Scene Attention Mechanism for Remote Sensing Image Caption Generation

Shiqi Wu
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
Xidian University
Xi'an 710071, China
sqwu@stu.xidian.edu.cn

Xiangrong Zhang
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
Xidian University
Xi'an 710071, China
xrzhang@mail.xidian.edu.cn

Xin Wang
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
Xidian University
Xi'an 710071, China
15091625752@163.com

Chen Li
School of Computer Science and Technology
Xi'an Jiaotong University
Xi'an 710049, China
cli@xjtu.edu.cn

Licheng Jiao
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
Xidian University
Xi'an 710071, China
lchjiao@mail.xidian.edu.cn

***Abstract*—Remote sensing images play an important role in various applications. To make it easier for humans to understand remote sensing images, the task of remote sensing image captioning attracts more and more researchers' attention. Inspired from the way human receives visual information, attention mechanism has been widely used in remote sensing image understanding. To catch more scene information and improve the stability of the generated sentences, a new attention mechanism called scene attention is proposed. Except for the current attention via the current hidden state of the long short-term memory network (LSTM), our proposed method simultaneously explores the global visual information from the mean feature of all convolutional features. The effectiveness of the proposed method is evaluated on UCM-captions, Sydney-captions and RSICD datasets. The results of our experiment show that comparing with some other captioning methods, our method is more stable and obtains a better performance.**

Keywords—remote sensing image captioning, convolutional neural network, long short-term memory network, scene attention mechanism

I. INTRODUCTION

Remote sensing, an important technology nowadays, can make humans know more about the earth that we live on [1]. Whether in geographic research, urban planning, environmental monitoring, or military intelligence gathering, remote sensing images play an indispensable role. Therefore, it is valuable for humans and machines to better understand the contents of remote sensing images.

The automatic image captioning task aims to describe the main content of the image with short sentences. For other computer vision tasks such as object detection, the most important part is identifying objects in the image. Unlike those tasks, image captioning not only identifies the elements in the image but also understands the relationship between them. There may be a lot of information in an image, but in general, the captioning task only pays attention to the overall and most prominent parts. This may lose some details, but keep the description concise and easy to understand, making it suitable for some specific applications. Automatic captioning of remote sensing images can be useful for both military and civilian use. An example in military use is that a computer can use the image captioning system to automatically convert captured battlefield images into text or voice information and send it to a command center or front line, thereby achieving more rapid information transfer [2]. In terms of civilian use, the captioning of remote sensing image can assist in remote sensing image retrieval and image understanding works such as easily and accurately searching in the remote sensing image library only based on the text.

The main differences between remote sensing images and natural images are as follows: (1) Different perspective. Remote sensing images are taken from overhead, while natural images are generally taken from a human perspective. Therefore, the remote sensing image is more about the top view of objects or scenes, which also means the description of the relative relationship of the objects can be different. (2) Different scale. Natural images are generally taken closer to the object, while remote sensing images are taken from far away. This means the area in the remote sensing image is much larger than the natural image, and the object in the image is smaller, even invisible. At

This work was supported by the National Natural Science Foundation of China (Nos. 61772400, 61772399, 61772409, 61871306), and the 111 Project (No. B07048).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

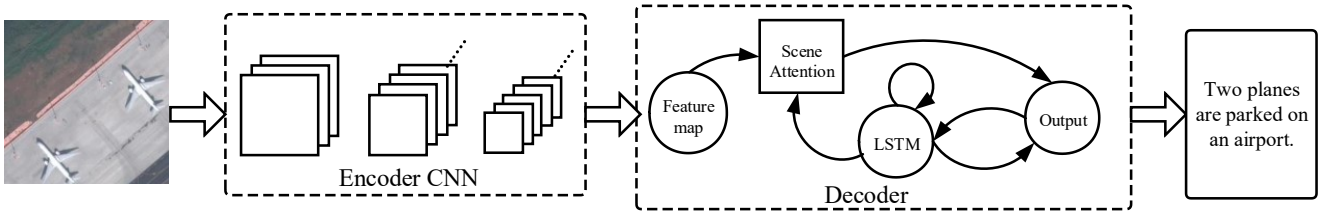


Fig. 1. The outline of the proposed method for remote sensing image captioning.

the same time, the distribution of objects in remote sensing images may be more complicated, while that in natural images is generally concentrated in the central location.

