A Hyperspectral Image Classification Method with CNN Based on attention-enhanced Spectral and Spatial Features

Yangming Zhang\textsuperscript{1,3,a}, Kun Yang\textsuperscript{2,3,*}, Lei Yuan\textsuperscript{2,3,b}

\textsuperscript{1} School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China
\textsuperscript{2} Faculty of Geography, Yunnan Normal University, Kunming 650500, China
\textsuperscript{3} Engineering Research Center of the Ministry of Education on Geography Information Technology of Western Resource Environment, Kunming 650500, China

\textsuperscript{a} Email: zym9529@163.com
\textsuperscript{b} Email: v_ict@163.com
\textsuperscript{*} Correspondence: Email: yangkun@ynnu.edu.cn

Abstract: In recent years, the convolutional neural network (CNN) has had a wide application in hyperspectral image (HSI) classification. HSI has many spectral and spatial features, which is well known that different spectral bands and spatial positions in the cubes have different discriminative abilities. Therefore, this paper proposes a classification method with CNN, which uses attention-enhanced spectral and spatial features (CNN-ASS). First, we use spectral and spatial subnetworks to extract spectral and spatial features. Then, we sum the weights of the classification results of the two subnetworks to get the final classification result. This paper conducts experiments on three typical hyperspectral image data sets, and the experiment results show the CNN-ASS has a competitive advantage compared with some advanced methods.

1. Introduction

1.1. HSI classification methods

In recent years, deep learning has made promising achievements in the machine learning field [1, 2], and it has been widely used for remote sensing image classification [3, 4]. CNN for depth representation based on spectral features, and its performance is better than SVM [5]. The typical CNN-based feature extraction or classification task approach is to learn ground object features directly from an original hyperspectral image [6]. If these methods only use spectral information, they often lead to noise classification results [7]. So we can use the characteristics of spatial neighbors in HSI spatial features to improve classification accuracy [8].

Recently, the attention mechanism has had a wide application in language modeling and computer vision tasks. [9-11]. Its success mainly depends on the reasonable assumption that human vision focuses only on particular parts of the whole visual space when and where needed [12]. So we propose an attention-aided spectral-spatial CNN model for hyperspectral image classification. It enhances the representation capacity of CNN by using an attention mechanism, making CNN focus on more discriminative spectral bands and spatial positions while suppressing unnecessary ones.
1.2. The main contributions of this paper

The main contributions of this paper can be summarized as follows:

1. We propose a two-branch network for HSI classification. The network can extract spectral and spatial features simultaneously to improve the utilization rate of hyperspectral images.

2. We add spectral and spatial attention modules to the spectral and spatial subnetworks, respectively, so that CNN can better extract spatial and spectral features.

2. Proposed Classification Method

2.1. Data preprocessing

In this paper, we use principal component analysis (PCA) to reduce the dimension of the original HSI data to remove spectral redundancy. Let \( M \in \mathbb{R}^{H \times W \times C} \) denotes the original HSI cube, and \( \tilde{M} \in \mathbb{R}^{H \times W \times B} \) represents the data cube after using the PCA method. The PCA method reduces spectral bands from \( C \) to \( B \) while maintaining the same width \( W \) and height \( H \).

2.2. Two-branch CNN structure

As shown in Figure 1, the CNN-ASS network consists of two subnetworks. The first subnetwork extracts spectral features and carries out convolution in spectral channels, and the designed CNN architecture consists of three convolution layers and one full connection layer. The second subnetwork extracts spatial features and carries out convolution in the spatial direction, and the designed CNN architecture also includes four convolution layers and one full connection layer. Each convolution layer of two subnetworks is followed by a corrected linear unit (RELU) layer.

