INSTANCE-AWARE REMOTE SENSING IMAGE CAPTIONING WITH CROSS-HIERARCHY ATTENTION

Chengze Wang, Zhiyu Jiang, Yuan Yuan*

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China

ABSTRACT

The spatial attention is a straightforward approach to enhance the performance for remote sensing image captioning. However, conventional spatial attention approaches consider only the attention distribution on one fixed coarse grid, resulting in the semantics of tiny objects can be easily ignored or disturbed during the visual feature extraction. Worse still, the fixed semantic level of conventional spatial attention limits the image understanding in different levels and perspectives, which is critical for tackling the huge diversity in remote sensing images. To address these issues, we propose a remote sensing image caption generator with instance-awareness and cross-hierarchy attention. 1) The instances awareness is achieved by introducing a multi-level feature architecture that contains the visual information of multi-level instance-possible regions and their surroundings. 2) Moreover, based on this multi-level feature extraction, a cross-hierarchy attention mechanism is proposed to prompt the decoder to dynamically focus on different semantic hierarchies and instances at each time step. The experimental results on public datasets demonstrate the superiority of proposed approach over existing methods.

Index Terms— Remote sensing image captioning, semantic understanding, visual attention

1. INTRODUCTION

Conventional remote sensing image analysis tasks usually focus on the object-level or pixel-level understanding, such as object classification, change detection, and image segmentation. Despite that all mentioned tasks have gained massive success and been deployed in many industrials, the lack of comprehensive description to high-level semantics limits the expansion of remote sensing applications. To access a comprehensive description of global semantic information contained in the captured scene intuitively, image captioning [1, 2, 3] is introduced to the remote sensing field, which generates sentences in human language to summarize the high-level semantic content from remote sensing images.

Image captioning is an interdisciplinary task emerged from the overlap of computer vision and natural language processing, and has aroused great attention from the community. It is complicated to accomplish automatically, since it not only requires a global understanding of objects and their relations, but also needs to transform the visual content into flexible and fluent sentences. Early image caption approaches were mainly template-based models [4] and retrial-based models [5], have been replaced by encoder-decoder based methods [6] for better performance and flexibility. Later on, to further enhance the correlation between the visual information and generated words during the encoding-decoding progress, the visual attention mechanism [7] was explored. Nowadays, most of image captioning approaches are designed based on the attention-enhanced encoder-decoder architecture.

When comes to remote sensing image captioning, there are several published works that have explored this field in recent years. The first attempt was made by Qu *et al.* [8] who utilized an encoder-decoder framework based on multimodal neural networks to prove the possibility of generating human sentences for remote sensing images. The attention mechanism for remote sensing image was firstly introduced in [1], and found out that the spatial attention is of crucial importance for captioning performance. Moreover, a multi-scale cropping mechanism [2] was designed to adapt different size instances in the image. Lately, Zhang *et al.* [3] introduced the classification label of remote sensing images into attention mechanism, which improved the performance considerably.

Though previous remote sensing image captioning methods have gained remarkable progress, there are still existing limitations as follows: 1) Mostly, the spatial attention is computed on the feature map with a low spatial-wise resolution grid, which leads to that the encoder is difficult to distinguish content with accurate, clear boundaries. At the same time, this grid-based attention is difficult to focus on tiny or irregular ground targets, causing the attended semantic feature usually containing quite a lot irrelevance visual information. 2) The attending standpoint is fixed to a particular scale and se-

*Corresponding author: Zhiyu Jiang (e-mail: zhiyu.jiang.chn@gmail.com)

