Improved AlexNet and embedded multi-attention for remote sensing scene image classification

Dongfu Dai, Weiheng Xu*, Shaodong Huang
College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650233, Yunnan, China

ABSTRACT

In order to improve the overall accuracy (OA) of the AlexNet model for high-resolution remote sensing scene images with complex backgrounds, we proposed an improved remote sensing scene image classification model. Firstly, we used layer normalization (LN) to replace the local response normalization (LRN) in AlexNet and changed the convolution kernel of the first convolution layer to 7 × 7. Secondly, to focus on critical information in the feature extraction process, and suppress irrelevant background information, the two attention modules of convolution block attention module (CBAM) and squeeze and excitation module (SEM) were combined. In this study, the classification verification was performed on three remote sensing scene datasets of NWPU-RESISC45, AID, and UCM, and achieved 96.29%, 96.02%, and 96.57% overall accuracy, respectively. Compared with AlexNet, the OA improved by 14.38%, 12.09%, and 9.9%, respectively, therefore, the improved model of this study can significantly distinguish between object information and background information in remote sensing scene imagery.

Keywords: Overall accuracy, AlexNet, remote sensing scene, attention module

1. INTRODUCTION

Recently, With the continuous emergence of various advanced remote sensing data acquisition equipment, the resolution of remote sensing scene images become clearer, and the Complexity of remote sensing image background become more obvious. This makes early classification algorithms unusable for current high-definition remote sensing images[1-3]. Recently, deep learning represented by convolutional neural networks (CNN) is a new theoretical learning framework that has made significant breakthroughs in natural language processing, speech recognition, image processing, and other fields[4, 5]. Concurrent, CNN had also been applied to the classification of remote sensing scenes.

AlexNet[6], which achieved excellent image classification in the ImageNet competition in 2012, is also extensively used in the classification of remote sensing scenes. However, AlexNet still has many problems in the classification of remote sensing scene images, such as the limited number of scenes, and excessive compression of spatial information in the process of convolution. Therefore, many scholars had improved AlexNet. Xiao et al.[7] improved the AlexNet algorithm that can effectively alleviate the excessive compression of spatial information in the feature extraction process, and the classification accuracy of the improved AlexNet was improved by 8.81%. A pre-trained AlexNet architecture was proposed[8], which can effectively resolve the problem that the AlexNet model did not converge due to the few remote sensing scenes and limited labeled samples, and achieved classification accuracy of 96.67%. A method to fuse AlexNet and SVM was proposed to detect land cover change and classify different land cover types with 96% accuracy[9]. With the continuous development of deep learning, an attention mechanism that imitates human vision has been proposed, it can quickly scan the global image, bring the target area into focus, and obtain more critical information, thereby suppressing other useless information. This mechanism was universally used in the classification of remote sensing scenes and has achieved excellent results. Zhang[10] proposed a multi-scale attention module, which embedded channel attention and position attention modules, effectively suppressed the useless information of remote sensing scene images, and achieved an accuracy of 92.52% in NWPU-RESISC45. An enhanced attention module (EAM) was proposed[11], which effectively improved the ability of the CNN network to extract critical information, and the method achieved

* weihengx@gmail.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250638
 · 0277-786X · doi: 10.1117/12.2661951

a classification accuracy of 94.29% on the NWPU-RESISC45 dataset. Haikel[12] proposed a deep attention CNN for remote sensing scene classification and realized better classification accuracy on multiple datasets. Ji[13] proposed a method for remote sensing scene localization using attention networks, a method for regional discrimination in multiple regions, with features learned from local regions. A residual attention network was proposed that could clearly distinguish critical and redundant information in remote sensing images[14]. Chen[15] proposed a method for the scene classification of remote sensing images based on multi-branch local attention networks to better extract critical information. However, the above methods fail to significantly distinguish the target object of the remote sensing scene during feature extraction, and the critical feature extraction ability of the model needs to be further improved.

To significantly differentiate the object information and background information in the remote sensing scene images, this paper improved AlexNet and embedded a multi-attention module to raise the feature extraction ability of this model and the classification accuracy of the scene, and performed classification verification on three scene datasets of NWPU-RESISC45, AID, and UCM.

