l。针对给定的图像 x，使生成的目标词序列的可能性最大，θ 为本文方法模型的参数，则训练目标可表示为：

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(c_1^i, \cdots, c_l^i | x^i) \tag{1}$$

首先以CLIP模型的Image Encoder为编码器对图像进行特征提取，再将特征转换至离散的词向量空间，最后通过Transformer架构的解码器将向量解码为描述文本（图1）。

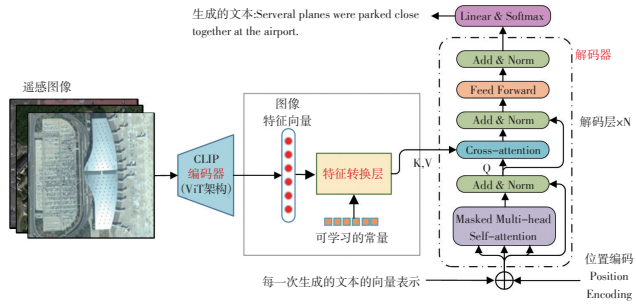


图1 基于CLIP模型和迁移学习的遥感图像描述方法流程图

1.2 编码器

1）图像编码。CLIP模型的Image Encoder采用多种架构，其中基于Transformer的ViT[14]架构表现较优，因此本文采用ViT架构对图像进行特征编码。对于图像 x^i，利用Image Encoder将图像转换为特征向量 $\boldsymbol{x}'^i \in \boldsymbol{R}^{1\times k}$，$k$ 为图像特征维度数，$\boldsymbol{x}'^i = \boldsymbol{x}'^i_1, \cdots, \boldsymbol{x}'^i_k = \mathrm{CLIP}(x^i)$。

2）文本编码。由于图像编码生成特征向量维度数为 k，因此将词嵌入（WE）维度数设置为 k，并将全部描述文本补长至 l，得到 $WE(c^i) \in \boldsymbol{R}^{1\times l\times k}$。另外，单词的顺序富含大量语义信息，根据参考文献[8]，本文采用正、余弦函数对文本进行位置编码。

1.3 特征转换层

特征转换层将编码器提取的图像特征转换至词向量所在的空间，包含 N 个自注意力模块。首先利用全连接层和矩阵变换将 $\boldsymbol{x}'^i$ 转换至词向量所在空间的维度 $\boldsymbol{R}^{1\times l\times k}$，该操作记为 f；再利用可学习的向量 $\mathrm{const} \in \boldsymbol{R}^{1\times l\times k}$ 与 $f(\boldsymbol{x}'^i)$ 拼接得到 $\boldsymbol{F}^i \in \boldsymbol{R}^{1\times 2l\times k}$，便于后续利用注意力机制检索图像特征向量中的特征信息，即 $\boldsymbol{F}^i = \mathrm{Concat}(\mathrm{const}, f(\boldsymbol{x}'^i))$。自注意力模块采用层归一化（$N$）和多头注意力机制（$M$），$M$ 可获得多个表示子空间，从多角度关注输入序列的特征，Q、K、V 分别为 $N(\boldsymbol{F}^i)$ 与可学习的参数 W_q、W_k、W_v 相乘。利用两个全连接层和ReLU激活函数将特征向

量转换至文本编码所在的维度 $\boldsymbol{R}^{1\times l\times k}$。令 σ 为Softmax函数，L 为全连接层，d 为 Q、K、V 的维度，I^i 为输出，则上述过程可表示为：

$$h^i=M(Q,K,V)=\sigma\left(\frac{Q\times K^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{2}$$