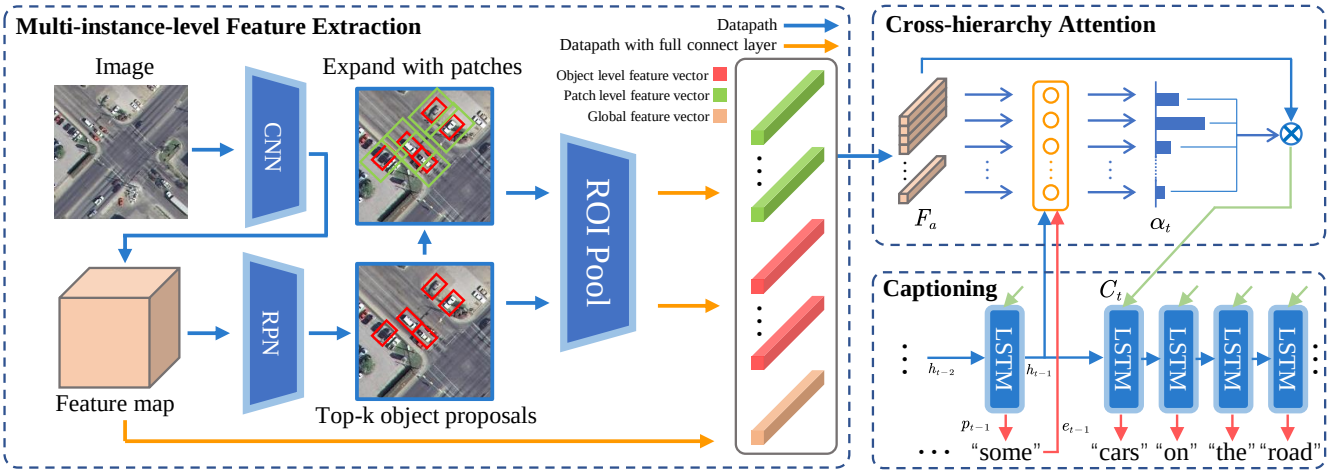

Fig. 1. The overall architecture of the proposed method. The model takes remote sensing images to the multi-instance-level feature extraction as encoding, then outputs descriptive sentences with the cross-hierarchy attention mechanism.

mantic level. However, in remote sensing image captioning, part of captured scenes are with extremely brief (*e.g.* ocean, desert) or extremely complex semantic information (*e.g.* city park, highway), and the fixed scale and semantic level will make it difficult to obtain a reasonable attention distribution.

Considering the limitations mentioned above, an instance-aware remote sensing image captioning method with cross-hierarchy attention is proposed. The contributions of this paper are summarized as follows:

1) Instance-awareness is for the first time realized in remote sensing image captioning. It is a more straightforward way to distinguish semantic information with the ground objects and their relations. The extracted object feature together with the neighboring patch feature and global feature constitutes the multi-instance-level feature, which improves the accuracy and possibility of feature attention at the spatial and semantic hierarchies.

2) We propose a cross-hierarchy attention mechanism to accommodate the multi-level feature inputs, prompting the decoder to dynamically focus on different semantic levels and different instances. This enhances the flexibility on semantics and scale when facing extreme remote sensing scenes.

The evaluation on mainstream datasets demonstrates the effectiveness of the proposed method.

2. METHODOLOGY

2.1. Multi-instance-level Feature Extraction

Following the basic encoder-decoder architecture, the framework of our proposed approach is shown in Fig.1. As the encoder part of the whole image caption generator, a multi-instance-level feature extractor is designed. These encoders simultaneously aggregate 3 different levels of instance visual feature, including the object-level feature, the patch-level feature and the global feature. After extraction, these features will be fed into the cross-hierarchy attention mechanism, achieving a more accurate and reasonable spatial focus on the feature when predicting words.

2.1.1. *Object-level Feature Extraction*

For describing most scenarios, salient ground objects are the key to description generation. To accurately locate salient targets in a remote sensing image, we select the top-n ROIs nominated by the Faster RCNN [9] as n possible regions of key objects. To minimize the gap between different datasets, the class-agnostic pre-trained model is adopted. ROI-pooling and full connect layers are applied to each proposed ROI so that the output feature vector sequence F_o has the same dimension for each object: $F_o = [F_o^1, F_o^2, \ldots, F_o^n]$.

2.1.2. *Patch-level Feature Extraction*

Remote sensing image caption is highly relative to the main objects and their spatial relations, which means the semantic information of the neighboring part of one salient object is as important as the object itself. On the other hand, a larger perception field enhances the robustness of visual feature extraction, especially for the inaccuracy caused by the region proposal module. In our approach, the patch of each object is defined as the rectangular area that has the same ratio, center, and direction of the responding object, but the scale of the patch is determined by a scaling-factor k. If the patch bound exceeds the image edge, the patch region is cropped to fit the image size. With the same after-ward processing as the object-level ROIs, the feature of multiple patches F_p is extracted.

