Attention-based DenseNet network for multi-source remote sensing classification

Lishuo Zhang¹, Hong Lin¹ and Fanyang Zeng¹, ²

¹Guangzhou Urban Planning & Design Survey Research Institute, No.10, Jianshe Damalu, Guangzhou 510060, China
²Corresponding author’s e-mail: fanyangz@126.com

Abstract. Remote sensing data are abundant in spatial, image information and widely applied in earth observation covering military and civil fields. As the hot spot and critical point, scene classification has attracted increasing attention. Convolutional neural networks (CNNs) behave as representative technologies in scene classification. Nevertheless, many CNNs are lack of the ability to distinguish key information from redundant information. To overcome the problem and achieve efficient feature extractions from multi-source data, we integrate DenseNet and attention mechanism into a CNN-based network. Multi-scale convolution and three-dimensional convolution are applied to extract features; and then, to explore and obtain feature information in a deeper level, a Dense structure are utilized; Afterwards, the features extracted from previous procedures are weighed due to importance and merged through the convolutional attention model. The experiment validates that the presented method poses advantages of efficiently merging various feature by Dense structure and the convolutional attention model.

1. Introduction

In recent years, with the continuous development of remote sensing technology and the continuous improvement of sensor hardware and software performance, remote sensing can obtain Earth observation data with higher resolution and more sources, which provides a better way to classify remote sensing features. Rich data resources. However, as the research on the classification of ground objects becomes more and more in-depth, it is difficult to meet the requirements of refined classification using a single remote sensing data. How to fuse multiple remote sensing data to extract complementary information and remove redundant information to better.

The proposal of AlexNet opened a precedent in the application of convolutional neural networks [1]. Subsequent networks such as GoogleNet and VGG used smaller convolution kernels and increased depth, proving that convolutional neural networks have better performance in processing image problems [2]. In 2015, the residual network was proposed to solve the problem of gradient dispersion caused by too deep network layers, and provided a new idea for network training [3]. Unlike the traditional method of directly learning network parameters, ResNet connects input and output, so as to learn only the changing disturbance, which greatly improves the training speed and training accuracy Its basic idea is the same as ResNet, but it establishes a dense connection between all the front layers and the back layers. Realize feature reuse (feature reuse) through the connection of features on the channel. These features allow DenseNet to achieve better performance than ResNet with fewer parameters and computational costs [4].
Inspired by the significant progress made in natural image processing, remote sensing image scene classification model using CNNs [5, 6], DenseNet, [7] ResNet and attention mechanism [8-10].

The proposed attention model contains multiple layers of Dense structure and an attention module. Considering that the classification needs to consider not only the weight of the feature dimension but also the mutual influence between different pixels, this paper introduces the CBAM module [11]. First, the extracted spectral features and spatial features are analysed through the attention mechanism on the channel. The importance of features and height information is learned, and then the influence of other pixels on the target pixel is considered through the spatial attention mechanism, so as to complete the classification of the features according to the weighted result.

The rest of this paper is structured as follows. Section 2 shows our methodology. Experiment and analysis are conducted in Section 3. The conclusion is obtained in Section 4. References are given at the end of the paper.

2. Methodology
The main architecture of our model is shown in Figure 1.

The proposed method contains two parts: Feature Extraction module and Attention Module.

![Figure 1. The main architecture of our model.](image)

2.1. Feature Extraction
The spectral information of high-resolution data and LiDAR data is weak, but they have strong spatial characteristics. The commonly used 3x3 convolution kernel can only perceive the information of the surrounding 8 pixels, which limits the extracted features. Therefore, in order to fully extract the spatial information of the data, this paper introduces the multi-scale convolution method, and superimposes the results of the convolution extraction under the multi-scale to obtain richer spatial information.

Aiming at the relatively continuous characteristics of the spectral response of hyperspectral data, the spatial and spectral features of hyperspectral data are extracted simultaneously by means of three-dimensional convolution in this article, and the features between adjacent bands can be learned by sliding on the channel to extract Rich band information.

2.2. DenseNet
DenseNet is a convolutional neural network with dense connections. In this network, there is a direct connection between any two layers, that is, the input of each layer of the network is the union of the outputs of all the previous layers, and the feature map learned by this layer will also be directly passed to All subsequent layers are used as input.
2.3. **Convolutional Block Attention Module (CBAM)**

At present, the commonly used fusion methods for the extracted features in the neural network are addition and concatenation, both of which are direct processing of various features, but in the actual remote sensing image classification, different types of ground objects have different features. The response is different. For example, vegetation and other ground objects with obvious spectral characteristics can get a better classification effect according to the spectral characteristics. Therefore, spectral characteristics should be considered more when classifying, but materials like cement roads and roofs It is often more accurate to use spatial information to classify similar ground objects with a large difference in shape and texture. At this time, the spatial characteristics need to be enhanced. At the same time, the same convolution kernel of the convolutional neural network shares spatial weights, so it is often necessary to increase the number of convolution kernels to extract different features, which leads to the generation of redundant information. This is achieved through simple feature superposition fusion. The classification of remote sensing images has caused some impact. Therefore, an attention mechanism is introduced in this article to weight the fusion features, so that different features have different importance in the classification of different features. Considering that the classification needs to consider not only the weight of the feature dimension but also the mutual influence between different pixels, this paper introduces the CBAM module. First, the extracted spectral features and spatial features are analyzed through the attention mechanism on the channel. The importance of features and height information is learned, and then the influence of other pixels on the target pixel is considered through the spatial attention mechanism, so as to complete the classification of the features according to the weighted result.