$$o^i=N(F^i+h^i) \tag{3}$$

$$I^i=L(\mathrm{ReLU}(L(o^i)))+o^i \tag{4}$$

1.4　解码器

解码器采用每次生成的文本向量 y^i 和来自特征转换层的 I^i 生成下一步的单词。解码器采用参考文献[8]中Transformer架构的解码器，由多个解码层堆叠构成。每个解码层包括含掩码的多头自注意力层、交叉注意力层和前馈网络层3个子层。与多头注意力不同，含掩码的多头自注意力在训练阶段预测第 t 个单词时采用掩码掩盖了位置 t 和后续位置的值，可确保预测每个单词时仅取决于已生成的单词。特别的是，交叉注意力层中的 Q 为含掩码的自注意力层的输出与 $W_q^{'}$ 的乘积，K、V 分别为 I^i 与 $W_k^{'}$、$W_v^{'}$ 的乘积。最终，解码器的输出被送至含Softmax函数的线性层来生成表示单词概率的向量 $\boldsymbol{P}=[p_1,p_2,\cdots,p_L]\in\boldsymbol{R}^{L\times Z}$，$L$ 为句子长度，Z 为单词表长度（单词表为英文描述文本分词构建）。单词概率向量 $\boldsymbol{p}_t=[p_t^{(1)},p_t^{(2)},\cdots,p_t^{(Z)}]$ 用于推测生成句子位置 t 处的单词，$\boldsymbol{y}_t=[y_t^{(1)},y_t^{(2)},\cdots,y_t^{(Z)}]$ 为线性层输出的向量。$\boldsymbol{p}_t$ 的生成过程可表述为：

$$\boldsymbol{p}_t=\sigma(\boldsymbol{y}_t)=\frac{\exp(\boldsymbol{y}_t)}{\sum_{i=1}^{Z}\exp(\boldsymbol{y}_t^{(i)})} \tag{5}$$

2　实验情况与数据分析

2.1　数据集和评价指标

描述数据集采用遥感图像描述领域常用的3个公开数据集RSICD[5]、Sydney-captions[10]和UCM-captions[10]。RSICD包含30类场景共10 921张图像，均为224像素×224像素；Sydney-captions包含7类场景共613张图像，均为500像素×500像素，分辨率为0.5 m；UCM-captions包含21类场景共2 100张图像，均为256像素×256像素，分辨率为0.3 m。3个数据集对每个图像给出5个描述参考句，本文按数据集原始划分，80%用于训练，10%用于验证，剩下10%用于测试。分类数据集采用NWPU-RESISC45[15]，包含21类场景共31 500张图像，均为256像素×256像素，用于微调CLIP模型的编码器。评价指标采用BLEU[16]、METEOR[17]、ROUGE_L[18]和CIDEr[19]4种指标。BLEU最初由IBM公司于2002年提出，主要应用于机器翻译，即针对给定机器生成的一个翻译，自动计算一个分数衡量机器翻译的好坏，取值范围为0~1。其中n-gram表示每次取句子里的 n 个连续单词的集合，n 取值1~4，分别对应BLEU-1到BLEU-4，本文采用BLEU-1和BLEU-4。BLEU的使用便捷，但并不涉及同义词和语法准确性。METEOR是基于BLEU的拓展，基于单精度的加权调和平均数和单字召回率，并利用WordNet等知识源扩充了同义词集，与BLEU相比，METEOR考虑了更多的因素，如同义词匹配、词干匹配、词序等。该指标考虑了整个语料库的准确率和召回率，最终得出测度。ROUGE主要应用于文本自动摘要，专注于召回率而非精度。本文采用其中的ROUGE-L，L为最长公共子序列的首字母，通过将BLEU的n-gram优化为最长公共子序列，计算时利用待评估句 c 和参考句 s 的最长公共子序列来生成相似度。该指标重点查看有多少个参考句的n-gram出现在待评估句中，但并不考量待评估句整体是否流畅通顺。CIDEr是专门为图像描述定制的指标，将BLEU与向量空间模型相结合，将句子作为文档处理，对每个n-gram进行TF-IDF[20]权重计算，并进一步计算 c 和 s 的余弦距离来衡量相似度。若一些n-gram频繁出现在描述图像的参考句中，TF对于这些n-gram将给出更高的权重，而IDF则降低那些在所有描述语句中都高频出现的n-gram的权重。不同的n-gram随着TF-IDF不同而具有不同的权重，因为整个语料里更常见的n-gram显然是包含了更小的信息量。

2.2　实验设置

1）损失函数。训练过程的损失函数采用经典的交叉熵损失函数，可表述为：

$$\mathrm{Loss}=-\sum_{t=1}^{L}\log\left(\sum_{z=1}^{Z}\tilde{\boldsymbol{y}}_t^{(z)}\boldsymbol{p}_t^{(z)}\right) \tag{6}$$

