SPECTRAL-SPATIAL ATTENTION NETWORKS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Erting Pan, Yong Ma, Xiaoguang Mei, Xiaobing Dai, Fan Fan, and Jiayi Ma

Electronic Information School, Wuhan University, Wuhan, 430072, China

ABSTRACT

Deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been successfully used to extract deep features for many hyperspectral tasks. In this study, we propose an spectral-spatial attention network for hyperspectral image classification. In our method, RNN with attention can learn interspectral correlations within a continuous spectrum, CNN with attention is designed to focus on similar features between neighbor pixels in spatial dimensions. Experimental results demonstrate that our method can fully utilize spectral and spatial information to obtain competitive performance.

***Index Terms*—** Hyperspectal image classification, attention mechanism, RNN, CNN

1. INTRODUCTION

Hyperspectral images captured from land surface-observing aircraft and satellites have become increasingly important in environmental monitoring, urban planning, mining, defense, and agriculture due to their rich spectral information [1, 2]. Hyperspectral imaging (HSI), also known as imaging spectroscopy, captures the electromagnetic energy that is reflected or emitted from the same area over hundreds of narrow, contiguous spectral bands from visible to middle infrared wavelength ranges [3, 4]. Each pixel in a hyperspectral image is composed of a vector of elements that measures spectral information as a function of wavelength, which is known as the spectrum. Each spectral band represents a gray-scale image, and all images make up a 3D hyperspectral cube, which causes a small patch to become a large data cube. A hyperspectral image can be construed as a 3D data structure with two spatial axes that carry information on the location of objects and one spectral axis that carry information on the objects' chemical composition.

Hyperspectral image classification, which assigns every pixel vector to a certain set of classes, is one of the major tasks in the analysis of hyperspectral images; it has received much attention from researchers. Numerous traditional methods, such as support vector machine (SVM) [5] and k-nearest neighbor (KNN) [4], have been proposed. However, these approaches disregard the correlations among pixels in spatial axes and cause a waste of spatial information. Thus, spectral-spatial-based methods improve classification performance because they incorporate additional spatial information from a hyperspectral image. For example, Li *et al*. [6] constructed a family of generalized composite kernels by utilizing spectral and spatial information from HSI data.

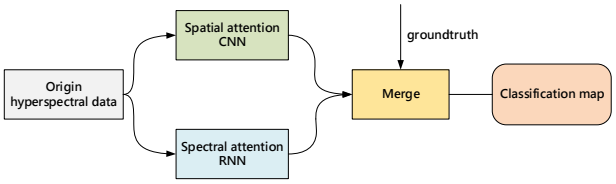


Fig. 1: Our framework for hyperspectral image classification.

Deep learning algorithms have become key tools in modern hyperspectral image analysis due to their outstanding predictive power; they can extract more discriminative features and achieve good performance than traditional shallow classifiers [7, 2]. Deep models, such as networks with 1D [8, 9], 2D, and 3D [10] convolutional layers, have been proposed for hyperspectral data analysis.

Methods with a 1D network use spectra as the input and learn features that capture spectral information only. Mou *et al*. [8] proposed the use of recurrent neural networks (RNNs) to model pixel spectra in a hyperspectral image as 1D sequences for classification, and they found that the modified gated recurrent unit (GRU) outperforms traditional approaches and the baseline convolutional neural network (CNN). Given that spatial information has been proven useful in improving the interpretation of HSI classification results, the study of classification models based on deep spectral-spatial features has been promoted. For example, SSUN [11] combined a spectral dimensional band grouping-based long short-term memory (LSTM) model with 2D CNN for spatial features and integrated the spectral finite element (FE), spatial FE, and classifier training into a unified neural network. The result showed that full use of spectral and spatial information can considerably improve accuracy.

With the same purpose, we designed a spectral-spatial

This work was supported by China Postdoctoral Science Foundation under Grant no. 2017M612504.

network with an attention mechanism. The contribution of this work can be summarized as follows: (1) We designed a joint network with a spectral attention bi-directional RNN branch and a spatial attention CNN branch to extract spectral-spatial features for hyperspectral image classification. An attention mechanism was used to emphasize meaningful features along the two branches, as shown in Fig. 1. Our goal was to increase representation power by using the attention mechanism, namely, enhance the correlations between adjacent spectral dimensions while focusing on important features, and suppressing unnecessary ones. (2) A bi-directional RNN with an attention mechanism was designed for spectral information. For each pixel, a spectral vector was divided into a set of single-ordered data and fed to a GRU one by one. Additional attention weights strengthened the spectral correlation between spectrum channels. (3) For spatial axes, we added attention to 2D CNN and trained this model on the image patch around the pixel. Compared with the average consideration of each image region, the attention parameter assigns a greater weight to the key parts to make the model pay more attention.

