Visual Question Answering of Remote Sensing Image Based on Attention Mechanism

Shihuai Zhang, Qiang Wei, Yangyang Li, Yanqiao Chen and Licheng Jiao

July 2, 2022

Visual Question Answering of Remote Sensing Image Based on Attention Mechanism

Shihuai.Zhang[1], Qiang.Wei[1], Yangyang.Li[1*], Yanqiao.Chen[2], and Licheng.Jiao[1]

[1] the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, Collaborative Innovation Center of Quantum Information of Shaanxi Province, School of Artificial Intelligence, Xidian University, Xi'an 710071, China

[2] the Key Laboratory of Aerospace Information Applications, The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China

*yyli@xidian.edu.cn

Abstract. In recent years, the research of attention mechanism has made significant progress in the field of computer vision.In the processing of visual problems of remote sensing images, the attention mechanism can make the computer focus on important image areas and improve the accuracy of question answering.Our research focuses on the role of synergistic attention mechanisms in the interaction of question representations and visual representations. On the basis of Modular Collaborative Attention (MCA), according to the complementary characteristics of global features and local features, the hybrid connection strategy is used to perceive global features at the same time without weakening the attention distribution of local features.The impact of attention mechanisms on various types of visual question answering questions has been evaluated:(i) scene classification (ii)object comparison (iii) quantitative statistics (iv) relational judgment.By fusing the global features and local features of different modalities, the model can obtain more information between modalities. Model performance evaluation under the RSVQA-LR dataset. Experimental results showthe method in this paper improves the global accuracy by 9.81% than RSVQA.

Keywords: Co-attention · VQA · feature fusion.

1 Intruduction

Visual Question Answering (VQA) is an emerging task in computer visionSimply put, it is given a question about the content of the image, and the computer gives an appropriate answer based on the text question and the input image. The problem can be split into sub-problems in computer vision, such as:

- Image classification - what area is this?
- Object comparison – is the farmland on the left larger or the farmland on the right larger?

- Quantity statistics - how many buildings are there?
- Relationship Judgment - Is there a small river next to the residential area?

Not only these, the VQA task also contains complex questions. Answering these questions requires certain thinking logic. The questions contain the spatial relationship between the targets, such as (how many villas are around this beach?). And some about behavioral reasoning, such as (Why is this boy running?). These complex problems require the VQA model to have a certain reasoning ability like a human.

In real life, VQA is useful for a wide range of application scenarios. As an open question answering system, it can obtain the specific information in the image in many aspects. For the blind, VQA can describe the content of the image in detail and realize the interaction between virtual and reality. It can also complete text-to-image retrieval through its analysis of image attributes. Usually, the information contained in the label of an image cannot cope with all scenarios. For example, to find a summer landscape of the Gold Coast, just ask the question (is there a coast, sand, hot sun?) instead of looking for answers from irrelevant tags. Therefore, VQA can improve the experience of human-computer interaction without relying on image labels.

Remote sensing image data is widely used in our daily life. Whether land detection, environmental protection, resource exploration, and urban monitoring require the direct or indirect participation of remote sensing images. For these applications, professional processing of remote sensing images is required, and effective information cannot be obtained directly from images. Conventional remote sensing image processing tasks only start from a single task (such as classification, detection, segmentation), Only a small number of experts can directly process this information. A VQA task for remote sensing images can directly obtain answers to questions related to remote sensing images, which greatly reduces the threshold for using remote sensing images in other fields.

Deep neural networks models learn a relational representation to generate a mapping from input to outputWhile the answer to the VQA task is open-ended (e.g. what's next to the beach? The answer can be a villa, a crowd, a car, etc.), This means that more attention needs to be paid to the potential connections in the question sentence, which is a difficult problem for the network model to learn. On the contrary, as long as we pay attention to the questions given in the VQA task, we can focus on the two key words of beach and beside, reducing the interference of other unnecessary information in the sentence. This paper adopts the strategy of stacking attention units and text features to guide visual features, and uses the category information in the question features to focus on relevant regions in the image.