In this paper, an end-to-end captioning method with scene attention mechanism for remote sensing image is proposed. As shown in Fig. 1, first, a convolutional neural network (CNN) extracts features from the input image, and then our proposed scene attention generates the context vector with the information from a long short-term memory network (LSTM) and the extracted features. Finally, the output sentence is generated by combining the hidden state of LSTM and the context vector. In our proposed method, CNN compresses the image and extracts top-layer features by operations such as convolution and pooling; LSTM uses the previous word to generate the next state in the sequence, ensuring that the method can generate a smooth sentence; the proposed scene attention module fuses the current hidden state of LSTM and the mean value of top-layer features to utilize the scene information and control which part of the input image should be focused on so that our method can generate more accurate descriptions.

The main contributions of this paper are as follows:

- Introduce an encoder-decoder based method which uses the information from current hidden state of LSTM for remote sensing image captioning task.
- A novel attention mechanism called scene attention is proposed. It uses both the semantic information from LSTM and the global visual information from features to generate the attention map combining with scene information of the image.
- Performance of our method is shown by experiments with several datasets. We also provide some visualized results of attention maps to show how our scene attention mechanism works on remote sensing images.

II. Related Work

A. Natural Image Captioning

With the rapid development of neural network technology, many image captioning methods based on neural networks have been proposed. For natural image captioning, there are currently three methods: retrieval-based methods, object detection-based methods, and encoder-decoder methods [3].

The retrieval-based method is an earlier method [4]. This method requires a huge image captioning data set. For an input image, the method searches for similar images in the dataset and generate the output sentence based on captions of these images. This method has high requirements for the quantity of the data, and the generated sentences are relatively limited.

The method based on detection needs to identify the objects, actions and scenes in the image first, and then analyze their relationships to generate captions [5][6]. Sentences generated by this method are generally relatively simple and rarely have a modified description. The factors affecting the performance of this method are mainly the performance of the relational model and the accuracy of detection.

The encoder-decoder method is the most common image captioning method and is first used in image captioning by Mao et al. [7]. It is inspired by machine translation, replacing the original text with the image. It uses a CNN as an encoder to extract features to contain the information of the input image and generate feature vectors, and uses a recurrent neural network (RNN) as a decoder to generate captions based on the encoded feature vectors. Vinyals et al. [8] use an LSTM as the decoder instead to improve the performance of the model. Comparing with retrieval-based method and object detection-based method, the structure of this method has more flexibility, and can use the information in the image to generate more diverse sentences, and has better performance. Therefore it is often used as a basic model for image captioning tasks.

Xu et al. [9] have introduced an attention mechanism based on the encoder-decoder framework. The attention mechanism makes the extraction of image information more believable, also it can transfer more image information. The attention mechanism provides an image processing method closer to humans, and each word is corresponding to an attention map so the generated sentences can also correspond exactly to the elements in the image. Since then, many attention mechanism based methods have been proposed. To make the model depend less on the attention mechanism when it comes to non-visual words, Lu et al. [10] have proposed an adaptive attention mechanism, using an visual sentinel so that the model can determine when to focus on the image. On the other hand, the above works are based on spatial attention. For utilizing the channel information of an image, Chen et al. [11] have proposed SCA-CNN model with both spatial attention mechanism and channel-wise attention mechanism. In addition, above attention mechanisms that use features extracted from CNN are top-down attention. With unexpected, novel or salient stimuli, bottom-up attention also has important effects on human visual perception. Therefore, Anderson et al. [12] use Faster R-CNN to get bottom-up attention features, and use an LSTM as top-down attention extractor.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

B. Remote Sensing Image Captioning

Many existing remote sensing image captioning methods follow the basic framework of natural image captioning, and have reached some achievements combined with the characteristics of remote sensing images. Shi and Zou [2] have proposed an object detection-based model for remote sensing image captioning without using LSTM. The model is divided into two parts for image analysis and description generation, respectively. The image analysis part uses a fully convolutional network (FCN). Comparing with traditional CNNs, FCN allows any size of the input image, and the output label image allows it to retain the spatial information of the original image. The language generation template is used in the description generation part. In addition to object detection-based methods, retrieval-based methods has also been used in remote sensing image captioning. Wang et al. [13] have proposed a retrieval-based captioning method using semantic embedding to measure the distance of image representations and collective sentence representations in order to generate appropriate sentences which are close to the input image.

