Journal of Systems Engineering and Electronics
Vol. 34, No. 1, February 2023, pp.9 – 18

VLCA: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning

WEI Tingting, YUAN Weilin, LUO Junren, ZHANG Wanpeng*, and LU Lina

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: In the field of satellite imagery, remote sensing image captioning (RSIC) is a hot topic with the challenge of overfitting and difficulty of image and text alignment. To address these issues, this paper proposes a vision-language aligning paradigm for RSIC to jointly represent vision and language. First, a new RSIC dataset DIOR-Captions is built for augmenting object detection in optical remote (DIOR) sensing images dataset with manually annotated Chinese and English contents. Second, a Vision-Language aligning model with Cross-modal Attention (VLCA) is presented to generate accurate and abundant bilingual descriptions for remote sensing images. Third, a cross-modal learning network is introduced to address the problem of visual-lingual alignment. Notably, VLCA is also applied to end-to-end Chinese captions generation by using the pre-training language model of Chinese. The experiments are carried out with various baselines to validate VLCA on the proposed dataset. The results demonstrate that the proposed algorithm is more descriptive and informative than existing algorithms in producing captions.

Keywords: remote sensing image captioning (RSIC), vision-language representation, remote sensing image caption dataset, attention mechanism.

DOI: 10.23919/JSEE.2023.000035

1. Introduction

Image captioning is concerned with the process of generating linguistic caption information for given images, which has been a hot topic in cross-modal learning. Aerial and satellite photography from remote sensing systems is crucial for interpreting geospatial data in various applications [1,2] and remote sensing image semantic information can be directly applied in defence fields such as disaster report generation [3], remote sensing image retrieval [4], and military intelligence generation [5]. To facilitate cross-modal information interpretation, remote sensing image captioning (RSIC) and the vision-language model merge data from remote sensing imagery and linguistic captions. Recent research has always struggled with how to effectively utilize the features of images and texts, and combine visual and linguistic representations.

Many captioning techniques for remote sensing scenes have emerged since RSIC proposed by Qu et al. [3]. Most of the techniques make use of natural image captioning (NIC) processing techniques. Researchers [6,7] used traditional convolutional neural networks (CNN) to extract image features and long-short term memory (LSTM) to generate words based on the standard encoder-decoder framework. The relationship between image and text is also learned simultaneously through attention mechanism [8,9].

Applying NIC methods directly to RSIC often requires some adjustments to the model, because remote sensing images differ from natural images in terms of the shooting angle and target scale size. In RSIC, fully processing visual information and integrating it with language generation is a difficult issue. Most of the existing RSIC methods directly use the output features and text features of CNN convolutional layer for learning but fail to align the high-level semantic information of the image with the text, or do not fully exploit high-level image visual information. Contrastive language-image pre-training (CLIP) [10], on the other hand, performs contrastive learning between text features and image features to train a transferable visual model. At present, CLIP has been applied to many vision-language multimodal tasks with quite outstanding performance [11−13], providing a method for solving RSIC vision-language alignment. The generated description statements are typically monotonous and uncorrelated to the pictorial content. The vision-language pretraining model has significant advantages in image captioning [14,15] and is used to solve the above problems in this paper.

In this paper, we propose a Vision-Language aligning method with Cross-modal Attention (VLCA) for bilingual remote sensing image captioning, which takes into account high-level visual semantics information. Our contributions are as follows:

Manuscript received August 30, 2022.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61702528;61806212).

(i) A new RSIC dataset is built with hand-annotated captions for object detection in optical remote sensing images dataset (DIOR) [16], named DIOR-Captions, which substantially enhances RSIC datasets.

(ii) To improve the variety and precision of description production, a vision-language pre-training architecture for bilingual RSIC is proposed.

(iii) To further jointly represent the fusion of visual and textual features and promote linguistic model efficiency, a cross-modal learning network is suggested for the alignment of vision and language, which maps both visual and semantic vectors to the same spatial location.

2. Related work

In this section, we revisit the development of RSIC and vision-language pretraining progress. A brief analysis of cross-lingual image captioning is also given for guiding the research of bilingual RSIC.

2.1 RSIC

Most current RSIC researches widely adopt the encoder-decoder framework and attention mechanism. Qu et al. [3] proposed a deep multimodal neural network model for semantic understanding of the high-resolution remote sensing images and investigated the effect of various types of CNN combined with recurrent neural network (RNN) on RSIC. In recent work, the researchers mostly aim to design attention mechanisms [6,17,18] for captions focusing on salient areas of the image.

Many cutting-edge vision-language pre-training methods, including vision-language pre-training (VLP) [15], Oscar [19], VilBert [20], and ClipCap [13] achieve extraordinary performance in challenging vision-language downstream tasks including NIC. The success can be attributed to the rapid development of pre-training techniques. However, few RSIC researches investigate the effects of vision-language pretraining models on RSIC, and the development of pretraining technology can benefit RSIC a lot, which is one of the motivations for this paper.