### 3. Experiment and analysis

#### 3.1. Datasets

The data used in this experiment is the public data set grss_dfc_2018 [12] provided by the 2018 IEEE Fusion Competition, including high-resolution aerial images (0.05 m), multispectral LiDAR point cloud data (0.45 m), and hyperspectral data (1m). The reference sample data contains 20 urban LULC classes with a spatial resolution of 0.5m.

#### 3.2. Experimental settings

80% of all samples are selected for model training and 20% for testing. The learning rate is set to 0.0001, including a total of 800 epochs. Due to the large image size, it cannot be directly imported into the network for training. Therefore, the training sample is first divided. The size of the original training image is 4172×1202, which is divided during training. 20 sub-blocks are trained, and the calculated loss function value is the average value of all sub-blocks, optimized by Adam algorithm. All implementations are based on TensorFlow and NVIDIA RTX2060.

#### 3.3. Experimental results

We evaluated our model on grss_dfc_2018 Datasets. Figure 2 show the accuracy and loss curves of our attention model.
The feature extraction of high-resolution images, hyperspectral images and LiDAR are shown in Figure 3. High-resolution data and hyperspectral data have spatial and spectral information, so the feature extraction results are relatively similar, the spatial characteristics of the high-resolution data are more obvious. The hyperspectral images can better express highly reflective objects. The features extracted from LiDAR data represent the height information of the object. By fusing different features using the attention mechanism, the fused features are shown in Figure 3(g). The final classification result is shown in Figure 4.

**Figure 2.** Accuracy and loss curves.

The feature extraction of high-resolution images, hyperspectral images and LiDAR are shown in Figure 3. High-resolution data and hyperspectral data have spatial and spectral information, so the feature extraction results are relatively similar, the spatial characteristics of the high-resolution data are more obvious. The hyperspectral images can better express highly reflective objects. The features extracted from LiDAR data represent the height information of the object. By fusing different features using the attention mechanism, the fused features are shown in Figure 3(g). The final classification result is shown in Figure 4.

**Figure 3.** Feature Extraction (a: High resolution image; b: Hyperspectral image; c: LiDAR image (Canopy height model); d-f: Feature Extraction of (a)-(c); g: CBAM).
As we can see from Figure 2, the training accuracy continues to rise, and the loss value steadily declines. 800 epochs have converged, indicating the feasibility of the model. The overall classification accuracy is 78.62%. This just is the rough classification result. More fine classification result can be obtained by using Markov random field and so on.

4. Conclusion
This paper uses deep learning as the basic framework and uses feature-level fusion to integrate multi-source remote sensing data to complete the classification. Using different feature extraction methods for different data can get richer feature information. By introducing the attention model for feature fusion, a better fusion effect can be achieved, which provides a new idea for remote sensing image fusion. The sample difference between different classes is large, and the problem of sample balance needs to be dealt with. In the future, we will set different parameters to test this model.

References
[1] A. Krizhevsky, I. Sutskever and G. E. Hinton 2012 Imagenet classification with deep convolutional neural networks
[2] C. Szegedy, S. Ioffe and V. Vanhoucke 2016 Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning CoRR vol. abs/1602. 07261
[3] K. He, X. Zhang, S. Ren and J. Sun 2016 Deep Residual Learning for Image Recognition
[4] G. Huang, Z. Liu, V. D. M. Laurens and K. Q. Weinberger 2016 Densely Connected Convolutional Networks
[5] W. Han, R. Feng, L. Wang and L. Gao 2018 Adaptive Spatial-Scale-Aware Deep Convolutional Neural Network for High-Resolution Remote Sensing Imagery Scene Classification in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium pp. 4736-4739
[6] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao and B. Zhang 2018 Multisource Remote Sensing Data Classification Based on Convolutional Neural Network vol. 56 pp. 937-949
[7] W. Tong, W. Chen, W. Han, X. Li and L. Wang 2020 Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification vol. 13 pp. 4121-4132
[8] F. Li, R. Feng, W. Han and L. Wang 2020 An Augmentation Attention Mechanism for High-Spatial-Resolution Remote Sensing Image Scene Classification vol. 13 pp. 3862-3878
[9] R. Fan, L. Wang, R. Feng and Y. Zhu 2019 Attention based Residual Network for High-
Resolution Remote Sensing Imagery Scene Classification in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* pp. 1346-1349

[10] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma and L. Jiao 2021 Attention Consistent Network for Remote Sensing Scene Classification p. 1-1

[11] S. Woo, J. Park, J. Lee and I. S. Kweon 2018 CBAM: Convolutional Block Attention Module *Cham* pp. 3-19

[12] 2018 IEEE GRSS Data Fusion Contest. Online: http://www.grss-ieee.org/community/technical-committees/data-fusion (accessed on 12 August 2020)