Similar to natural images, the encoder-decoder model has also been used widely in the task of remote sensing image captioning. Qu et al. [14] combine the image features of high-resolution remote sensing images and their corresponding text information, and use the encoder-decoder model composed of a CNN and an RNN or a CNN and an LSTM for caption generation. Then Lu et al. [3] add the attention mechanism to the encoder-decoder model for better use of visual information of remote sensing images. As an improvement of basic attention mechanism, Zhang et al. [15] have proposed an attribute attention mechanism, which combines both the low-level and high-level features to generate more satisfying sentences by using attributes of remote sensing images.

A problem of captioning tasks for remote sensing images is that, since this topic is rarely studied, there are few suitable datasets for training and testing. Therefore, Qu et al. [14] proposed the UCM-Captions dataset and the Sydney-Captions dataset, which are based on the UC Merced Land-Use dataset [16] and the Sydney dataset, respectively [17]. However, data samples and scene categories in these two datasets are still not enough. So in [3] the authors proposed RSICD dataset, which greatly expands the size of data, and the sample images in the dataset have higher intra-class diversity and lower inter-class differences, providing researchers with valuable data resource for research on remote sensing image captioning tasks.

III. Methodology

The proposed method is based on the encoder-decoder model which has shown good performance in previous researches. In this section, we first introduce the overall structure of our proposed method. Then we focus on our proposed scene attention mechanism.

A. Overall Structure

In the CNN-LSTM encoder-decoder model, given the input image and its corresponding description, the model needs to maximize the probability of the words to be generated as following:

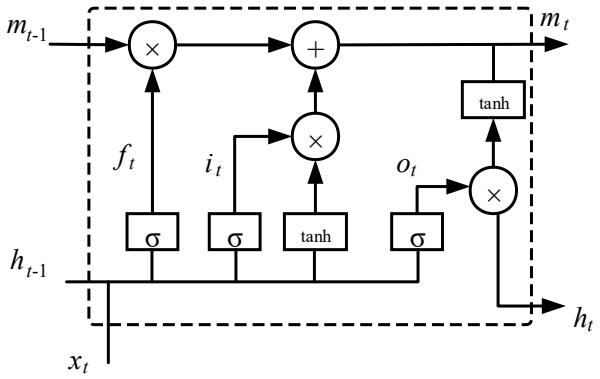


Fig. 2. The basic structure of LSTM.

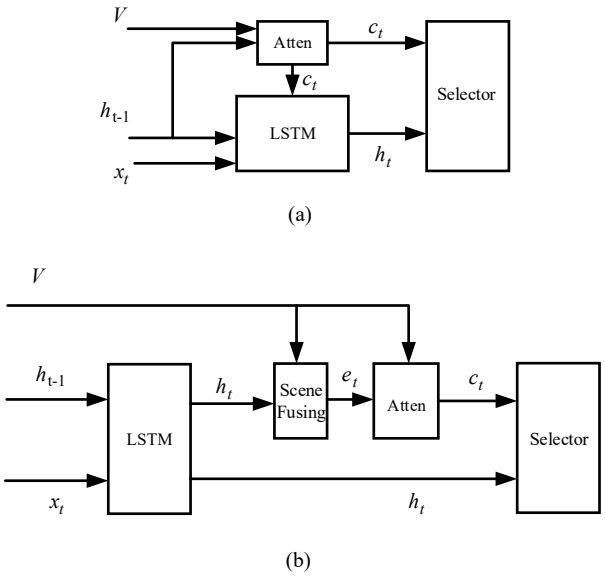


Fig. 3. (a) Structure of traditional soft-attention. (b) Our proposed scene attention method.