2.2 Bilingual captions generation

Only a few studies focus on the cross-language description [21,22], and the relevant datasets and studies in the field of remote sensing have not been retrieved. However, cross-language description is crucial for cross-language information retrieval and other applications [23] and non-native English speakers, which can bridge the semantic gap to a certain extent. Previous studies have proved that the end-to-end training method is better than the translation result, where the language of training and target captions are the same.

3. Methodology

VLCA based on encoder-decoder architecture is proposed by the following efforts Firstly, the high-level semantic information about the image is obtained by using CLIP and then the cross-modal attention module aligns the image features and text features for language generation. The language model generative pre-training (GPT-2) [24] generates correlated captions for given images. The overall framework of VLCA is shown in Fig. 1.

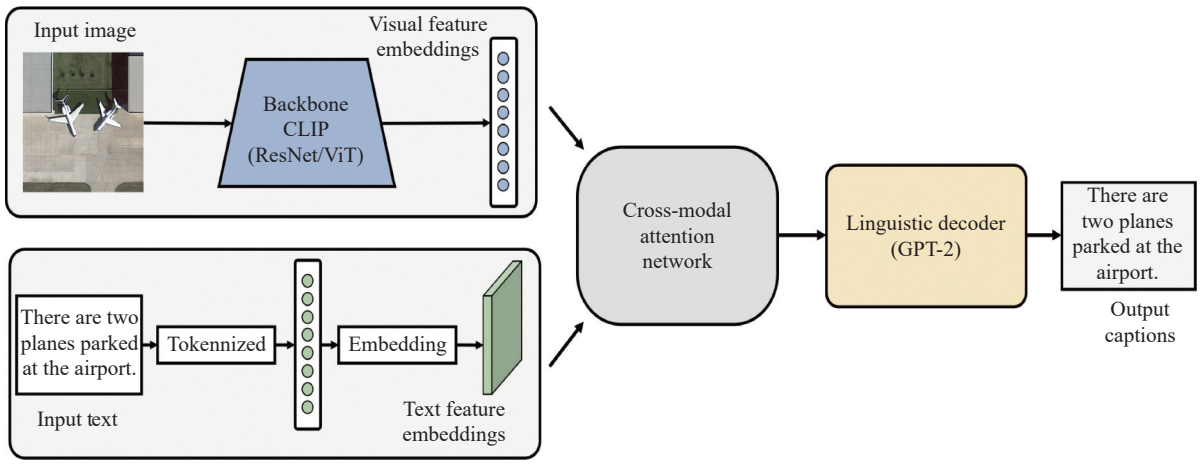


Fig. 1 Framework of the VLCA for bilingual RSIC

3.1 Image feature extraction

For downstream tasks (e.g., image captioning), image feature extraction is a critical step, in which many current methods show remarkable performance, such as CLIP. Based on a sizeable dataset of image-text pairs, CLIP makes great success in vision-language pre-training study, with advanced techniques including training contrast loss and building inextricably links between visual and textual representations. Vision Transformer (ViT) [25] and residual network (ResNet) [26] serve as

the backbone network of visual features extraction to transform image features to image embeddings. And image features are treated as input to the language model to guide the generation of sentences. Meanwhile, the input text is encoded as text embeddings, and the progress is shown in Fig. 1.

3.2 Language model fine-tuning

Translating image representations provided by CLIP to language is the main challenge in this paper. To address this issue, we fine-tune GPT-2 [24] for generating accurate and diverse descriptions, which has made great success in the text generation task in recent years. GPT-2 was constructed by Transformer [27] and trained on enormous datasets, which generates captions by combining the word embeddings with image features extracted by CLIP. When the language model is fixed, the embedded features can be a set of embeddings without textual meaning. For this case, additional language pre-training model fine-tuning is necessary for addressing the remote sensing scene language generation problem. The fine-tune pre-trained language model effectively learns the relationship between image data and corresponding contents, which can increase the number of training parameters but produce better descriptions.

3.3 Cross-modal learning network

A cross-modal learning network (CMN) is introduced for learning the relationship between text features and image features and mapping these features to GPT-2 space, as shown in Fig. 2. The external attention multi-layer perception (EAMLP) [28] is considered as one of the primary constituents, which is a lightweight network and can attend potential correlation between different samples compared with self-attention [27].