2. METHOD

2.1 The proposed network structure

AlexNet consists of a convolutional layer, a pooled layer, a fully connected layer, and an LRN. The size of the convolution kernel in the first two convolution layers is 11×11 and 5×5 respectively, and the size of the convolution kernel in the other convolution layers is 3×3. This study improved the AlexNet and embedded the multi-attention module. Firstly, LN was used instead of LRN, the convolution kernel with the size of 11×11 is replaced by 7×7 to relieve the excessive compression of spatial information in the model. Second, multi-attention modules are embedded to make the model more focused on the key information of the scene image. The improved network structure of this study is shown in Figure 1. The enhanced data were randomly separated into train set, validation set, and test set at an 8 to 1 to 1 ratio, after that, the model was trained and optimized using the training set and the validation set, and then the model was evaluated using the test set.

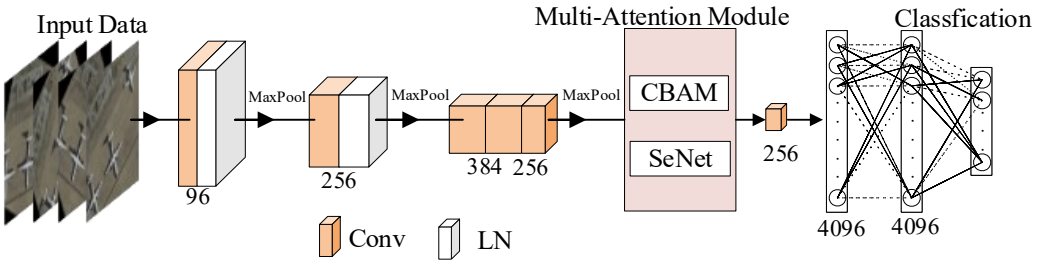

Figure 1. The network structure diagram of this study.

2.2 Layer normalization

LN was a method similar to batch normalization (BN) and was also a network layer with learnable parameters. LN can improve the generalization ability of CNN. Unlike the BN, LN was normalized of all neurons in a certain layer. Suppose there are n neurons in a layer, then the input of this layer is equation (1).

$$\{z_1^l, z_2^l, \ldots\ldots\ldots, z_n^l\} \tag{1}$$

In equation (1), z_n^l represents the n neurons in layer l.

(1) The mean of all neurons in layer l is calculated:

$$\mu^{(l)} = \frac{1}{n^{(l)}} \sum_{k=1}^{n^{(l)}} z_k^{(l)} \tag{2}$$

In equation (2), $n^{(l)}$ is the number of neurons at layer l.

(2) The variance of all neurons in layer l is calculated:

$$\sigma^{(l)^2} = \frac{1}{n^{(l)}} \sum_{k=1}^{n^{(l)}} \left(z_k^l - \mu\right)^2 \tag{3}$$

(3) The LN is calculated:

$$Z^{(l)} = \frac{z^{(l)} - \mu^{(l)}}{\sqrt{\sigma^{(l)} + t}} \gamma + \beta \tag{4}$$

In equation (4), γ and β are the scaling and translation coefficients of LN respectively, which are also the learning parameters of LN.

2.3 Fusion attention module

The fusion attention module of this study combined two attention modules of CBAM and SEM. The input features passed through the two attention modules of CBAM and SEM, respectively, then the outputs of CBAM and SEM were pooled by average pooling and stacked on the channel, the stacked features were convoluted with a 3×1 convolution kernel, as shown in Figure 2.