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p\left(S \middle| I; \theta\right) \tag{1}$$

where I is the input image, and $S = \{S_1, ..., S_T\}$ is the ground truth of the output sentence which have T words. θ is the parameters of the model. $p\left(S \middle| I; \theta\right)$ represents the probability of the output S generated from input I under parameters θ. According to the chain rule, the log-likelihood can be decomposed as:

$$\log p\left(S \middle| I\right) = \sum_{t=1}^{T} \log p\left(S_t \middle| S_1, \ldots, S_{t-1}, I\right) \tag{2}$$

where θ is omitted for convenience.

Through the encoder, the top-layer features V of the input image can be extracted as:

$$V = \mathrm{CNN}\left(I\right) \tag{3}$$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I. RESULTS OF DIFFERENT METHODS ON UCM-CAPTIONS DATASETS

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
CCSMLF[13]	0.387	0.215	0.125	0.091	0.095	0.360	0.370
Multimodal [14]	0.787	0.710	0.649	0.594	0.404	0.754	2.927
Soft-att [3]	**0.826**	0.756	0.699	0.649	0.417	0.761	3.092
Ours	0.822	**0.765**	**0.717**	**0.674**	**0.440**	**0.778**	**3.228**

TABLE II. RESULTS OF DIFFERENT METHODS ON SYDNEY-CAPTIONS DATASETS

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
CCSMLF[13]	0.444	0.337	0.282	0.241	0.158	0.402	0.938
Multimodal [14]	0.715	0.608	0.529	0.462	0.348	0.641	1.956
Soft-att [3]	0.744	0.662	0.599	0.542	0.380	0.676	2.323
Ours	**0.786**	**0.698**	**0.626**	**0.561**	**0.381**	**0.709**	**2.505**

TABLE III. RESULTS OF DIFFERENT METHODS ON RSICD

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
CCSMLF[13]	0.576	0.386	0.283	0.222	0.213	0.446	0.530
Multimodal [14]	0.609	0.439	0.337	0.268	0.244	0.459	0.738
Soft-att [3]	**0.629**	0.460	0.359	0.292	**0.254**	0.470	0.788
Ours	0.625	**0.463**	**0.364**	**0.297**	0.253	**0.474**	**0.809**

then the probability $p\left(S_t \middle| S_1,\ldots,S_{t-1},V\right)$ can be modeled with an LSTM.

The basic structure of an LSTM is shown in Fig. 2. The gates in LSTM and the outputs can be defined as:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{4}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{5}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{6}$$

$$m_t = f_t \odot m_{t-1} + i_t \odot \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{7}$$

$$h_t = o_t \odot \tanh\left(m_t\right) \tag{8}$$

where f_t, i_t, o_t, h_t, m_t are the forget gate, input gate, output gate, hidden state and the memory cell, respectively. σ is the sigmoid function. The various W, U, and b are trainable parameters. x_t is the embedded words of the caption, which is the ground truth in the training set while training or the previous generated word while testing. The word embedding maps words to a vector space so that images and text can be processed in the same space. x_t can be calculated as:

$$x_t = W_s S_t \tag{9}$$

where W_s is the embedding weight to be trained.

Adding the visual context c_t, the output probability can be calculated with a non-linear function g including attention module and attention selector as follows:

$$p\left(S_t \middle| S_1,\ldots,S_{t-1},V\right) = g\left(h_t, c_t\right) \tag{10}$$

A traditional RNN can also be used as a decoder, but we use an LSTM instead. It is because an LSTM is able to memorize information of a few words before, related to the word being generated, which makes it suitable for generating longer sentences. Also, LSTM has been proven with good performance in previous research [18][8].

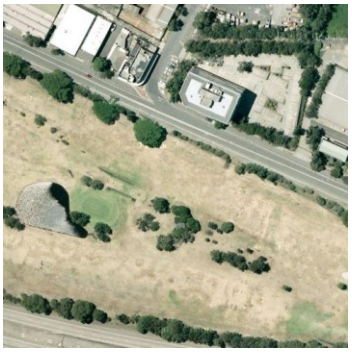
B. Scene Attention Mechanism

The visual context vector c_t is an important coefficient in the encoder-decoder model. It contains the visual information of the image and provides the basis for the description work. In the model using the attention mechanism, c_t has a relationship with both the encoder and the decoder. When predicting each word, the attention module in decoder fuses the feature map and the current hidden state to produce a feature weight α_t to decide which part of the input image to be focused on at time t, and

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Soft-att: Lots of boats docked neatly at the harbor.
Scene att: Lots of boats docked neatly at the harbor.