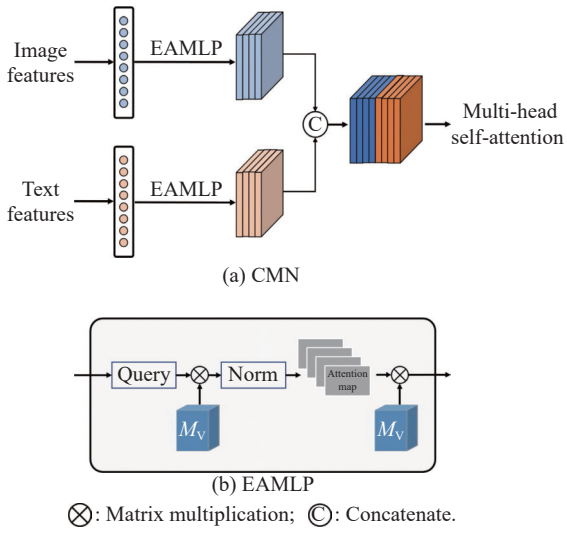


Fig. 2 Structure of the CMN

External attention computes attention between the input features and the external memory unit $\boldsymbol{M}$. Given an image x^i, the input feature is

$$\boldsymbol{F}_i = \{\boldsymbol{q}_1^i, \boldsymbol{q}_2^i, \cdots, \boldsymbol{q}_k^i\} = \text{CLIP}(x^i). \tag{1}$$

The attention matrix $\boldsymbol{A}$ is expressed as

$$\boldsymbol{A} = (\alpha)_{i,j} = \text{Norm}\left(\boldsymbol{F}\boldsymbol{M}^{\text{T}}\right) \tag{2}$$

where $(\alpha)_{i,j}$ is the pair-wise similarity of the ith and jth elements. Norm$(\cdot)$ means normalizing the external attention output so that $\sum_j (\alpha)_{i,j} = 1$. Unlike softmax adopted in self-attention, double-normalization [29] is employed in external attention, which avoids the characteristic of sensitivity to input scale. This double-normalization is formulated as

$$\tilde{\alpha}_{i,j} = \boldsymbol{F}\boldsymbol{M}_k^{\text{T}}, \tag{3}$$

$$\hat{\alpha}_{i,j} = \frac{\exp(\tilde{\alpha}_{i,j})}{\sum_k \exp(\tilde{\alpha}_{k,j})}, \tag{4}$$

$$\alpha_{i,j} = \frac{\hat{\alpha}_{i,j}}{\sum_k \hat{\alpha}_{i,k}}. \tag{5}$$

Multi-head external attention network uses two external memory units $\boldsymbol{M}_k$ and $\boldsymbol{M}_v$ to calculate the attention matrix $\boldsymbol{A}$ as follows:

$$h_i = \text{ExternalAttention}(\boldsymbol{F}_i, \boldsymbol{M}_k, \boldsymbol{M}_v), \tag{6}$$

$$\begin{aligned} \boldsymbol{F}_{\text{out}} &= \boldsymbol{p}_1^i, \boldsymbol{p}_2^i, \cdots, \boldsymbol{p}_k^i = \\ &\text{MultiHead}(\boldsymbol{F}, \boldsymbol{M}_k, \boldsymbol{M}_v) = \\ &\text{Concat}(h_1, h_2, \cdots, h_H)\boldsymbol{W}_o, \end{aligned} \tag{7}$$

where h_i is the ith head, H is the number of heads, and $\boldsymbol{W}_o$ is a linear transformation matrix.

To fuse image and text t^i features, the sentences are split into words called tokens, and then convert tokens into vectors $\boldsymbol{c}_1^i, \boldsymbol{c}_2^i, \cdots, \boldsymbol{c}_k^i$. Then the visual embeddings and caption embeddings are concatenated:

$$\begin{aligned} \boldsymbol{Z}^i &= \boldsymbol{P}^i \oplus \boldsymbol{C}^i = \\ &(\boldsymbol{p}_1^i, \boldsymbol{p}_2^i, \cdots, \boldsymbol{p}_k^i, \boldsymbol{c}_1^i, \boldsymbol{c}_2^i, \cdots, \boldsymbol{c}_\ell^i). \end{aligned} \tag{8}$$

In the training phase, $\{\boldsymbol{Z}^i\}_{i=1}^N$ is chosen as the input of VLCA; meanwhile, outputs of VLCA are predicting captions. And we use cross-entropy loss to train the model, which calculates as follows:

$$\mathcal{L}_X = -\sum_{i=1}^{N}\sum_{j=1}^{\ell}\log_2 p_\theta\left(\boldsymbol{c}_j^i \mid \boldsymbol{p}_1^i, \boldsymbol{p}_2^i, \cdots, \boldsymbol{p}_k^i, \boldsymbol{c}_1^i, \boldsymbol{c}_2^i, \cdots, \boldsymbol{c}_{j-1}^i\right). \tag{9}$$

3.4 DIOR-Captions dataset

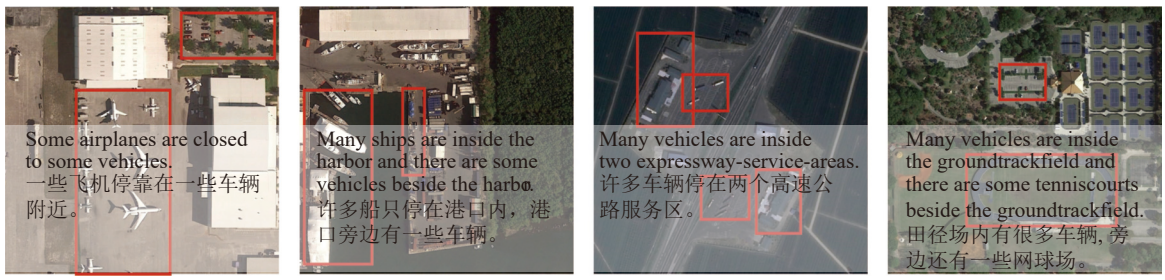
There are few publicly available RSIC datasets, with none annotated in Chinese. To the best of our knowledge, the annotated dataset DIOR-Captions is the first RSIC dataset with both Chinese and English captions.