For the VQA task of natural images, it is necessary to extract features and eliminate redundant information in the visual channel. Natural images cover a small range of realistic scales, generally ranging from a few meters to hundreds of meters. However, for remote sensing images, the scene information contained in them covers far more ground objects than natural images, and the number of targets in a scene is several times that of natural images. In answering visual questions about remote sensing images, there are both macroscopic global problems and microscopic details. Therefore, it is necessary to focus on both local features and global perception. This paper adopts the strategy of cross-modal feature fusion, which combines global information with local information, makes full use of the complementarity of information between modalities, and solves the shortcomings of insufficient global perception of attention-based visual question answering models.

To this end, we propose a model for the remote sensing image VQA task and complete the performance evaluation on the RSVQA-LR dataset. In Section III, HMCAN (Hybrid Moudlar Co-attention Network) is introduced in detail. In Section IV, HMCAN is evaluated and discussed on the RSVQA-LR dataset. The contributions of this paper are as follows:

- A Modular Collaborative Attention Mechanism with Hybrid Connections
- Cross-modal global feature and local feature fusion strategy

2 Related Work

With the development of deep neural networks, the visual question answering task also adopts the method of stacking deep network models to obtain deep features. Relying on the powerful feature extraction ability of deep networks, only using global features alone can achieve a good effect. However, this model blurs the task-related areas of the input information, and the model with attention mechanism can effectively overcome this shortcoming, these models have achieved great success in natural language processing (NLP) and other computer vision tasks, Such as object recognition [1], subtitle generation [13] and machine translation [10, 2].

I. Ilievski [5] et al proposed FDA,In this paper, the bounding box with object labels is obtained through ResNet, using word2vec [11] to calculate the similarity between the words in the question and the bounding box, and the region proposal boxes relevant to the question are generated. In [14], to emphasize the global image representation, the authors propose a stackable attention layer, using Softmax with CNN(Convolutional Neural Network) to compute the attention distribution across image locations in a single layer by weighting and focusing on spatially important locations. C. Xiong [12] et al. added a dynamic memory network DMN [6] to the VQA task, the text is fed into the recurrent neural network to generate "fact" features, and the CNN image features of each region are regarded as the words in the sentence and sequentially sent to the recurrent neural network to generate visual "fact" features. Finally, the answer is predicted from the text and image "facts".

In recent years, the neural network model based on the transformer structure has been widely used in the field of computer vision, and many scholars [7, 3] introduced the transformer into the VQA task. In [9], the authors add a cross attention layer and a collaborative attention layer based on a two-way transformer to fully learn the context content and achieve efficient exchange of information between modalities. This paper mainly studies the attention distribution method that uses text features to generate image features, and uses text representations to guide the learning of image features.

2.1 Co-Attention

Scaled Dot-Product Attention The scaled dot-product attention on the left side of Figure 1 is a normalized dot product attention, where Q, $K$and V represent *query, key* and *value*, respectively.

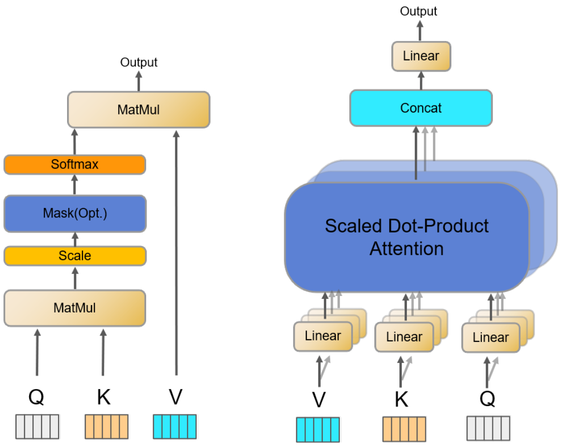

Fig. 1. The scaled dot-product attention model on the left, and the multi-head attention model on the right consists of several parallel attention layers.

Assuming that the input *query* and *key* dimensions are d_k , and the *value* dimension is d_v , first calculate the dot product of the query, and each *key* and divide by $\sqrt{d_k}$, and then apply the *softmax* function to calculate the weight.

$$Attention\left(Q, K_i, V_i\right) = softmax\left(\frac{Q^T K_i}{\sqrt{d_k}}\right) V_i \tag{1}$$

In the actual operation, each *query*, *key* and *value* are processed into Q,K and V matrices respectively.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{2}$$

Where $Q \in R^{m \times d_k}, K \in R^{m \times d_k}, V \in R^{m \times d_v}$,the dimension of the output matrix is $R^{m \times d_v}$.