(a)

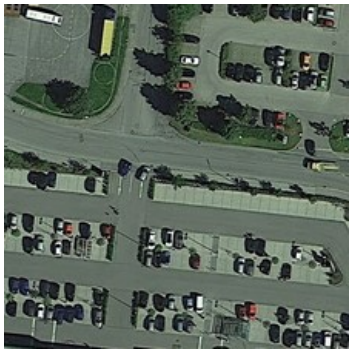


Soft-att: Some plants on the wide river.
Scene att: There are some green bushes and white bunkers on the meadow.

(b)

Soft-att: There is a curved river with some plants on both sides of the river.
Scene att: A meadow with some white marking lines on it while some plants beside.

(c)



Soft-att: Many cars are parked in a parking lot with some green trees.
Scene att: Many cars are parked in a parking lot with several green trees.

(d)

Soft-att: A baseball field is near several green trees.
Scene att: A baseball field is next to a road.

(e)

Soft-att: Some storage tanks are near some green trees and a building.
Scene att: Some storage tanks are near a road.

(f)

Fig. 4. Several input images and the corresponding sentences generated by soft-attention and our proposed method.

uses image features from the CNN convolutional layer to calculate c_t. So it will constantly change during the calculation process.

As shown in Fig. 3(a), the traditional soft-attention uses the image features and the previous hidden state of LSTM as the input of the attention module. It also uses the output context of the attention as a part of LSTM inputs. Different from the traditional attention-based method, we don't use the context generated by attention module to train the LSTM, but use the current hidden state of LSTM and a fusing module to generate the attention map, as shown in Fig. 3(b). This makes the hidden state a residual structure, which can reduce the instability and maintain the scene information to make the attention module supplement the current hidden state for better predicting the output at time t.

In the attention module, first, we use the mean of the top-layer features extracted by the encoder to represent the scene features of the image, and then concatenate it with the hidden state h_t of the LSTM in the decoder. The fused feature e_t can be represented as following:

$$e_t = \mathrm{concat}\left(W_m \overline{V}, W_h h_t\right) \tag{11}$$

where $\overline{V}$ is the mean of the features V. W_m, W_h are weights to be trained in linear layers.

We use a single fully connected layer to combine V and the fused feature e_t. Set k is the number of regions of the feature map V and the dimension of the weight α_t, and the softmax function generates the attention distribution of k regions. At time t, the weights of features α_t can be calculated as:

$$\alpha_t = \mathrm{softmax}\left(W_a \tanh\left(W_v V + \left(W_e e_t\right)\mathbf{1}^{\mathrm{T}}\right)\right) \tag{12}$$

where W_a, W_v, W_e are weights to be trained in linear layers, and $\mathbf{1}^{\mathrm{T}}$ is a vector of k dimensions whose elements are all 1. The output context c_t of attention module is:

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{ti} \tag{13}$$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

where v_{ti} and α_{ti} are i-th region of the feature map and its weight at time t.

After the new visual context vector is calculated according to the feature weights, the model can better know which parts of the image should be focused on and generate corresponding words based on these areas of interest with the scene attention mechanism. We use an attention gate as in [9], in which a scalar β_t to represent the proportion of the attention mechanism at time t. The larger β_t is, the more the generated word dependent on attention module. It is generated by a single fully connected layer as follow:

$$\beta_t = \sigma\left(W_\beta h_t + b_\beta\right) \tag{14}$$

where W_β and b_β are trainable parameters. Then the selected context $\hat{c}_t$ is:

$$\hat{c}_t = \beta_t \cdot c_t \tag{15}$$

Through attention gate, the model can give more weight to image information when predicting words related to the object, thus can predict the next word based on the image content. In addition, it can reduce the weight of image information when predicting words with no visual information, such as prepositions, conjunctions, etc. At this time the model pays more attention to the words that have been generated before, and predict the next word based on the above. Therefore, the attention mechanism can focus more on and emphasize the objects in the image.