We use 2D-CNN to extract spectral and spatial features in both subnetworks. In 2D-CNN, 2D convolution is performed at the convolutional layers to extract features from the local neighborhood on feature maps in the previous layer. Then an additive bias is applied, and the result is passed through an activation function. Formally, the value of unit at position \((x, y)\) in the \(j\)th feature map in the \(i\)th layer, denotes as \(v_{i,j}^{x,y}\) is given by Equation (1):

\[
v_{i,j}^{x,y} = \varphi(b_{i,j} + \sum_{m=0}^{P-1} \sum_{p=0}^{Q-1} w_{i,j,m}^{p,q} v_{(i-1)_{m}}^{(p)+x(q)+y})
\]

where \(\varphi\) is the active function, \(b_{i,j}\) is the bias parameter for the \(j\)th feature map of the \(i\)th layer, \(m\) indexes over the set of feature maps in the \((i-1)\)th layer connected to the current feature map, \(w_{i,j,m}^{p,q}\) is the value at the position \((p, q)\) of the kernel connected to the \(k\)th feature map, \(P_i\) and \(Q_i\) are the height and width of the kernel, respectively.
2.3. Add attention mechanism
This paper aims to better use spectral information among bands and semantic information in space. Two kinds of attention modules are designed to achieve this goal: the channel attention module and the spatial attention module. The spectral attention module focuses the channel subnet on the more discriminating channels and suppresses the unnecessary channels. Similarly, the spatial attention module can make the spatial subnetwork pay more attention to the semantic location and extract spatial information more effectively.

In general, we use transformation $F_r$ mapping an input $X \in \mathbb{R}^{H \times W \times C}$ to feature maps $U \in \mathbb{R}^{H \times W \times C}$. In the Equation (2), we take $F_r$ to be a convolutional operator and use $V = [v_1, v_2, ..., v_C]$ to denote the learned set of filter kernels, where $v_c$ refers to the parameters of the $c$th filter. Then we can denote the outputs as $U = [u_1, u_2, ..., u_C]$.

$$u_c = v_c * X = \sum_{s=1}^{V} v_c^s * x^s$$

here $*$ denotes convolution, $v_c^s$ is the $s$th channel 2D spatial kernel representing a single channel of $v_c$ that acts on the corresponding channel of $X$. However, since the sum is taken for the convolution result of each channel, the channel characteristic relation is mixed with the spatial relation learned from the convolution kernel. We need to pull out such confounding so that the model can directly learn the channel characteristic relationship and space characteristic relationship.

2.3.1. Spectral attention module
As shown in Figure 2, we use the spectral attention module to learn the relationship among spectral bands better. It can make the model automatically learn the importance of different spectral features. First of all, we extract the global features of each channel, adopting global average pooling as the following Equation (3) to achieve:

$$z_c = F_{aq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), z \in \mathbb{R}^c$$

Then, we use convolution operations to extract spectral features. After we have the global description characteristics, we begin to learn the relationship among the various channels. Here we use the gating mechanism of sigmoid function as the following Equation (4):

$$s = F_{as}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \Re L(U(W_1)z))$$

Figure 2. Spectral attention module

2.3.2. Spatial attention module
As shown in Figure 3, spatial attention focuses on the most informative part so that the model can better learn the spatial characteristics of the object. Similar to channel attention, first, we use the channel average pooling operation to reduce the channel dimension, as the following Equation (5):

$$p_n' = F_{aq}(p_n) = \frac{1}{C} \sum_{c=1}^{C} u_c^n, p'_n \in \mathbb{R}^n$$

where $u_c^n$ represents $n$th pixel in the $n$th layer feature map. Then we adopt convolution operations to extract
spatial features. At last, we utilize the gating mechanism of the sigmoid to learn the spatial semantic relationship among pixels, as the following Equation (6):
\[ s = F_{ex}(p,W) = \sigma(g(p,W)) = \sigma(W_2 \text{ReLU}(W_1 p)) \]  

(6)

![Spatial attention module](image)

Figure 3. Spatial attention module

3. Experiment and Results

3.1. Experimental environment
For the CNN-ASS model proposed in this paper, all programs are implemented in Python language, the network is constructed by using the TensorFlow 1.14.0 deep learning framework, and the calculation is deployed on the GPU RTX2060.