Multi-head-attention In the scaled dot-product attention, only one weight operation is performed on Q, K and V, which is obviously not sufficient. Therefore, Multi-head-attention adopts the method of parallel splicing of multiple attention layers to strengthen the attention generation of global features. First perform linear mapping on the Q, K and V matrices, and convert the input dimension from d_{model} to $Q \in R^{m \times d_k}, K \in R^{m \times d_k}, V \in R^{m \times d_v}$.Then the mapping is sent into the attention in parallel, the formula is as follows:

$$MultiHead(Q, K, V) = Concat\left(head_1, \ldots, head_h\right) W^0 \tag{3}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{4}$$

Among them, the weight matrix $W_i^Q \in R^{d_{model} \times d_k}, W_i^K \in R^{d_{model} \times d_k}, W^O \in R^{hd_v \times d_{model}}, W_i^V \in R^{d_{model} \times d_v}$.In this part, h represents the number of parallel attention layers, so the conversion relationship of the input dimension is: $d_k = d_v = d_{model}$ /h , and finally a splicing operation is performed at the end, so that the input and output dimensions are consistent.

Self-attention and guided-attention Self-attention (SA) is based on Multi-head attention, adding residual connections between outputs, adding a normalization layer and a fully connected forward network. This forward network consists of two linear layers, in which the activation The function is Relu, and the formula is as follows:

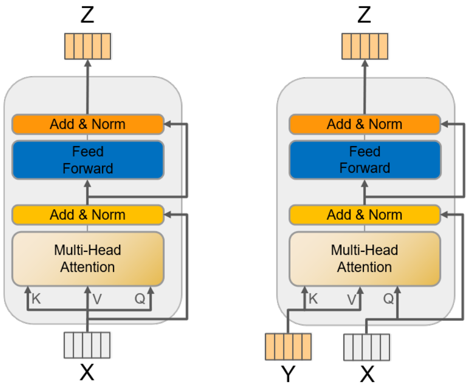

Fig. 2. The left side is the self-attention with multi-head attention as the main structure, and the right side is the guide-attention that introduces the dual modalities of text and images.

$$FFN(x) = \max\left(0, xW_1 + b_1\right) W_2 + b_2 \tag{5}$$

W_1,W_1 adjust the input and output dimensions to d_{model} =2224, and the input feature of the middle layer is d_{ff}=2048.

As shown in Figure 2 (right), Guide-attention (GA) introduces another input, Y and X correspond to text input and image input, respectively, where $X \in R^{m \times d_x}$, $Y = [y_1; \ldots; y_n] \in R^{n \times d_y}$.The method of obtaining attention distribution by SA is used here, and the extraction of image features is guided by the input queryand keyof the text, and then the GA model is established according to the feature correlation of $< x_i, y_j >$,On the basis of the original, the interaction between modalities has been increased.

Moudlar Co-attention Network The modular collaborative attention network (MCAN) in Figure 4 (left) is composed of multiple modular collaborative attention layers $\left(MCA^{(1)}, MCA^{(2)}\right)$ stacked. Each MCA layer is composed of one SA unit in the text branch and one SA and GA unit in the image branch. Suppose that the input $Y^{(0)}, X^{(0)}$ correspond to text features, image features Y,X, respectively, then the single-layer output is as follows:

$$\left[X^{(l)}, Y^{(l)}\right] = MCA^{(l)}\left(\left[X^{(l-1)}, Y^{(l-1)}\right]\right) \tag{6}$$

The input of each layer of MCA is the output of the upper layer of MCA.