IV. Experiments

A. Experiment Details

We use UCM-captions [14], Sydney-captions [14] and RSICD [3] datasets for testing the performance of our proposed method. The UCM-captions dataset contains a total of 2100 images with a size of 256 × 256, including 21 categories, and 100 images per category. The Sydney-captions dataset contains a total of 613 images with a size of 500 × 500, including 7 categories. The RSICD dataset contains a total of 10916 images with a size of 224 × 224, including 30 categories. Each image corresponds to 5 descriptions in the above three datasets. For each dataset, about ten percent of the samples are used as the validation set and about ten percent are the test set, and the rest are used as the training set. It is to be noted that we use the default partition.

Before training, each input image is resized to 224 × 224. For all captions in the training set, count the words that appear in them, and sort the numbers according to the total number of times the words appear in the training set to make a vocabulary. We use a pre-trained on the ImageNet dataset VGG-19 network [19] as the encoder. Feature extraction is performed on the input image at the Conv5-4 layer. Each pooling layer reduces the image to a quarter and finally generates a feature map of 14 × 14 × 512. After passing through 4 pooling layers, the inputted feature vector of the decoder is in size of 196 × 512 finally. When generating words, we use a beam search in which beam width is set as 5. This can help the model to find a better solution and generate more appropriate sentences.

B. Results

We use BLEU, METEOR, ROUGE_L and CIDEr, commonly used evaluation metrics in the image captioning task, to evaluate the captioning results. The evaluation results shown in Table I-III are automatically generated using COCO-captions [20]. It can be seen that comparing with the traditional multimodal method, methods with attention mechanism can bring obvious improvement to the performance. Though for UCM-captions and RSICD datasets there are not much differences on BLEU-1 score, our method has a better performance than soft-attention method on CIDEr, which is a metric that designed for image captioning [21]. CIDEr performs TF-IDF weighting to give higher weights to the n-tuples that often appear in the original description and reduce the weights of those often appear in all descriptions, so it can increase the importance of the unique description of each image in the evaluation. This shows that our method can catch the unique information of images better and is more suitable for practical applications.

Fig. 4 shows some images from UCM-captions dataset, Sydney-captions dataset and RSICD dataset and captions generated by soft-attention and our proposed scene attention method. The results show that our method can generate smooth sentences with the main objects in the scene of the image and the relationship between them also well described. While for Fig.4(b), caption generated by soft-attention model mistakes the meadow for the river. And for Fig.4(c) it mistakes the roads for the river. In remote sensing images, meadows and roads are similar to rivers, so that it is difficult to distinguish for soft-attention method. However, since there are few buildings around rivers in general, from the perspective of the entire scene, our method can determine the grass and road correctly according to the buildings in the image, which shows the superiority of our scene attention mechanism. It can also be noticed that for Fig.4(e) and Fig.4(f), both soft-attention and scene attention method can generate reasonable captions, but with different elements. In our method, the fused features increase the proportion of some specific scene information. Therefore, the description generated by our method is not only limited to the main object and tends to focus on the entire scene content. So that our scene attention can identify the roads, which are closer to the edge in Fig.4(e) and Fig.4(f).

The generated attention maps of our proposed method are shown in Fig. 5. Each image shows the attention map of our proposed method when predicting different word in the sentence. Highlight areas in the images signify where the attention mechanism focuses on. For the reason we use a visual selector instead of the attention itself to inhibit the visual information when predicting these words, some randomness of the attention mechanism emerges, so the maps show that some prepositions like “with” and “in” also have large highlight areas. It also needs to be noted that some phrases like “green trees” describe only one element, but have visual information in both words. In this case, generating several different attention maps may not be a good solution.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