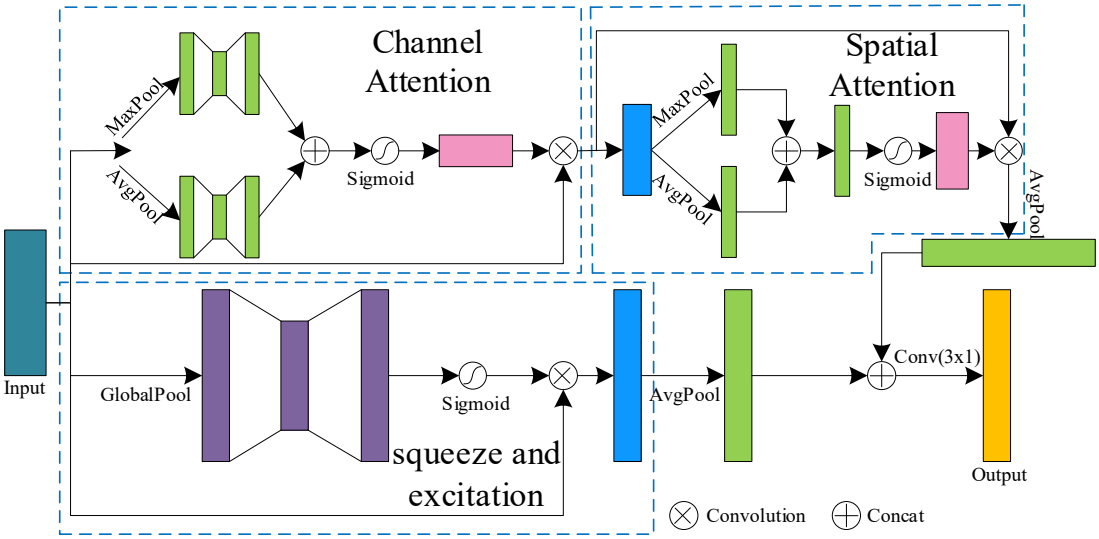

Figure 2. Fusion attention module.

3. EXPERIMENT

3.1 Dataset

The NWPU-RESISC45 dataset was a remote sensing scene data, collected by North Western Polytechnic University (NWPU). NWPU-RESISC45 includes altogether 31,500 images and covers 45 scene categories. Each scene contains 700 images. Each image is 256 × 256.

Aerial Image Dataset (AID), was jointly collected by Huazhong University of Science and Technology and Wuhan University. It contains a total of 10,000 images, including 30 scene categories. Each category has about 200-400 samples, and each sample is 600 × 600 pixels.

UC Merced Land Use (UCM), comes from the national map urban area images collected by the US Geological Survey with 21 categories of scene images. Each category has 100 images, a total of 2100 images. Each image is 256 × 256.

To ensure that the study model can be fully trained, the datasets of the three scenes were enhanced by rotation at 30, 60, 90, and 120 angles, then the enhanced image size was adjusted to 256×256 pixels.

3.2 Experimental design

The hardware environment for this experiment was Intel(R) Xeon(R), 4G memory, and NVIDIA GeForce RTX 1080Ti GPU for accelerated computing, the software environment was based on Tensorflow 2.3 and python3 version. The three remote sensing scene datasets of NWPU-RESISC45, AID, and UCM were data enhanced and divided into training set, validation set, and test set at an 8 to 1 to 1 ratio. The input data was shaped as 256 × 256 pixels RGB image, using an SGD optimizer, and the Nesterov momentum method was used to accelerate model convergence. The momentum, batch size, and learning rate were set to 0.9, 32, and 0.001, respectively. The OA was used as an evaluation index for model classification performance.

$$OA = TP/(TP + FP) \tag{5}$$

In equation (5), TP and FP represent the number of images correctly and inaccurately classified by the model on the test set, respectively.

3.3 Ablation experiment

To demonstrate the effects of this study, we conducted ablation experiments on three datasets: NWPU-RESISC45, AID, and UCM. Table 1 shows the performance of models with four different settings:

(1) Scheme 1: LRN was replaced with LN and Replaced the convolution kernel size of 11×11 with 7×7

(2) Scheme 2: The CBAM was embedded in the improved AlexNet.

(3) Scheme 3: The SEM was embedded in the improved AlexNet.

(4) Ours: The CBAM and SEM were embedded in the improved AlexNet.

Table 1. Comparing OA of different schemes on three remote sensing scene datasets.

Scheme	NWPU-RESISC45	AID	UCM
AlexNet	81.91%	83.12%	86.67%
Scheme 1	93.99%	92.84%	94.86%
Scheme 2	93.05%	93.50%	92.67%
Scheme 3	93.47%	93.24%	95.52%
Ours	96.29%	96.02%	96.57%

From Table 1, we can see that compared with the AlexNet, Scheme 1 has the highest OA of the three datasets, although Scheme 2 only has the highest OA on the AID dataset, however, the OA of Scheme 3 in the three datasets has been improved. The OA of our model on the three datasets was 14.38%, 12.09%, and 9.9% higher than AlexNet, respectively, therefore the proposed improvement method was remarkable.

4. ANALYSIS OF RESULTS

To verify the classification accuracy of the enhanced model in this study, The classification accuracy of NWPU-RESISC45, AID, and UCM datasets were compared and analyzed using four different models. as shown in Table 2. The OA of the improved model in this paper was 96.26%, 96.02% and 96.57% in the three datasets, and higher than the other three models.

Table 2. Comparison of OA of different models on three remote sensing scene datasets.