3 Methods

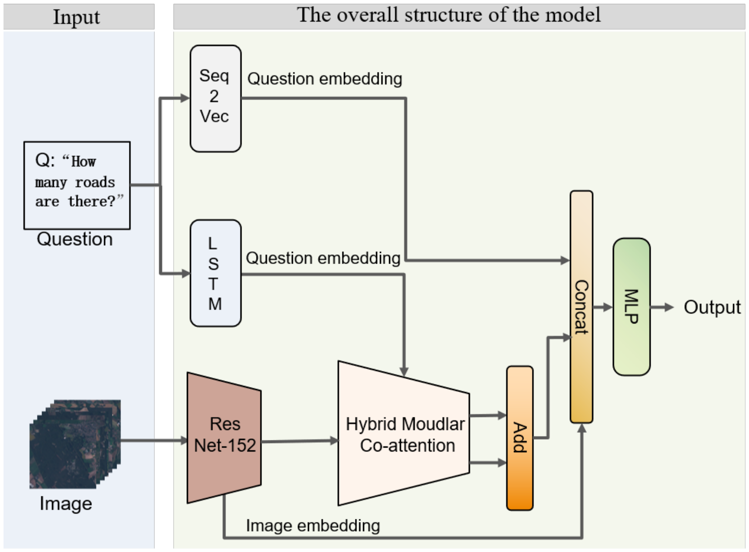

Fig. 3. The overall structure of the model.

We propose a novel structure for visual question answering for remote sensing data using a hybrid-connected Modular Collaborative Attention (HMCA) component, which is outlined in Figure 4. The model takes question representation and image representation as input, uses a cross-layer attention-guided strategy to guide image representation with question representation, and finally implements visual question answering through a multi-layer perceptron (MLP). The detailed description of the components is expanded next.

3.1 Hybrid Moudlar Co-Attention

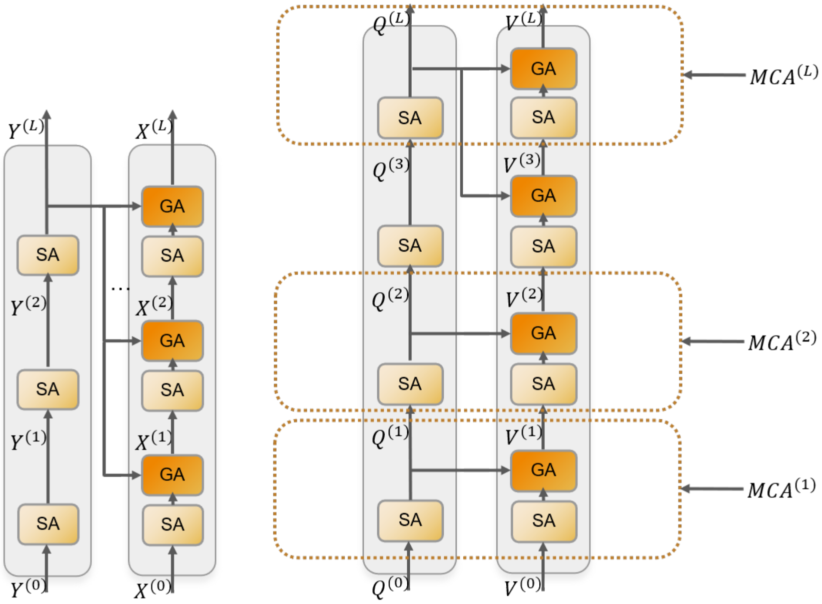

Fig. 4. The left figure shows MCA and the right figure shows hmca with hybrid connection strategy.

In the MCA unit, the output of the last SA block of the Y branch is used as the input of the GA block of the X branch $\left[X^{(0)}, X^{(1)} \ldots X^{(L)}\right]$. On the whole, the text features with deep attention distribution can guide the image features effectively, making the network pay more attention to the text-related regions in the image, which is a good strategy. But at the same time, this approach will ignore some global information in the text features. In HMCA, the Y and X branches in MCA are replaced by the question representation Q and the visual representation V, where the number of stacking layers of attention layers is four, Using the parallel connection method in the first two layers, in $MCA^{(1)}, Q^{(1)}$ is used as the guide of $V^{(0)}$ to obtain $V^{(1)}$. The formula is as follows:

$$\left[V^{(l)}, Q^{(l)}\right] = MCA^{(l)}\left(\left[V^{(l-1)}, Q^{(l-1)}\right]\right) \tag{7}$$

In the first two layers of CMCA, the problem-based shallow attention distribution is used to guide the visual representation, which ensures that the visual features have a global perception of the problem category, and the extreme distribution of the attention area will not occur due to the attention distribution. In the last two layers, the method of *MCA* is used:

$$\left[V^{(i)}, Q^{(i)}\right] = MCA^{(i)}\left(\left[V^{(i-1)}, Q^{(l)}\right]\right) \tag{8}$$