式中，L 为解码器生成描述句的长度；Z 为单词表长度；$\tilde{\boldsymbol{y}}_t^{(z)}$ 为模型解码器输出描述句在位置 t 处的单词向量中的第 z 个单词；$\boldsymbol{p}_t^{(z)}$ 为解码器输出描述句在位置 t 处的单词概率向量中的第 z 个单词。

2）微调编码器。本文选取CLIP模型ViT[14]架构下的ViT-B/16、ViT-B/32、ViT-L/14编码器模型。模型L的参数量大于模型B，数字代表对图像划分Patch的尺寸，如ViT-B/16表示将输入像素为224×224的图片按照16×16大小的Patch进行划分。采用上述3种Image Encoder在NWPU-RESISC45[15]数据集上进行分类训练。数据集按照原设定的6∶1进行划分，训练迭

代次数（epoch）设置为10，初始学习率为0.000 2，结果见表1，可以看出，利用遥感图像分类数据集进行微调，CLIP模型的编码器在训练5 epoch后，分类正确率均能达到93%以上，训练10 epoch后，正确率均约为95%。此时，默认编码器已具备提取遥感图像各类地理场景浅层基础特征和深层抽象特征的能力，可进一步应用于遥感图像描述任务。

表1 NWPU-RESISC45分类数据集训练编码器的结果/%

CLIP模型	训练前分类正确率	训练5 epoch分类正确率	训练10 epoch分类正确率
ViT-B/16	53.11	96.15	94.86
ViT-B/32	47.24	94.93	94.95
ViT-L/14	59.51	93.88	96.55

3）实现细节。模型均在NVIDIA RTX 3090 GPU上进行训练和评估，设置最大epoch为100，初始学习率为0.000 2，并利用AdamW优化器和cosine_schedule_with_warmup学习率调整策略进行优化。每训练1 epoch，在验证集上计算模型的4个评价指标，最后选择在验证集上表现最佳的模型在测试集上进行评估。测试时采用Beam Search算法，Beam Size设置为5。

2.3 结果分析

2.3.1 对比实验

为验证本文方法的有效性，采用3个遥感图像描述数据集的测试集对模型进行评估，并与性能较好的其他主流遥感图像描述方法进行对比，结果见表2~4，可以看出，本文方法在部分指标上取得了更高的分数。在所有的训练模型中，ViT-L/14的参数量最大，而各评价指标分数却并非最高，可能是受限于本文方法实验环境的显存，设定的Batch Size较小，模型欠拟合。

表2 RSICD上不同模型对比实验结果

方法	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr
Soft-att[5]	0.629	0.292	0.254	0.470	0.788
CCSMLF[10]	0.576	0.222	0.213	0.446	0.530
Scene-att[12]	0.625	0.297	0.253	0.474	0.809
ViT-B/16	0.649	0.309	0.253	0.495	0.861
ViT-B/32	0.644	0.304	0.252	0.490	0.849
ViT-L/14	0.656	0.325	0.262	0.503	0.897

表3 Sydney-captions上不同模型对比实验结果

方法	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr
Soft-att[5]	0.744	0.542	0.380	0.676	2.323
CCSMLF[10]	0.444	0.241	0.158	0.402	0.938
Scene-att[12]	0.786	0.561	0.381	0.709	2.505
ViT-B/16	0.746	0.501	0.350	0.662	2.045
ViT-B/32	0.758	0.578	0.387	0.698	2.323
ViT-L/14	0.789	0.563	0.383	0.707	2.376

表4 UCM-captions上不同模型对比实验结果

方法	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr
Soft-att[5]	0.826	0.649	0.417	0.761	3.092
CCSMLF[10]	0.387	0.091	0.095	0.360	0.370
Scene-att[12]	0.822	0.674	0.440	0.778	3.228
ViT-B/16	0.865	0.715	0.463	0.818	3.368
ViT-B/32	0.851	0.706	0.453	0.805	3.374
ViT-L/14	0.856	0.735	0.476	0.812	3.551

由本文模型（ViT-B/32）在RSICD数据集上部分预测描述结果（图2）可知，通用场景中预测描述与参考描述接近；但涉及描述图像中的对象几何形状和数量属性时，模型则倾向生成概括性描述，性能相对较差。