3.2. Experimental data
The Indian Pines data set contains 16 different ground object classes acquired by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor in 1992. The Pavia University data set contains 9 ground object classes acquired by the ROSIS sensor on the Pavia University campus in Italy. The Salinas data set contains 16 ground object classes, and it is also collected by AVIRIS sensor in the Salinas Valley, California, USA. The number of training samples for Indian Pines, Pavia University, and Salinas is 10%, 3%, and 3% of the total samples, respectively.

3.3. Experimental result
In this paper, we use the overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa) evaluation measures to judge the HSI classification performance. As shown in Table 1, the results of the proposed CNN-ASS model are compared with the most widely used supervised methods, such as SVM [13], 2D-CNN [14], 3D-CNN [15], SSUN [16], and HybridCN [17], CNN-ASS has achieved better classification results on the three public data sets. Figure 4 to Figure 6 show the classification effect diagrams on three data sets by using different methods. Table 2 shows the execution time of different methods on the three data sets. The training time of the CNN-ASS network is less than that of 3D-CNN and HybridCN, and the test time is only less than that of 3D-CNN.

Table 1. Average classification results (%) of 10 experiments performed on the three data sets by using different methods

| Data           | Performance | SVM  | 2D-CNN  | 3D-CNN  | SSUN  | HybridCN | CNN-ASS |
|---------------|-------------|------|---------|---------|-------|----------|---------|
| Indian Pines  | OA          | 62.55| 83.11.  | 91.86   | 97.11 | 98.13    | 98.79   |
|               | AA          | 54.02| 78.22   | 86.93   | 94.52 | 98.31    | 98.93   |
|               | Kappa       | 55.80| 79.59   | 90.71   | 96.70 | 97.68    | 98.27   |
| Pavia University | OA       | 79.97| 90.72   | 96.84   | 96.56 | 98.21    | 99.28   |
|               | AA          | 66.53| 88.04   | 94.85   | 95.93 | 97.67    | 99.02   |
|               | Kappa       | 72.07| 87.64   | 95.83   | 95.44 | 98.14    | 99.04   |
| Salinas       | OA          | 83.39| 90.28   | 92.59   | 96.67 | 98.12    | 99.58   |
|               | AA          | 87.53| 94.29   | 94.85   | 98.16 | 98.03    | 99.54   |
|               | Kappa       | 81.31| 89.17   | 91.73   | 96.29 | 98.35    | 99.41   |
(a) (b) (c) (d) (e) (f) (g)
Figure 4. Classification maps of the Indian Pines. (a) Ground truth. (b) SVM. (c) 2D-CNN. (d) 3D-CNN. (e) SSUN. (f) HybridCN. (g) CNN-ASS.

(a) (b) (c) (d) (e) (f) (g)
Figure 5. Classification maps of the Pavia University. (a) Ground truth. (b) SVM. (c) 2D-CNN. (d) 3D-CNN. (e) SSUN. (f) HybridCN. (g) CNN-ASS.

(a) (b) (c) (d) (e) (f) (g)
Figure 6. Classification maps of the Salinas. (a) Ground truth. (b) SVM. (c) 2D-CNN. (d) 3D-CNN. (e) SSUN. (f) HybridCN. (g) CNN-ASS.

| Data         | Experiment | SVM   | 2D-CNN | 3D-CNN | SSUN  | HybridCN | CNN-ASS |
|--------------|------------|-------|--------|--------|--------|----------|---------|
| Pavia University | train     | 0.06  | 21.66  | 324.57 | 46.83  | 174.49   | 119.66  |
|               | test       | 1.28  | 1.17   | 2.84   | 0.76   | 0.64     | 2.43    |
| Indian Pines  | train     | 0.05  | 38.89  | 275.18 | 56.03  | 155.69   | 138.31  |
|               | test       | 2.91  | 2.05   | 6.86   | 2.99   | 2.22     | 5.09    |
| Salinas      | train     | 0.05  | 53.73  | 520.61 | 69.32  | 170.12   | 140.54  |
|               | test       | 7.48  | 1.25   | 15.62  | 3.80   | 2.71     | 6.32    |