2. METHODOLOGY

Two sections play crucial roles in our methodology; they are a bidirectional RNN-based spectral attention feature learner and a CNN-based spatial attention feature learner.

The attention mechanism, which become a vital part in human perception, is based on a reasonable assumption that human vision does not process an entire image at once and only focuses on selective parts of the entire visual space [12, 13]. Several attempts have been exerted to incorporate attention processing into visual tasks, including hyperspectral image classification.

In our work, attention is of much concern. For spectral classification, considering that each pixel can be represented as a continuous spectral curve that contains rich spectrum characters, we can focus on the inter-band relationship of features by attention. In spatial dimensions, we regard spatial features as complementary to spectral ones; this branch improves the representation of interests and focus on the inter-spatial relationships of features by exploiting spatial attention to CNN. Then, we concatenate two branches and feed them to the fully connected layers to learn high-level joint spectral-spatial features and acquire a prediction class after a softmax layer.

2.1. Attention with RNN for Spectral Classification

RNNs are popular architectures for modeling various sequential problems. They contain feedback loops that allow the current output to depend on the current input and the previous inputs; therefore, they have a strong capability to capture contextual information within a sequence. By considering all

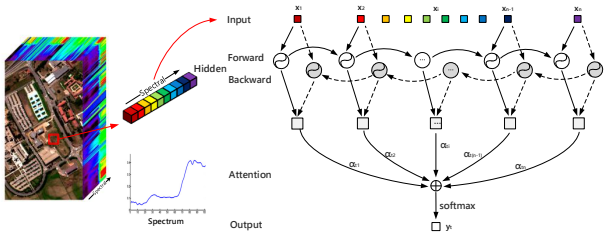


Fig. 2: Bi-RNN model with attention mechanism for Spectral Classification.

spectra of a hyperspectral pixel as a sequence, we develop a bi-RNN model, as illustrated in Fig. 2.

GRU is introduced to learn long-term dependencies and alleviate the vanishing gradient problem, and it has fewer parameters than LSTM. In our RNN model, we use bi-RNN as an encoder. Its input is a spectral vector of one hyperspectral vector x, $x = (x_1, x_2, \ldots, x_n)$, and the bidirectional hidden vector is calculated as

$$\overrightarrow{h_t} = f(\overrightarrow{W}x_t + \overrightarrow{V}h_{t-1} + \overrightarrow{b}), \quad (1)$$

$$\overleftarrow{h_t} = f(\overleftarrow{W}x_t + \overleftarrow{V}h_{t+1} + \overleftarrow{b}), \quad (2)$$

$$\overrightarrow{g_t} = concat[\overrightarrow{h_t}, \overleftarrow{h_t}], \quad (3)$$

where t ranges from the first spectral band 1 to the last n one, the coefficient matrices $\overleftarrow{W}$, $\overrightarrow{W}$ are from the input at the present step, $\overrightarrow{V}$ is from the hidden state h_{t-1} at the previous step, $\overleftarrow{V}$ is from h_{t+1} at the succeeding step, f is the nonlinear activation of the hidden layer, and g_t is the memory of the input as the output of this encoder.

Compared with the traditional RNN model that treats the input in the same manner, we add an attention layer to decode different spectral information to learn many characters. Our attention layer can be defined as follows:

$$e_{it} = tanh(W_i g_t + U_i h_i + b_i), \quad (4)$$

$$\alpha_{it} = softmax(W_i{}' e_{it} + b_i{}'), \quad (5)$$

$$y_t = U[g_t, \alpha], \quad (6)$$

where $W_i, U_i, W_i{}'$ are transformation matrices and $b_i, b_i{}'$ are bias terms. Output y_t is the predicted label of pixel x.