The RSIC dataset (RSICD) named DIOR-Captions is constructed based on DIOR, which is a large dataset for object detection in optical remote sensing imagery with abundant image and scenarios. DIOR contains abundant scenarios and we randomly select some images for detailed annotation. The comparison of DIOR-Captions with existing public datasets is shown in Table 1 (The number in bold indicates the highest value in each column, and the table below is identical to this). As can be seen from Table 1, DIOR-Captions extensively enriches RSIC datasets and depicts more common remote sensing scenarios.

Table 1 Comparison of different datasets

Dataset	Category number	Average caption length	Vocabulary size	Number of images
Sydney [3]	21	11.5	315	2,100
University of California Merced [3]	7	**13.2**	231	613
RSICD [6]	**30**	11.4	**2 695**	10 000
DIOR-Captions	20	11.2	376	**16 565**

In DIOR-Captions, remote sensing image is annotated with a sentence both in English and Chinese, based on the original object detection bounding boxes in DIOR. Some examples of DIOR-Captions are shown in Fig. 3. For each image in DIOR-Captions, we refer to the object bounding box in DIOR during labeling. The most prominent (or the most numerous) objects are described in sentences, where the linguistic meanings of English and Chinese texts are consistent. The DIOR-Captions dataset is used for bilingual scenarios in this paper, and also can be applied to cross-lingual RSIC research.



(a) Image 1 (b) Image 2 (c) Image 3 (d) Image 4

Fig. 3 Examples of DIOR-Captions

In addition, DIOR-Captions can provide support for other RSIC methods (e.g., the scene graph generation method in NIC) due to the specific object label in DIOR-Captions. Therefore, the DIOR-Captions dataset is of great practical significance and can provide important support for future research and we will publish it soon.

4. Experiments

4.1 Training details

All the experiments are running on a single server with 80 CPUs and four NVIDIA TITAN RTX GPUs. The proposed DIOR-Captions dataset is used to verify VLCA. In which the ratio of the number of training images to the number of test images is 8∶2. In the training phase, the CLIP pretrained models (ResNet50 and ViT) are used to extract image features, in which image feature representation size (or clip size) is set to 512. For text generation, the GPT-2 pretrained model provided by OpenAI is used and the model dimensionality is 768. The nucleus sampling method Top-P [30] is adopted to generated accurate and diverse captions. Additionally, we use Adam [31] for optimization. Parameters and hyperparameters are summarized in Table 2.

Table 2 Details of the training process

Parameter	Value
Initial learning rate	$2e^{-5}$
Batch size	40
Epoch	50
Warmup step	4 000
Number of external attention heads	8
Word embedding dimension	512
Model dimensionality of GPT-2	768
CLIP embedding dimension	512
Probability value of TopP	0.9
Max captions length	20

4.2 Baselines

VLCA is compared with various baselines for captioning images. We implement the following method based on open-sourced code [32].

(i) VGG16-based multimodal (VMM) [3]: Multimodal method explores the impact of different types of CNN combined with LSTM. The combination of VGG16 and LSTM is selected to compare with VLCA.

(ii) Soft attention (SOA) [6]: SOA adopts VGG16 to encode image features and soft attention-based LSTM to generate captions.

(iii) Attribute attention [17]: Attribute attention focuses on the text generation process, which leverages the output of CNN fully connected layers (FC) and the output of convolution layer through softmax (SMA) as the result of feature extraction.

(iv) Scene attention (SCA) [18]: SCA encodes the image with VGG-16 and generates captions by LSTM. The attention weights are computed by fusing convolutional layer outputs and hidden states of LSTM.

4.3 Evaluation metrics

To validate VLCA, the following image captioning indicators are used to assess algorithm performance.

(i) Bilingual evaluation understudy (BLEU) [33] computes the co-occurrences of consecutive words (*N*-gram) to determine the similarity of two sentences and BLEU4 computes a cumulative score from 1-gram to 4-gram with equal weights.

(ii) Recall understudy for gisting evaluation (ROUGE) [34] and BLEU differ in that ROUGE only takes recall into account rather than fluency and accuracy.

(iii) Metric for evaluating translation with explicit ordering (METEOR) [35] is based on recall and accuracy and used to address the issue that BLEU does not take into account synonyms or similar expressions.

(iv) Consensus-based image description evaluation (CIDEr) [36] measures image captioning consistency by applying term frequency-inverse document frequency (TF-IDF) weights to each *N*-tuple to compute the cosine similarity between the reference and the generated caption.