图2 RSICD数据集部分描述结果

注：例1~3来源于RSICD: mediumresidential_96, industrial_46, storagetanks_76；例4~6来源于RSICD:port_39, airport_84, square_41。

2.3.2 消融实验

为体现微调编码器阶段对模型性能的影响，本文采用未经遥感图像分类数据集微调的ViT-B/16、ViT-B/32、ViT-L/14编码器模型，在RSICD数据集上进行实验。结果表明（表5），相较于原模型，各指标评分存在1%~4%的下降。

表5 未经分类数据集微调模型在RSICD数据集上的实验结果

方法	BLEU-1	BLEU-4	METEOR	ROUGE_L	CIDEr
ViT-B/16	0.635	0.295	0.248	0.487	0.819
ViT-B/32	0.627	0.286	0.241	0.477	0.800
ViT-L/14	0.646	0.302	0.250	0.490	0.864

3 结 语

本文提出了基于CLIP模型和迁移学习的遥感图像

描述方法，通过遥感图像分类数据集对预训练的CLIP模型进行微调，构建了高效的遥感图像特征编码器，可在当前遥感图像描述领域小样本数据集场景下获得较好的效果；再基于注意力机制，设计了图像—文本的特征转码层，以平衡有效的方式实现了视觉—语言空间的对齐；最后采用Transformer架构的解码器生成文本描述。在3个公开遥感图像数据集上的实验结果证明，本文方法可在数据和计算资源有限的情况下取得较好结果，且在多项测评指标上超过了主流方法，能有效提升遥感图像描述语句的质量。

参考文献

[1] 苗益,赵增顺,杨雨露,等. 图像描述技术综述[J]. 计算机科学,2020,47(12):149-160

[2] 袁鹏,刘芳,朱永泰,等. 一种改进的马尔可夫随机场遥感图像分割方法[J]. 地理空间信息,2024,22(6):34-38

[3] 何民华,张润达,赵胜利. 基于Att-DConv的遥感舰船检测方法研究[J]. 地理空间信息,2024,22(3):24-28

[4] 刘琦,杨小华,陈柏行,等. 基于Transformer的成都市地物分类技术[J]. 地理空间信息,2023,21(12):51-53

[5] Lu X, Wang B, Zhang X, et al. Exploring Models and Data for Remote Sensing Image Caption Generation[J]. IEEE Transactions on Geoscience and Remote Sensing,2017(12): 2 183-2 195

[6] 张孝慈. 递归神经网络模型的若干关键问题研究[D]. 合肥:中国科学技术大学,2019

[7] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models from Natural Language Supervision[C]// Proceedings of the International Conference on Machine Learning,2021:8 748-8 763

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C]. Neural Information Processing Systems , 2017: 5 998-6 008

[9] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille,2015:2 048-2 057

[10] Qu B, Li X, Tao D, et al. Deep Semantic Understanding of High Resolution Remote Sensing Image[C]. International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, 2016

[11] Wang B, Lu X, Zheng X, et al. Semantic Descriptions of High-resolution Remote Sensing Images[J]. IEEE Geoscience and Remote Sensing Letters,2019(2):1 274-1 278

[12] Wu S, Zhang X, Wang X, et al. Scene Attention Mechanism for Remote Sensing Image Caption Generation[C]. International Joint Conference on Neural Networks (IJCNN), Glasgow, 2020

[13] Liu C, Zhao R, Chen H, et al. Remote Sensing Image Change Captioning with Dual-branch Transformers: A New Method and a Large Scale Dataset[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60:1-20

[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[C]. International Conference on Learning Representations, 2020

[15] Cheng G, Han J, Lu X. Remote Sensing Image Scene Classification: Benchmark and State of the Art[J]. Proceedings of the IEEE, 2017, 105(10):1 865-1 883

[16] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics, 2002:311-318

[17] Denkowski M, Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language[C]// Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014:26-27

[18] Lin C Y. ROUGE: A Package for Automatic Evaluation of Summaries[C]//Proceedings Text Summarization Branches Out, 2004:74-81

[19] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based Image Description Evaluation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015:4 566-4 575

[20] Robertson S. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF[J]. Journal of Documentation, 2004(7):503-520