2.2. Attention with CNN for Spatial Classification

Our CNN model aims to extract robust spatial features. The dimensions of a hyperspectral image are reduced to a low-dimensional subspace via principal component analysis (PCA), which can reduce the dimensionality of a dataset with interrelated variables while retaining as much of the variation in the dataset as possible. The tight relationships between a target pixel and its neighbor regions are considered, and a small patch is created for every target pixel. With the addition

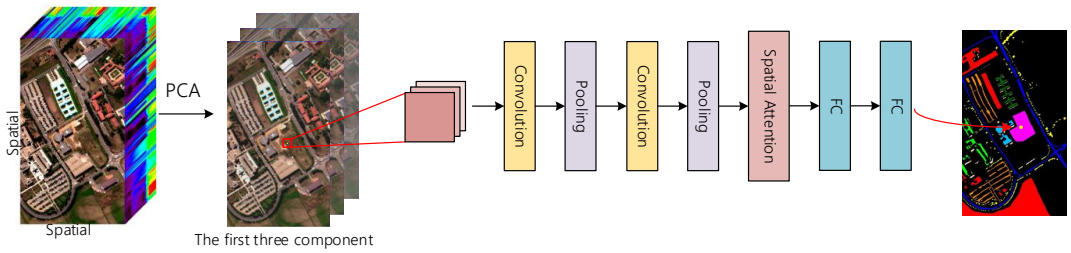


Fig. 3: CNN model with attention mechanism for Spatial Classification.

of the attention mechanism, our CNN model can focus on the primary meaningful features and ignore the rest.

The architecture is shown in Fig. 3. After PCA, for instance, the first three components of the Pavia University dataset are reserved because they have almost 99.3% information. Around each pixel, we create a patch with a size of $k * k * 3$ as a neighbor region. For each patch, the convolutional layer uses a sliding window as a kernel to move across, and it can locate similar features in this patch by calculating the point-to-point inner product. The pooling layer selects values to reduce the feature map dimensions. The kernels of the convolutional layers are $5 * 5$, and the strides of the max pooling layers are 2. As for the spatial attention layer, instead of considering each region equally, it pays attention to feature-related regions. With the previous feature map, we generate the attention distributions α to a new feature map. The fully connected layer FC owns $1,024$ units, and the last FC's unit number is equal to the classes. We obtain the final result by using a softmax function.

In our method, the last step is concatenating the two branches then co-training them, as shown in Fig. 1. If we only use the spectral RNN model, the result will be uneven due to the lack of spatial information; if we use the single spatial CNN model, an unlabeled area may be incorrectly labeled due to neighbor information. The merge layer fuses and balances the spatial and spectral information, and its result has the largest diversity in class probability estimation.

3. EXPERIMENT RESULTS

To evaluate our method, we train and test it on two public hyperspectral image classification datasets, namely, the Pavia University dataset and the Pavia Center dataset. Nine land cover classes of urban areas are contained in them. The datasets are split into training, validation, and test sets. To overcome the class imbalance problem, we randomly select 100 samples of each annotated class for training and 100 samples for validation instead of splitting them by an average percentage from each class.

We compare our method with traditional advanced machine-learning methods, such as KNN, linear SVM with radial basis function kernel, CNN, RNN, RNN with attention (ARNN), and CNN with attention (ACNN). For a fair comparison, we utilize the same training and testing datasets for all methods, and all algorithms are executed five times; the average results are reported to reduce random selection effects. Overall accuracy (OA), average accuracy, and the kappa coefficient k are used as the evaluation measurements for the compared methods. The experimental results of the Pavia University dataset are shown in Table RNum1, and the results of the Pavia Center dataset are presented in Table RNum2. The classification results of both datasets show that our proposed method, SSAN, exhibits the best performance among all compared methods in all scenarios.

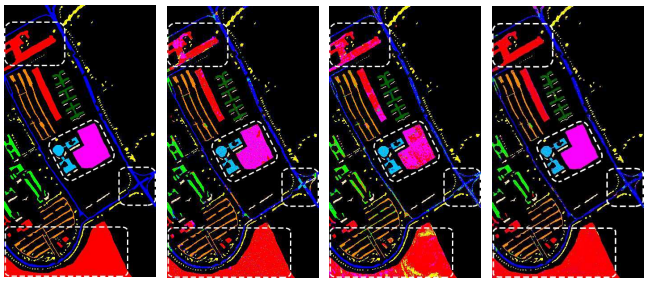
Table 1: Classification performance of different methods for the Pavia University dataset. Bold indicates the best result.