In the abovementioned metrics, BLEU, ROUGE-L, and METEOR are introduced from machine translation task and the ranges are limited in [0,1]. CIDEr is designed for image captioning and the range is set to [0,5]. The higher the score of evaluation metrics, the more effective the captioning method is.

4.4 Results

Three different experiments are designed in this subsection to validate the effectiveness of the algorithm framework (Subsection 4.4.1), the performance of the cross-modal attention module and the fine-tuning pretrained language model (Subsection 4.4.2), and the extensive application of bilingual captioning (Subsection 4.4.3).

4.4.1 Performance against baselines

VLCA is compared with VMM, SOA, attribute attention with FC (FCA), SMA, and SCA. The results are shown in Table 3 and Fig. 4. The comparison between the sentences generated by VLCA and other models is shown in Fig. 5, where words in blue are accurate descriptions, words in red are inaccurate descriptions, and words in bold are the results of the proposed method.

VLCA is extended to bilingual RSIC to demonstrate the actual application value of VLCA and DIOR-Captions in this part.

Table 3 and Fig. 4 show that the multimodal and attention methods perform poorly. The indicators of attribute attention and scene attention are higher. By transferring the visual-language pretraining model, VLCA model outperforms all baselines in both METEOR and CIDEr indicators, which verifies the effectiveness of the proposed model. Notably, VLCA has an average performance in BLEU4 and ROUGE-L and is higher than the best scene attention method in METEOR and CIDEr.

Table 3 Metric results of different method

Method	BLEU4	ROUGE-L	METEOR	CIDEr
VMM [3]	0.293	0.660	0.286	1.292
SOA [6]	0.294	0.631	0.279	1.469
FCA [17]	0.355	0.655	0.308	2.149
SMA [17]	0.342	0.659	0.307	2.033
SCA [18]	0.369	0.685	0.317	2.200
VLCA-ViT	0.309	0.631	**0.372**	2.293
VLCA-ResNet	**0.401**	**0.696**	**0.330**	**2.591**