3.2 Cross-modal fusion of global and local information

It is found that after adopting the CMAC component, the overall effect has been significantly improved, but it has a negative impact on specific problem categories, and has a good performance in such problems as quantity calculation, area estimation, and target comparison. However, the accuracy of the global scene judgment problem (is the area to the left of the forest urban or rural?) has declined. Due to the highly concentrated attention distribution obtained by the CMAC component, part of the global information is ignored, and the global features have not received due attention in large scene problems. Therefore, a cross-modal fusion strategy that combines global information with local information is deliberately adopted. In the question feature extraction stage, LSTM and Seq2Vec are used as two-way branches to extract the embedding vector of the question, respectively. The output of the LSTM branch is sent to CMCA as a guide, and the other branch that retains the global problem information is spliced and fused with shallow visual features and deep visual features.

Finally we map the 2224-dimensional vector to the answer space using a single hidden layer MLP with 512 hidden layer units. In the answer output stage, the open-ended question answer is transformed into a classification question, and each possible answer is classified into a category, so the dimension of the output is related to the question.

4 Experiments

4.1 Dataset

The dataset used in the experiments is RSVQA (Real Dataset for Remote Sensing Visual Question Answering Task), which is based on images acquired by Sentinel-2 satellite over the Netherlands.The Sentinel-2 satellite provides imagery at 10-meter resolution (for the visible band used in this dataset) globally and is updated frequently (approximately 5 days).The RSVQA dataset is a real remote sensing image of the Netherlands, which includes complex terrains such as mountains, hills, rivers, coasts, towns, rural areas, and wasteland. Since the dataset covers most geographical scenes, the richness of its geographic information is conducive to the migration of this method to other application scenarios.

Fig. 5. RSVQA datasets.

4.2 Experimental setup and hyperparameters

Our experimental setup and model hyperparameter details are as follows: Using pretrained Resnet-152 [4] for the image channel, Using pretrained Seq2Vec in one branch of the problem channel, the other branch needs to be trained by itself. Adam is used as the optimizer in the experiment, the learning rate is 0.0001, and the number of iterations is 20 epochs. Overall classification accuracy (OA), average classification accuracy (AA), and various classification accuracies are used in performance evaluation.

4.3 Experimental results and analysis

Table 1. Comparison of the performance of the two models on the RSVQA-LR dataset using RSVQA as the baseline.

TypeModel	RSVQA[8]	HMCAN	Improvement
Count	67.01%	71.80%	↑ 4.79%
Presence	87.46%	91.45%	↑ 3.99%
Comparison	81.50%	93.01%	↑ 11.51%
Rural/Urban	90.00%	91.00%	↑ 1.00%
AA	81.49%	86.81%	↑ 5.32%
OA	79.08%	88.89%	↑ 9.81%

In Table 1, the performance of our method is presented. On this problem category based on comparison, our method achieves a significant improvement, which is 11.51% higher than the baseline. Due to the use of a pretrained network, 20 epochs ran for four hours on an NVIDIA 1080 configuration. On the overall accuracy OA, it is 9.81% higher than the benchmark.

5 conclusion

In this work, we propose a hybrid connection unit based on a modular collaborative attention mechanism, employing different connection strategies at different positions of the stacked attention layers. In feature fusion, we adopt a cross-modal feature fusion strategy to obtain global and local information between modalities. For the remote sensing visual question answering task, our model outperforms previous models on the RSVQA-LR dataset. In the future, the attention mechanism should have more room for development. The performance of a model largely depends on the quality of the data. The next step is to apply the attention mechanism to remote sensing images with high noise,and compress the network as much as possible to realize a lightweight and highly portable remote sensing visual question answering model.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61772399 and Grant 62101517, in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY09-05, and in part by the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project).

References

1. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755 (2014)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Ilievski, I., Yan, S., Feng, J.: A focused dynamic attention model for visual question answering. arXiv preprint arXiv:1604.01485 (2016)
6. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: International conference on machine learning. pp. 1378–1387. PMLR (2016)
7. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)
8. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **58**(12), 8555–8566 (2020)
9. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)

10. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
12. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: International conference on machine learning. pp. 2397–2406. PMLR (2016)
13. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
14. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)