Model	NWPU-RESISC45	AID	UCM
Wu[16]	88.29%	93.61%	95.81%
Liu[17]	\	91.92%	93.52%
Liang[18]	92.90%	95.80%	\
This study	96.29%	96.02%	96.57%

5. CONCLUSION

In this study, we improved the AlexNet model. Firstly, we used the LN to replace LRN in AlexNet and changed the convolution kernel size of the first convolution layer to 7 × 7 to reduce excessive compression of spatial information during feature extraction. Secondly, the two attention modules of CBAM and SEM were embedded to make the model more focused on critical information in the feature extraction process. This study achieved the OA of 96.29%, 96.02%, and 96.57% on the NWPU-RESISC45, AID, and UCM remote sensing scene datasets, respectively. Compared with the unimproved AlexNet, the OA was significant improvement. Compared with the three excellent models above, the model in this study has a higher classification accuracy. In future research, this study will further improve the network structure, extract more critical information, and train the model with fewer samples to achieve an excellent classification effect.

Funding: This research was supported in part by research grants from the National Natural Science Foundation of China (32060320, 31860181, 32160368); Research Foundation for Basic Research of Yunnan Province (202101AT070039); "Ten Thousand Talents Program" Special Project for Young Top-notch Talents of Yunnan Province (YNWR-QNBJ-2020047); Joint Special Project for Agriculture of Yunnan Province, China (202101BD070001-066).

REFERENCES

[1] Wang, X. and Shen, S., “Multi-class remote sensing object recognition based on discriminative sparse representation,” Applied Optics, 55(6), 1381-1394(2016).
[2] Ning, C. and Liu, W. B., “Enhanced synthetic aperture radar automatic target recognition method based on novel feature,” Applied Optics, 55(31), 8893-8904(2016).
[3] Wang, X. and Xiong, X., “Integration of heterogeneous features for remote sensing scene classification,” Journal of Applied Remote Sensing, 12(1), 15-23(2018).
[4] Feng, Z. Y., “Research and application of image feature learning and classification methods based on deep learning,” South China University of Technology, 2, 1-137(2016).
[5] Xi, X. F. and Zhou, G. D., “A survey on learning for natural language processing,” Acta Automatica Sinica, 42(10), 1445-1465(2016).
[6] Krizhevsky, A. and Sutskever, I., “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, 25, 1097-1105(2012).
[7] Xiao, L. and Yan, Q., “Scene classification with improved AlexNet model,” Proc. of the IEEE Inter. Conf. on Intelligent Systems and Knowledge Engineering (ISKE), 1-6(2017).
[8] Han, X. and Zhong, Y., “Pre-trained Alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification,” Remote Sensing, 9(8), 848(2017).
[9] Fu, Q., Luo, W. L. and Lv, J. X., “Land utilization change detection of satellite remote sensing image based on AlexNet and support vector machine,” Laser & Optoelectronics Progress, 57(17), 282-289(2020).
[10] Zhang, G., Xu, W. and Zhao, W., “A multiscale attention network for remote sensing scene images classification,” Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 9530-9545(2021).
[11] Zhao, Z. and Li, J., “Remote sensing image scene classification based on an enhanced attention module,” Proc. of the IEEE Conf. on Geoscience and Remote Sensing Letters, 18(11), 1926-1930(2020).
[12] Alhichri, H. and Alswayed, A. S., “Classification of remote sensing images using EfficientNet-B3 CNN model with attention,” IEEE Access, 9, 14078-14094(2021).
[13] Ji, J. and Zhang, T., “Combining multilevel features for remote sensing image scene classification with attention model,” Proc. of the IEEE Conf. on Geoscience and Remote Sensing Letters, 17(9), 1647-1651(2019).
[14] Fan, R. and Wang, L., “Attention based residual network for high-resolution remote sensing imagery scene classification,” Proc. of the IEEE Conf. on International Geoscience and Remote Sensing, 1346-1349(2019).

[15] Chen, S. B. and Wei, Q. S., “Remote sensing scene classification via multi-branch local attention network,” Proc. of the IEEE Conf. on Transactions on Image Processing, 31, 99-109(2021).
[16] Wu, H., Zhao, S., Li, L., Lu, C. and Chen, W., “Self-attention network with joint loss for remote sensing image scene classification,” IEEE Access, 8, 210347-210359(2020).
[17] Liu, C. R. and Ning, Q., “Application of improved residual network in sensing image classification,” Science Technology and Engineering, 21(31), 13421-13429(2021).
[18] Liang, W. T. and Kang, Y., “Adaptive remote sensing scene classification based on complexity clustering,” Computer Engineering, 46(12), 256-261+269(2020).