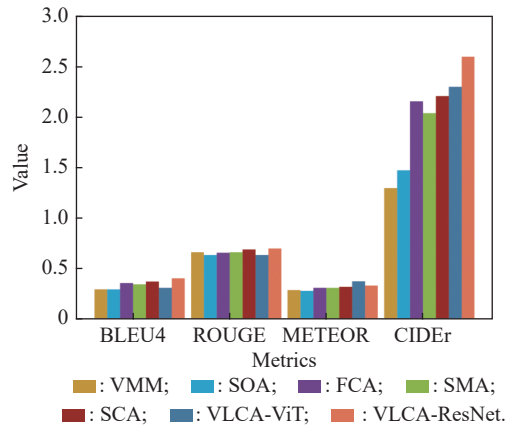


Fig. 4 Comparison of different methods

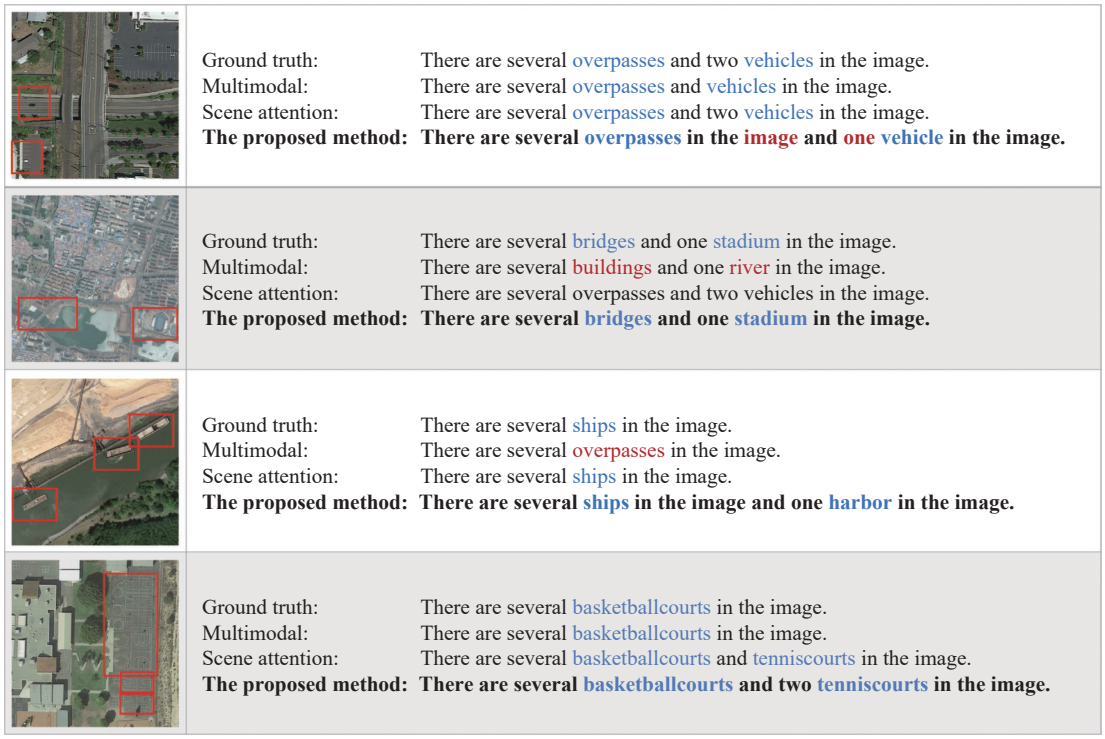


Fig. 5 English captions results of VLCA compared with VMM and SCA

The initial goal of this research is to develop a new approach for RSIC by using the existing vision-language pretraining model. Although VLCA performs better than the baselines, the pre-training method for RSIC still needs to be explored. Because CLIP is trained on large-scale natural image-text pairs and ViT requires a large number of training samples to achieve a comparable effect to CNN, the effect of CLIP (ViT) applied to remote sensing scenes may be inferior to CNN. Another reason is that VLCA is designed to produce more diverse captions, and some of them are comprehensive and accurate. Due to resource constraints, the reference captions are not annotated with a comprehensive description of all contents in every image. The indicators measure the similarity between results and annotations. According to the evaluation index calculation principle, there is still a substantial difference between the image description index and manual evaluation. The evaluation index can be used as a reference, and the algorithm should be developed in conjunction with actual demand.

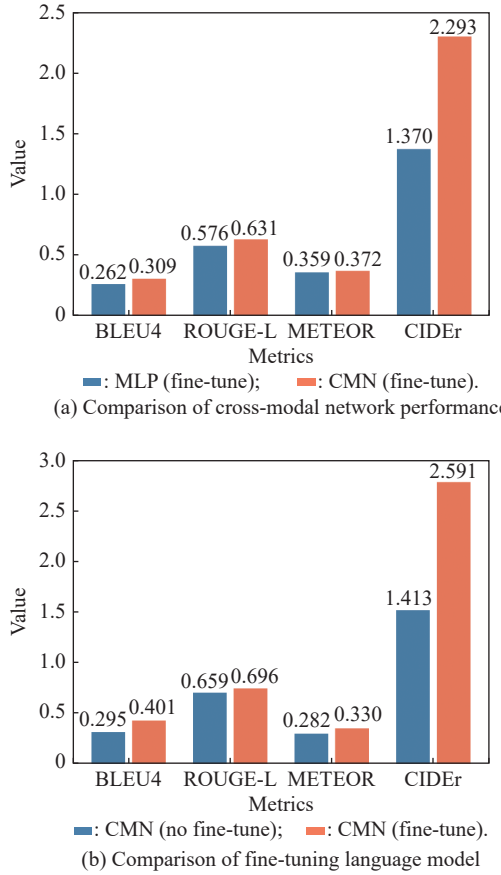
As a result of the aforementioned factors, the improvement of VLCA indicators is not high. However, it is important to note that VLCA is just as accurate as and even more thorough than those of other baselines. It is worth noting that the proposed framework is intended to investigate the effects of visual-linguistic large-scale image-text pairs pretraining in remote sensing. RSIC can benefit from the advancement of remote sensing pretraining technology. VLCA is scalable and can be easily extended to bilingual descriptions.

As shown in Fig. 5, VLCA contains more abundant information and can be seen intuitively from the fact that sentences generated by VLCA are longer than that of the baselines. This is because the nuclear sampling method is employed for sentence generation, which means that words with a low frequency have a chance of occurring. The baselines take beam search [37] as captioning strategy, which generates sentence by several words with the highest probability. The results generated by beam search are often simple and rigid and tend to describe commonsense in the image while ignoring important features. TopP increases the range of words selected and introduces more randomness. VLCA does have a few syntax errors, but this has little impact on the overall performance of the model.

4.4.2 Ablation experiments

Then, to test the role of network structure and fine-tuning pretrained model in VLCA, two sets of ablation experiments are conducted. The external attention module is replaced with two fully connected layers to test the effec-

tiveness of the proposed cross-modal attention module. This basic network can also map image and text features into the same space for language generation. Fig. 6 displays the comparison of each metric for two ablation experiments. To compare cross-modal network performance, ViT is used as backbone of VLCA, and ResNet50 is the backbone to compare the performance of the fine-tuning language model.



(a) Comparison of cross-modal network performance

(b) Comparison of fine-tuning language model

Fig. 6 Comparison of each metric for ablation experiments

Fig. 6(a) shows that the cross-modal learning module in VLCA outperforms MLP, which proves the effectiveness of the cross-modal learning module constructed in this paper. Fig. 6(b) illustrates the results achieved by fine-tuning the pretrained language model GPT-2. Although fine-tuning the language model can improve performance, growing training parameters will also increase computing resource consumption.