2.1.3. Global Feature Extraction

In scenarios lack of texture diversity (*e.g.* ocean, forest, and desert), so an independent global feature extraction branch is involved to provide a more comprehensive feature hierarchy.

Unlike other spatial-attention-based methods, our above mentioned two feature extraction levels have brought the visual information of key objects and their surroundings to the decoder, so the global feature is the feature vector output by full connection layer instead of the feature cube acquired from the end of convolutional layers, which also boosts computation efficiency of the decoder at the same time. Specifically, we utilize the output of the last full-connect layer of a Resnet-101 as the global feature, which has the same dimension of F_p and F_p, and is denoted as F_g. At last, the overall feature F_a is the stack of all three level features:

$$F_a = [F_o^1, \ldots, F_o^n, F_p^1, \ldots, F_p^n, F_g]. \tag{1}$$

2.2. Cross-hierarchy Attention and Caption Generation

The proposed cross-hierarchy attention mechanism is similar to the spatial attention mechanism, while having a higher diversity of focusable scale and semantic levels. The core idea of spatial attention is using the former state of LSTM to decide the next focus point or area on a uniform feature map grid. With the more oriented image feature as the input of each word's prediction, the performance and interpretability of image caption have been significantly improved.

Following this basic way, the cross-hierarchy attention dynamically re-weight the input multi-instance-level feature to focus on different area or the whole image at each time step. In the cross-hierarchy attention mechanism, the extracted feature vector of each instance can be regard as the feature of a cell in conventional spatial attention grid.

More specifically, given a stacked multi-level feature F_a, each of which is d dimensional, F_a is fed as a concatenation with the text feature and the hidden state on the last time step to a single layer neural network to calculate the attention score map over different levels and instances,

$$a_t = w_h^T \tanh(W_a[F_a, h_{t-1}, e_{t-1}]), \tag{2}$$

where W_a, w_h are network weights to be trained. e_{t-1} is the text feature vector, which is extracted from the last generated word by embedding its one-hot feature into same d dimensional. And the hidden state h_{t-1} comes from the LSTM outputs at last time step,

$$h_{t-1} = LSTM(C_{t-1}, h_{t-2}), \tag{3}$$

where C_{t-1} is the corresponding derived image feature.

Then, the attention distribution mask α_t is computed by a softmax layer,

$$\alpha_t^i = \exp\left(a_t^i\right) / \textstyle\sum_j^n \exp\left(a_t^j\right). \tag{4}$$

At the beginning of each generation, the attention distribution is uniformly initialized. Subsequently, based on this attention distribution α_t^i, the cross-hierarchy attention derived image feature C_t is computed as follows,

$$C_t = \textstyle\sum_{i=1}^N \alpha_t^i F_i. \tag{5}$$

At time step t, given the attended feature F_d, hidden state h_{t-1}, the probability of word prediction p_t is

$$p_t = \mathrm{softmax}\left(U_p \tanh\left(W_p\left[C_t, h_{t-1}\right] + b_p\right)\right). \tag{6}$$

At last, the t-th word of the description is chosen from the pre-embedding vocabulary vector by the maximum probability from p_t.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

We adopt all 3 remote sensing image caption datasets to our quantitative evaluation. Each dataset is divided into three parts by a default 80%:10%:10% ratio, which is for training, evaluation and test, respectively.

UCM-Captions [8] is a mid-scale remote sensing image caption dataset. This dataset is originally used for scene classification, and then extended with manual descriptions to become the most popular caption dataset. There are 21 categories of scenes and 100 images for each category, and each image contains 5 sentences as description.

Sydney-caption [8] is a relatively small-scale remote sensing image caption dataset, like UCM-Caption, it is also extended from a scene classification dataset. The dataset includes 613 pictures, each with 5 sentences as captions.

RSICD [1] is nowadays the most complicated remote sensing image caption dataset because of the largest scale and high difficulty. RSICD includes 10,921 pictures, but there are only 24,333 ground truth descriptions. For incompletely described images, copies of existing sentences are used to meet the model input requirements of 5 sentences per picture.

Since the manual estimation of quality and accuracy of the generated sentences is susceptible to subjective influences, a series of quantitative evaluation criteria from natural language processing are introduced to our quantitative evaluation. They are BLEU-n (n=1,2,3,4), CIDEr, and ROUGE-L, which abbreviated as B-n, C and R in Table 1.