Table 2. Average operation time of different methods on three data sets (unit: second)

4. Conclusion
In this paper, a method named the CNN-ASS is proposed for HSI classification. First, to better use the hyperspectral image's spectral and spatial information, we propose a two-branch network to extract spectral and spatial features, respectively. Then we add spectral and spatial attentions to the spectral and spatial subnetworks, respectively. It enables the model to extract useful features and suppress some redundant features, making the model more robust. The experimental results show that the CNN-ASS
can provide higher classification accuracy. However, the time performance of the CNN-ASS model is ordinary, so the model needs to be further improved in future work.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (No. 41761081) and in part by the Science Research Foundation of Yunnan Education Bureau 2020 (No. 2020J0098).

References

[1] A Krizhevsky, I. Sutskever, and G. E. Hinton. (2012) “ImageNet classification with deep convolutional neural networks,” Advances in neural information processing systems, pp. 1097–1105.
[2] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, (2015) “Recurrent network models for human dynamics,” IEEE International Conference on Computer Vision, pp. 4346–4354.
[3] X. Ma, J. Geng, and H. Wang, (2015) “Hyperspectral image classification via contextual deep learning,” EURASIP Journal on Image and Video Processing.
[4] Q. Zou, L. Ni, T. Zhang et al., (2015) “Deep Learning Based Feature Selection for Remote Sensing Scene Classification,” IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 11, pp. 1-5.
[5] H. Wei, H. Yangyu, W. Li et al., (2015) “Deep Convolutional Neural Networks for Hyperspectral Image Classification,” Journal of Sensors, pp. 1-12.
[6] W. Li, G. Wu, F. Zhang et al., (2017) “Hyperspectral Image Classification Using Deep Pixel-Pair Features,” IEEE Transactions on Geoscience and Remote Sensing.
[7] S. A. Ahmadi, N. Mehrshad, and S. M. Arghavan, (2020) “Spectral-spatial classification method for hyperspectral images using stacked sparse autoencoder suitable in limited labelled samples situation.” Geocarto International.
[8] W. Li and Q. Du, (2014) “Joint within-class collaborative representation for hyperspectral image classification,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 6, pp. 2200–2208.
[9] V. Mnih et al., (2014) “Recurrent models of visual attention,” Advances in Neural Information Processing Systems, pp. 2204–2212.
[10] K. Xu et al., (2015) “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., pp. 2048–2057.
[11] J. Hu, L. Shen, and G. Sun, (2018) “Squeeze-and-excitation networks,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141.
[12] L. Chen et al., (2017) “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5659–5667.
[13] F. Melgani, and L. Bruzzone, (2004) “Classification of hyperspectral remote sensing images with support vector machines,” IEEE Transactions on Geoscience and Remote Sensing, vol. 42, no. 8, pp. 1778-1790.
[14] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, (2015) “Deep supervised learning for hyperspectral data classification through convolutional neural networks,” IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4959–4962.
[15] Zilong, Zhong, Jonathan, Li, Zhiming, Luo, Michael, and Chapman, (2017) “Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework,” IEEE Transactions on Geoscience and Remote Sensing.
[16] Y. Xu, L. Zhang, B. Du et al.,(2018) “Spectral–Spatial Unified Networks for Hyperspectral Image Classification,” IEEE Transactions on Geoscience and Remote Sensing, pp. 1-17.
[17] S. K. Roy, G. Krishna, S. R. Dubey et al., (2020) “HybridSN: Exploring 3D-2D CNN Feature Hierarchy for Hyperspectral Image Classification,” IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 277-281.