Label	KNN	SVM	RNN	CNN	ARNN	ACNN	SSAN
OA	84.48	84.43	91.2	89.20	96.54	92.61	**99.54**
AA	84.88	88.59	88.6	93.20	86.52	97.51	**98.41**
Kappa	83.0	79.94	89.3	85.91	90.90	82.01	**99.12**

Table 2: Classification performance of different methods for the Pavia Center dataset. Bold indicates the best result.

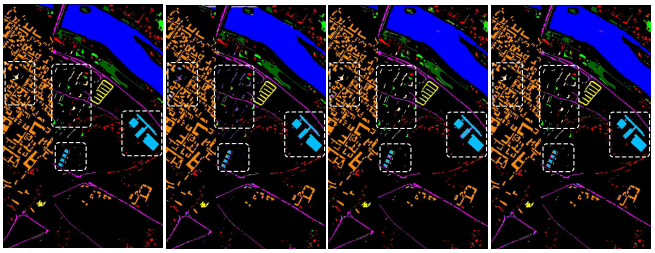
Label	KNN	SVM	RNN	CNN	ARNN	ACNN	SSAN
OA	92.5	93.05	92.3	86.20	99.47	96.38	**99.64**
AA	92.5	85.89	89.5	91.20	91.31	93.37	**98.06**
Kappa	91.6	90.18	91.4	68.91	98.41	94.83	**98.92**

The results indicate that the proposed method with the attention mechanism in two branches is effective in hyperspectral image classification. The traditional methods, such as SVM and KNN, demonstrate poor performance. Deep learning methods, such as CNN and RNN, are effective because of their discriminative features. A comparison of RNN and ARNN or CNN and ACNN indicates that the attention mechanism plays an important role in our method. Within the attention weights, CNN focuses on similar features between neighbor pixels, and RNN learns interspectral correlations. Better than a single CNN or RNN network, which only takes spatial information or spectral curve features, CNN appears to be more homogeneous and smoother than RNN, but RNN performs better in terms of OA. Our fusion network combines spatial and spectral dimensions and acquires well-balanced results. We show the classification maps for our proposed



(a)groundtruth (b) ACNN (c) ARNN (d) SSAN

Fig. 4: Visual results on the Pavia University dataset



(a)groundtruth (b) ACNN (c) ARNN (d) SSAN

Fig. 5: Visual results on the Pavia Center dataset

method in Figs. 4 and 5.

4. CONCLUSION

In this study, a two-branch co-training method is proposed to extract spectral-spatial features based on ARNN and ACNN for hyperspectral image classification. By adding attention weights to CNN and RNN, we can learn numerous interspectral correlations in the continuous spectrum domain and focus on similar spatial features between neighbor pixels in spatial dimensions. Analysis of experimental results on two datasets shows that our method not only performs better than the other methods but also extracts more homogeneous discriminative feature representations by combining ACNN and ARNN. We will generalize our method for other remote sensing applications, such as unmixing and change detection, in the future.

5. REFERENCES

[1] Fan Fan, Yong Ma, Chang Li, Xiaoguang Mei, Jun Huang, and Jiayi Ma, "Hyperspectral image denoising with superpixel segmentation and low-rank representation," *Information Sciences*, vol. 397, pp. 48–68, 2017.

[2] Liangpei Zhang, Lefei Zhang, and Bo Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[3] ME Paoletti, JM Haut, J Plaza, and A Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 120–147, 2018.

[4] Jon Atli Benediktsson and Pedram Ghamisi, *Spectral-spatial classification of hyperspectral remote sensing images*, Artech House, 2015.

[5] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[6] Jun Li, Prashanth Reddy Marpu, Antonio Plaza, Jose M Bioucas-Dias, and Jon Atli Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4816–4829, 2013.

[7] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, 2015.

[8] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[9] Hao Wu and Saurabh Prasad, "Convolutional recurrent neural networks forhyperspectral data classification," *Remote Sensing*, vol. 9, no. 3, pp. 298, 2017.

[10] Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[11] Yonghao Xu, Liangpei Zhang, Bo Du, and Fan Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5893–5909, 2018.

[12] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.

[13] Rudong Xu, Yiting Tao, Zhongyuan Lu, and Yanfei Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sens.*, vol. 10, no. 10, pp. 1602, 2018.