3.2. Implementation Details and Results

During the training, the self-critical training mechanism [10] and beam search are both utilized with the beam size of 2. For the salient object region proposal, a Faster RCNN with Resnet-101 backbone is pre-trained on MSCOCO detection datasets while class-agnostic is enabled. For global feature extraction, another Resnet-101 model pre-trained on

Table 1. Evaluation results on UCM, Sydney and RSICD dataset. Bold represents the top one.

Dataset	UCM						Sydney						RSICD					
Criteria	B-1	B-2	B-3	B-4	C	R	B-1	B-2	B-3	B-4	C	R	B-1	B-2	B-3	B-4	C	R
Zhang *et al.* [2]	0.594	0.532	0.481	0.429	/	/	0.615	0.540	0.473	0.400	/	/	/	/	/	/	/	/
Attention [1]	0.745	0.655	0.586	0.525	2.612	0.724	0.732	0.667	0.622	0.582	**2.499**	0.713	0.676	0.531	0.433	0.36	1.964	0.611
FC-ATT [3]	0.814	0.750	0.685	0.635	2.996	0.750	0.808	0.716	0.628	0.554	2.203	0.711	0.746	0.625	0.534	0.457	**2.366**	0.633
SM-ATT [3]	0.815	0.758	0.694	0.646	3.186	**0.763**	0.814	0.735	**0.659**	0.580	2.302	0.719	0.757	0.634	**0.538**	0.461	2.356	0.646
Proposed	**0.823**	**0.768**	**0.710**	**0.659**	**3.192**	0.756	**0.817**	**0.742**	0.657	**0.591**	2.291	**0.721**	**0.770**	**0.649**	0.532	**0.471**	2.363	**0.651**

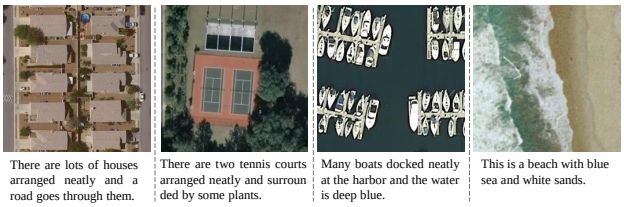

Fig. 2. Example results of our proposed method.

ImageNet classification dataset is chosen to conduct comprehensive semantic information. The caption decoder is trained under the cross-entropy objective using ADAM optimizer with the learning rate of a fixed 10^{-4}. The patch scale factor k is fixed to 2.0 during the test, and the number of detection objects n is set to 5.

Table 1 shows the quantitative results of various approaches on 3 datasets. It should be noted that the listed approaches utilized different CNN encoder architecture to extract features. To be specific, "Attention" [1], "FC-ATT" [3] and "SM-ATT" [3] utilized VGG16 as backbone model, Zhang *et al.* [2] utilized ResNet-152, and all of them were pre-trained on ImageNet. However, the utilized object detector in our implementation was pre-trained on MSCOCO, which is not a remote sensing related dataset, the gap of different datasets definitely draws our performance back.

It can be observed from Table 1, the comparison results indicate that our proposed method surpasses the mainstream methods at most metrics in all datasets. Practically, on the BLEU-n series metrics, our method has shown its superiority most. while the proposed method is much more likely to lag behind on CIDEr and ROGUE-L by other attention-based methods.

4. CONCLUSION

In this paper, we propose a remote sensing image caption method with instances-awareness and cross-hierarchy attention. In this work, the Faster RCNN is utilized to accurately locate the key objects and their surroundings. In this way, the encoder is capable of extracting key element regions accurately, instead of estimating the content based on the uniform spatial grid. To deal with single-textured scenes like deserts and oceans, a global visual feature is also introduced. To adapt to this multi-instance-level feature form, a cross-hierarchy attention mechanism is proposed to prompt the decoder dynamically focus on different semantic levels and different instances of visual feature. The experimental results on mainstream datasets show the effectiveness of the proposed approach.

5. REFERENCES

[1] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, 2017.

[2] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 10039–10042.

[3] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, pp. 612, 2019.

[4] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proc. Conf. Comput. Natural Lang. Learn.*, 2011, pp. 220–228.

[5] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[8] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput. Inf. and Telecom. Syst.*, 2016, pp. 1–5.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7008–7024.