4.4.3 Extensive application

VLCA is extended to bilingual RSIC to demonstrate the actual application value of VLCA and DIOR-Captions in this part.

The difference between Chinese and English captions generation is the word tokenizer and language pre-training models. Jieba [38] is applied to word segmentation and the GPT-2 Chinese pre-training model [39] is used for text generation. The captioning mechanism in Chinese and English is essentially the same: the sentence is segmented into the smallest language unit token, and the token is then expressed as a vector for learning. Because VLCA has been tested and verified in English, only captioning results for Chinese description generation are shown. The results of English and Chinese captioning model are compared in this section, in which the output of English captioning model is translated into Chinese by Google and Youdao translate for comparison, as shown in Fig. 7, where CN refers to Chinese, and EN refers to English.

As shown in Fig. 7, the end-to-end Chinese captioning method outperforms translation software results. Because Chinese and English annotations in DIOR-Captions refer to each other, Chinese sentences for the same content are shorter than English sentences. In the task of RSIC, it is generally required to use concise sentences (usually limited to 20 words) to accurately describe the image. Therefore, for short sentences, translation software is easy to make mistakes due to lack of context information. The results of the translation may contain some imprecise nouns and sentences that do not adhere to Chinese grammatical conventions because of the absence of context. In addition, the bilingual description dataset and model can be used for cross-modal and cross-language remote sensing image retrieval in artificial intelligence (AI) systems.

In conclusion, the experiments show that VLCA is effective in RSIC tasks. The model not only performs well, but also the framework has good scalability. The cross-modal attention module can effectively realize cross-modal learning and modal transformation. Because the visual encoder is trained on the very large dataset collected from the internet and the linguistic decoder is also trained on large-scale text dataset, therefore when applied to a particular scenario, fine-tuning model can make the model speedily adapted to specific situations. To some extent, the disadvantage of RSIC requiring a large number of carefully annotated data sets is alleviated. At present, the large-scale pre-training technology of remote sensing images is immature, and the detection performance of many complex scenes needs to be improved. In future research, the encoder can be replaced by the remote sensing image pre-trained model, which will greatly improve the accuracy and comprehensiveness of descriptions generated.

Fig. 7 Chinese captions results of VLCA compared with Google and Youdao translate

5. Conclusions

In this paper, a solution for remote sensing image captioning is provided. We propose a new dataset DIOR-Captions and the problem of lack of RSICD has been alleviated to a certain extent. For remote sensing visual-semantic understanding, a framework and a cross-modal learning module is designed to achieve visual-sematic alignment. VLCA can generate bilingual descriptions well, which can also be applied to cross-language image retrieval. However, VLCA still has some limitations, such as the implicit alignment of visual and linguistic features is uninterpretable. The scene graph generation approach is expected to be introduced for reasoning to generate descriptions explicitly in the future.

References

[1] GAO L N, BI F K, YANG J. Visual attention based model for target detection in large-field images. Journal of Systems Engineering and Electronics, 2011, 22(1): 150–156.

[2] MOGADALA A, KALIMUTHU M, KLAKOW D. Trends in integration of vision and language research: a survey of tasks, datasets, and methods. Journal of Artificial Intelligence Research, 2021, 71: 1183–1317.

[3] QU B, LI X L, TAO D C, et al. Deep semantic understanding of high resolution remote sensing image. Proc. of the IEEE International conference on Computer, Information and Telecommunication Systems, 2016. DOI: 10.1109/CITS.2016.7546397.

[4] YUAN Z Q, ZHANG W K, TIAN C Y, et al. Remote sensing cross-modal text-image retrieval based on global and local information. IEEE Trans. on Geoscience and Remote