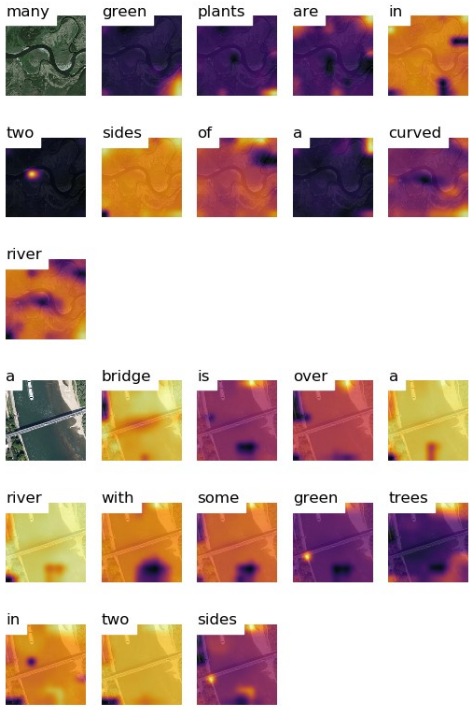


Fig. 5. The visualized attention maps which shows the correspondence between the model's focus area and the generated words.

V. Conclusion

In this paper, a remote sensing captioning method with scene attention mechanism has been proposed and achieved good performance on UCM-captions dataset, Sydney-captions dataset and RSICD dataset. The results have shown that our scene attention method can catch the scene information and effectively improve the performance of the captioning method compared with the traditional soft attention mechanism. Although our method has considered the characteristic of remote sensing images, e.g. exploring scene information to fit the large scale of the image, how to design models being suitable to other characteristics of remote sensing images, such as various views, different sizes of objects and high similarity between scenes, is still a challenge for remote sensing image captioning.

References

[1] Zhang Liangpei, Zhang Lefei, and Du B. "Deep learning for remote sensing data: A technical tutorial on the state of the art." IEEE Geoscience and Remote Sensing Magazine 4.2 (2016): 22-40.

[2] Shi Z, and Zou Z. "Can a machine generate humanlike language descriptions for a remote sensing image?." IEEE Transactions on Geoscience and Remote Sensing 55.6 (2017): 3623-3634.

[3] Lu X, Wang B, Zheng X, and Li X. "Exploring models and data for remote sensing image caption generation." IEEE Transactions on Geoscience and Remote Sensing 56.4 (2017): 2183-2195.

[4] Ordonez V, Kulkarni G, and Berg T L. "Im2text: Describing images using 1 million captioned photographs." Advances in neural information processing systems. 2011.

[5] Li S, Kulkarni G, Berg T L, Berg A C, and Choi Y. "Composing simple image descriptions using web-scale n-grams." Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011.

[6] Yang Y, Teo C L, Daumé III H, and Aloimonos Y. "Corpus-guided sentence generation of natural images." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[7] Mao J, Xu W, Yang Y, Wang J, and Yuille A L. "Explain images with multimodal recurrent neural networks." arXiv preprint arXiv:1410.1090 (2014).

[8] Vinyals O, Toshev A, Bengio S, and Erhan D. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[9] Xu K, Ba J, Kiros R, Cho K, Courville A, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

[10] Lu J, Xiong C, Parikh D, and Socher R. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[11] Chen L, Zhang H, Xiao J, Nie L, Shao J, et al. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] Anderson P, He X, Buehler C, Teney D, Johnson M, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[13] Wang B, Lu X, Zheng X, and Li X. "Semantic descriptions of high-resolution remote sensing images." IEEE Geoscience and Remote Sensing Letters (2019).

[14] Qu Bo, Li X, Tao D, and Lu X. "Deep semantic understanding of high resolution remote sensing image." 2016 International Conference on Computer, Information and Telecommunication Systems (CITS). IEEE, 2016.

[15] Zhang X, Wang X, Tang X, Zhou H, and Li C. "Description generation for remote sensing images using attribute attention mechanism." Remote Sensing 11.6 (2019): 612.

[16] Yang Y, and Shawn N. "Bag-of-visual-words and spatial extensions for land-use classification." Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. ACM, 2010.

[17] Zhang F, Du B, and Zhang L. "Saliency-guided unsupervised feature learning for scene classification." IEEE Transactions on Geoscience and Remote Sensing 53.4 (2014): 2175-2184.

[18] Gers F. Long short-term memory in recurrent neural networks. Diss. Verlag nicht ermittelbar, 2001.

[19] Simonyan K, and Andrew Z. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[20] Chen X, Fang H, Lin T Y, Vedantam R, Gupta S, et al. "Microsoft coco captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325 (2015).

[21] Vedantam R, Lawrence Zitnick C, and Parikh D. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE