

Remote sensing image scene classification based on stackable attention structure

Haonan Zhou^{1, a}, Lurui Xia^{2, b}, Xiaoping Du^{2, c}, and Sen Li^{2, d}

¹ Space Engineering University, China;

^azhouhaonan011@163.com, ^bxlurui522@163.com, ^cdxp8600@163.com, ^dls_space@163.com

Abstract. This work proposes a remote sensing image scene classification method based on stackable attention structure. This method utilizes a stackable attention structure to represent the feature information in remote sensing images as a linear function, changing the previous processing methods for global and local information in remote sensing images to avoid the need for memory intensive attention maps. Through experiments, it has been shown that the scene classification method we proposed based on stackable attention structure has higher accuracy and smaller model size than ResNet152. This provides a new approach for constructing remote sensing image scene classification models, especially for processing large-sized remote sensing images.

Keywords: remote sensing images; scene classification; attention mechanism; stackable structure

1. Introduction

Remote sensing imagery is characterized by its high amplitude and resolution. The data volume of a single remote sensing image can reach tens of gigabytes, with millions of pixels. And the target to be processed is usually small and relatively hidden in the whole remote sensing image. The characteristics of remote sensing image data limit the usual remote sensing scene classification (RSC) algorithms, which are key links in remote sensing image interpretation. The commonly used methods are mostly based on convolutional neural networks. Convolutional neural networks are a good inductive bias for extracting local feature information, but their speed for extracting global feature information is slow and their effectiveness is poor. Therefore, when computer vision processing is applied to a remote sensing image, attention mechanisms are usually added to improve the processing ability of convolutional neural networks.

Attention mechanism is a commonly used paradigm for extracting large-scale feature information in images. The use of attention mechanisms can enable convolutional neural networks to better extract local and global feature information in images, allowing algorithms to focus attention on specific local regions of the image, greatly improving processing speed and results. Typical methods based on attention mechanisms include Faster R-CNN [1], which has achieved excellent results; In recent years, the self-attention mechanism has been widely applied. The Detection Transformer (DETR) [2] applied the self-attention mechanism to object detection tasks and achieved results comparable to Faster R-CNN; OpenAI has released the iGPT model for image classification [3], which trains GPT-2 on images to understand 2D image features such as object appearance and category. The features generated by this model have achieved great performance on several classification datasets and good image completion results; Vision Transformer (ViT) [4] approaches or exceeds the SOTA level on several image recognition benchmarks.

However, the introduction of attention mechanisms and self-attention can increase the size of the trained models, with some models reaching GB in size. In addition, the attention mechanism generates a large number of feature maps and attention maps during the computation process, occupying a large amount of memory. This is very unfavorable for the application and deployment of the model in practical scenarios, and hinders the development of related fields.

Therefore, we propose a method based on a stackable attention mechanism. It based on lambda layer [5] and coordinate attention [6] to optimize the classical classification model ResNet [7]. Lambda layer can replace the self-attention mechanism to capture global features in remote sensing images. In addition, coordinate attention is added to further improve the network's ability to extract

features from remote sensing images. Experimental results show that our proposed model has higher classification accuracy and computational efficiency, smaller model size, and flexible network structure compared to ResNet.

2. Scene classification based on stackable attention structure

2.1 Attention structure

The purpose of using attention mechanisms in computer vision tasks is to enable models to extract key information from a large amount of information and focus on it. The characteristics of large amplitude and sparse information in remote sensing images require attention mechanisms to capture the internal correlations of global features. To achieve this feature, attention structures are usually introduced, and lambda layers are one of them.

For each query point, the lambda layer calculates the relationship between it and other points, and then aggregates them through weighted summation to obtain global features related to that query point. Add global features to the features of each query point separately to complete long-range dependency modeling. The attention structure usually projects input X and context C separately:

$$Q = XW_Q \quad (1)$$

$$K = CW_K \quad (2)$$

$$V = CW_V \quad (3)$$

W_K and W_V are the linear projection tensor of context C, and W_Q is the linear projection tensor of input X. This calculation process can be seen as a process of extracting feature information.

2.2 Stackable attention structure

In the process of extracting key feature information through attention mechanisms, the proportion of contextual information in different channel domains varies. This article uses coordinated attention and lambda layers to form a stackable attention structure for scene classification. Coordinate attention is an efficient attention mechanism that embeds position information into channel attention, encoding the features of the input image into the attention map in both vertical and horizontal directions, making it easier for the model to find key features while avoiding significant computational overhead. By using coordinate attention processing on the contextual information P in different channel domains, the proportion of different channel domains is obtained:

$$\bar{P} = \text{coordinate attention}(P) \quad (4)$$

At the same time, using the lambda layer, a function can extract information about the image context content from each local range:

$$\mu_m^c = \overline{K_m P_m} V_m^T \quad (5)$$

We also can extract a function for contextual position information about images:

$$\mu_{nm}^p = E_{nm} V_m^T \quad (6)$$

By extracting the contributions of all local contextual regions according to the above method, a parameter λ_n is obtained:

$$\lambda_c = \sum_m \mu_m^c = \sum_m \overline{K_m P_m} V_m^T \quad (7)$$

$$\lambda_n^p = \sum_m \mu_{nm}^p = \sum_m E_{nm} V_m^T \quad (8)$$

$$\lambda^n = \alpha \lambda_c + \lambda_n^p \quad (9)$$

α is a hyperparameter. The relationship between the output and input of the lambda layer can be expressed as a linear function as follows:

$$y_n = \lambda_n q_n = [\alpha \lambda_c + \lambda_n^p] q_n = [\alpha \lambda_c + \lambda_n^p] W_Q x_n = \sum_m [E_{nm} + \alpha \overline{K_m P_m}] V_m^T W_Q x_n \quad (10)$$

The above process only generates a linear function, and does not generate attention maps, greatly reducing the occupation of computer resources. Divide the input remote sensing image into local regions, and use the linear functions mentioned above to extract feature information within the local range, respectively, to obtain a linear function. Because it is a linear function and different local regions of the image are independent to each other, it is possible to directly stack and integrate the feature information of each local region to obtain the output:

$$y = \lambda_1 q_1 + \lambda_2 q_2 + \lambda_3 q_3 + \dots + \lambda_{t-1} q_{t-1} + \lambda_t q_t \quad (11)$$

Calculate linear functions separately in each local area of remote sensing images, reduce the size of the calculation object of attention mechanism, and improve the efficiency of extracting local feature information within the local range. The entire calculation process is dynamic, and only linear functions are saved at the end of the calculation. The linear functions extracted from different regions will be integrated through superposition, which will greatly reduce the cost of attention mechanisms. The use of linear functions makes the entire network highly scalable, especially in the processing of high-resolution and large-sized images, as the model can process different parts of remote sensing images separately and then stack them, rather than processing the entire remote sensing image. Based on the characteristics of the method, we named it co-lambdaNet, and the schematic diagram of the algorithm is shown in Figure 1.

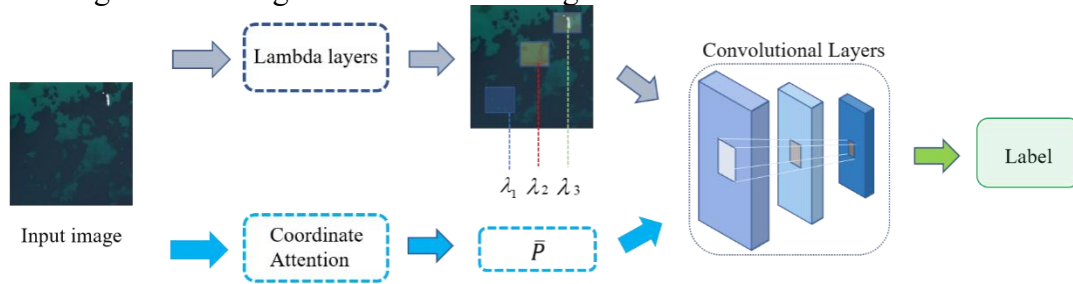


Fig. 1 Schematic diagram of co-lambdaNet

3. Experiment

This section introduces the dataset used to train co-lambdaNet152, the parameters set during model training, and finally analyzes the experimental results.

3.1 Dataset

We chose the v2 version of the MARITIME satellite image dataset to train and test the model. This dataset contains satellite images of 7 different visible light sea scenes, totaling 7337 samples. These images are all selected from Microsoft's Bing Maps in PNG format with a resolution of 512 x 512.

3.2 Model training parameters

During model training, the size of the remote sensing image will be adjusted to 224x224 when input into the model. The activation function is ReLU, and the batchsize is set to 64. The optimizer is AdamW, the weight decay is set to 0.01, and the learning rate is 0.0025. During training, the pytorch framework was used, with a graphics card of RTX 3090 24GB, a CPU of i9-10900k, a hard drive of 2TB, and an operating system of ubuntu 20.04. It has trained a total of 300 epochs.

3.3 Experimental results and analysis

ResNet-152 is a widely used structure for RSC. Although it has a simple structure, it can achieve excellent performance and is often used as a foundation for improvement and comparison; Our proposed co lambdaNet152 is built on the ResNet-152 structure, so we chose ResNet-152 as the

basic model for comparison in the experiment to better demonstrate the effectiveness of our series of designs.

We compared ResNet-152 and co-lambdaNet152 in the same dataset and environment, focusing on their accuracies and model size. The comparison results are shown in the table below.

Table 1. Comparison of experimental results of scene classification models

Model	Classification accuracy %	Model size MB
ResNet-152	86.5	377
co-lambdaNet152	86.7	152

From the comparison results, it can be seen that the co lambdaNet152 designed in this article has a higher scene classification accuracy than ResNet-152, indicating the effectiveness of the design in co lambdaNet152 for extracting global features; The size of the co lambdaNet152 model is much smaller than that of the ResNet-152 model. That is to say, co lambdaNet152 has a smaller model size of ResNet-152 and achieves stronger scene classification performance.

4. Conclusion

This article proposes an RSC method based on stackable attention structure, which includes lambda layer and coordinate attention. This stackable attention structure enables classification models to better capture global features in remote sensing images and significantly reduce the size of the model, achieving better classification performance. The stackable attention structure can process different parts of remote sensing images separately, and then stack them through simple addition, which is more advantageous for processing larger remote sensing images. At the same time, this structure can be easily expanded and optimized. In subsequent research, this structure can be used as a basis for downstream tasks to promote the development of intelligent interpretation of remote sensing images.

References

- [1] Ren, Shaoqing, Kaiming He, Ross B. Girshick and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015): 1137-1149.
- [2] Zhu, Xizhou, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang and Jifeng Dai. "Deformable DETR: Deformable Transformers for End-to-End Object Detection." *ArXiv abs/2010.04159* (2020): n. pag.
- [3] Chen, Mark, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan and Ilya Sutskever. "Generative Pretraining From Pixels." *International Conference on Machine Learning* (2020).
- [4] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ArXiv abs/2010.11929* (2020): n. pag.
- [5] Bello, Irwan. "LambdaNetworks: Modeling Long-Range Interactions Without Attention." *ArXiv abs/2102.08602* (2021): n. pag.
- [6] Hou, Qibin, Daquan Zhou and Jiashi Feng. "Coordinate Attention for Efficient Mobile Network Design." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 13708-13717.
- [7] He, Kaiming, X. Zhang, Shaoqing Ren and Jian Sun. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 770-778.