Sensing, 2022, 60: 5620616.
[5] SHI Z W, ZOU Z X. Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans. on Geoscience and Remote Sensing, 2017, 55(6): 3623–3634.
[6] LU X Q, WANG B Q, ZHENG X T, et al. Exploring models and data for remote sensing image caption generation. IEEE Trans. on Geoscience and Remote Sensing, 2017, 56(4): 2183–2195.
[7] WANG B Q, LU X Q, ZHENG X T, et al. Semantic descriptions of high-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters, 2019, 16(8): 1274–1278.
[8] LU X Q, WANG B Q, ZHENG X T. Sound active attention framework for remote sensing image captioning. IEEE Trans. on Geoscience and Remote Sensing, 2019, 58(3): 1985–2000.
[9] ZHAO R, SHI Z W, ZOU Z X. High-resolution remote sensing image captioning based on structured attention. IEEE Trans. on Geoscience and Remote Sensing, 2021, 60: 5603814.
[10] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision. Proc. of the International Conference on Machine Learning, 2021: 8748–8763.
[11] SHEN S, LI L H, TAN H, et al. How much can CLIP benefit vision-and-language tasks? http://arxiv.org/abs/2107.06383.
[12] LU J S, GOSWAMI V, ROHRBACH M, et al. 12-in-1: multi-task vision and language representation learning. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10437−10446.
[13] MOKADY R, HERTZ A, BERMANO A H. Clipcap: clip prefix for image captioning. https://arxiv.org/abs/2111.09734.
[14] DU Y F, LIU Z K, LI J Y, et al. A survey of vision-language pre-trained models. https://arxiv.org/abs/2202.10936.
[15] ZHOU L W, PALANGI H, ZHANG L, et al. Unified vision-language pre-training for image captioning and VQA. Proc. of the AAAI Conference on Artificial Intelligence, 2020, 34: 13041−13049.
[16] LI K, WAN G, CHENG G, et al. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296–307.
[17] ZHANG X R, WANG X, TANG X, et al. Description generation for remote sensing images using attribute attention mechanism. Remote Sensing, 2019, 11(6): 612.
[18] WU S Q, ZHANG X R, WANG X, et al. Scene attention mechanism for remote sensing image caption generation. Proc. of IEEE the International Joint Conference on Neural Networks, 2020. DOI: 10.1109/IJCNN48605.2020.9207381.
[19] LI X J, YIN X, LI C Y, et al. Oscar: object-semantics aligned pre-training for vision-language tasks. Proc. of the European Conference on Computer Vision, 2020: 121–137.
[20] LU J S, BATRA D, PARIKH D, et al. Vilbert: pretraining task-agnostic vision linguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 2019. DOI: 10.48550/arXiv.1908.02265.
[21] MIYAZAKI T, SHIMIZU N. Cross-lingual image caption generation. Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016: 1780−1790.
[22] LI X R, XU C X, WANG X X, et al. COCO-CN for cross-lingual image tagging, captioning, and retrieval. IEEE Trans. on Multimedia, 2019, 21(9): 2347–2360.
[23] WANG B, WANG C G, ZHANG Q, et al. Cross-lingual image caption generation based on visual attention model. IEEE Access, 2020, 8: 104543–104554.
[24] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners. OpenAI Blog, 2019, 1(8): 9.
[25] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale. https://arxiv.org/abs/2010.11929.
[26] HE K M, ZHANG X, REN S Q, et al. Deep residual learning for image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017: 6000–6010.
[28] GUO M H, LIU Z N, MU T J, et al. Beyond self-attention: external attention using two linear layers for visual tasks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022. DOI: 10.1109/TPAMI.2022.3211006.
[29] GUO M H, CAI J X, LIU Z N, et al. Pct: point cloud transformer. Computational Visual Media, 2021, 7(2): 187–199.
[30] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration. https://arxiv.org/abs/1904.09751.
[31] DIEDERIK P K, BA J. Adam: a method for stochastic optimization. http://arxiv.org/abs.1412.6980.
[32] LUO R T. Image captioning. https://github.com/ruotianluo/ImageCaptioning.pytorch.
[33] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation. Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311−318.
[34] LIN C Y. 2004. ROUGE: a package for automatic evaluation of summaries. Proc. of the ACL-04 Workshop, 2004: 74–81.
[35] DENKOWSKI M, LAVIE A. Meteor universal: language specific translation evaluation for any target language. Proc. of the Ninth Workshop on Statistical Machine Translation, 2014: 376−380.
[36] VEDANTAM R, LAWRENCE ZITNICK C, et al. Cider: consensus-based image description evaluation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4566–4575.
[37] HANNUN A Y, MAAS A L, JURAFSKY D, et al. First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. https://arxiv.org/abs/1408.2873.
[38] SUN J Y. Jieba. https://github.com/fxsjy/jieba.
[39] DU Z Y. GPT2-Chinese. https://github.com/Morizeyao/GPT2-Chinese.

Biographies

WEI Tingting was born in 1997. She received her M.S. degree from the College of Intelligence Science and Technology, National University of Defense Technology. She is pursuing her Ph.D. degree in National University of Defense Technology. Her research interests are pattern recognition and knowledge inference.
E-mail: weitingting20@nudt.edu.cn

YUAN Weilin was born in 1994. He received his M.S. degree in control science and engineering from National University of Defence Technology, where he is pursing his Ph.D. degree. His research interests include cognitive decision-making and intelligent gaming, reinforcement learning, and multi-agent system.
E-mail: yuanweilin12@nudt.edu.cn

LUO Junren was born in 1989. He received his M.S. degree in control science and engineering from National University of Defense Technology where he is pursuing his Ph.D. degree. His research interests include multi-agent learning and game confrontation.
E-mail: luojunren17@nudt.edu.cn

ZHANG Wanpeng was born in 1981. He received his M.S. and Ph.D. degrees in control science and engineering from National University of Defense Technology (NUDT). He is a professor at NUDT . His research interests include mission planning and intelligent control.
E-mail: wpzhang@nudt.edu.cn

LU Lina was born in 1984. She received her Ph.D. degree in control science and engineering from National University of Defense Technology (NUDT). She is a lecturer at NUDT. Her main research interests include machine learning, and multi-agent cooperation and confrontation.
E-mail: lulina16@nudt.edu.cn