JOINT MULTILAYER SPATIAL-SPECTRAL CLASSIFICATION OF HYPERSPECTRAL IMAGES BASED ON CNN AND CONVLSTM

Jie Feng, Xiande Wu, Jiantong Chen, Xiangrong Zhang, Xu Tang, Di Li

Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China

ABSTRACT

In this paper, a novel method based on convolutional long short-term and convolutional nerual network (MCNN-ConvLSTM) is proposed for hyperspectral image (HSI) classification. Firstly, due to powerful hierarchical feature extraction ability, CNN is devised to extract spatial features of HSIs. Then, shallow, middle and deep spatial features of CNN are used as the input of several ConvLSTMs. ConvLSTMs are devised to extract different layers of joint spatial-spectral features due to the ability of global spectral feature extraction. Finally, multilayer spatial-spectral features are fused to achieve an end-to-end classification, which learn complementary information among the shallow layers with basic information and the deep layers with abstract information. The experimental results demonstrate that the proposed algorithm can yield competitive classification performance compared with existing methods.

Index Terms— feature extraction (FE), convolutional neural network (CNN), hyperspectral image (HSI) classification, convolutional long short-term memory (convLSTM), multilayer spatial-spectral information.

1. INTRODUCTION

Hyperspectral imagery (HSI) contains hundreds of narrow and contiguous spectral bands, with wavelengths spanning the visible to infrared spectrum [1]. Due to rich spectral information, HSI has been widely applied in many fields, such as environmental management, agriculture and mineralogy. The task of pixel-wise classification in HSIs is of great significant in these applications.

Recently, deep learning-based methods have been developed to deal with HSI classification. In [2], stacked autoencoder (SAE) learns intrinsic representations of the data in an unsupervised way. Deep belief network was used to extract spectral-spatial features along with logistic regression classification [3]. For these two methods, image patches are flattened into vectors as the input, which results in the loss of spatial information. In order to alleviate this problem, convolutional neual network (CNN) was introduced into HSI classification task [4]. In order to make full use of spatial and spectral information, some joint spectral-spatial methods based on CNN were proposed. In [5], a dual-channel CNN (DC-CNN) is constructed to extract spectral and spatial information respectively, then the spectral and spatial features are concatenated for classification. Chen et al. presented a 3-dimensional CNN (3DCNN) model, which extracts spatial-spectral features simultaneously [6]. However, the size of convolutinal kenels in DC-CNN and 3DCNN is fixed, which only takes local information into consideration in spectral dimension [7].

Recently, some advances in recurrent neural networks (RNNs) and long short-term (FC-LSTM) models provide some useful insights on modeling sequential data [8][9]. Considering the spectral continuity of hyperspectral data, HSIs are regarded as sequential data. In [10], RNN is used to extract contextually spectral information. To extract joint spatial-spectral features, 1DCNN and RNN (CRNN) are combined [7]. In CRNN, features extracted by the last layer of 1DCNN are used as the input of RNN. However, CRNN only extracts spectral features in the deep layers. It ignores complementary information among different layers.

In this paper, a novel joint multilayer spatial-spectral classification based on CNN and convolutional long short term memory networks (ConvLSTM) is designed for HSI classification. It is abbreviated as MCNN-ConvLSTM. Specifically, CNN is utilized to extract features from original image patches. Because full-connected LSTM cannot capture spatial information, it is replaced by ConvLSTM. The feature maps extracted by CNN are input into ConvLSTM and ConvLSTM is applied to model the dependencies among these feature maps. To make full use of complementary information among different layers, several ConvLSTM units are applied to fuse both shallow and deep features. Compared with DC-CNN and 3DCNN, the proposed method takes global spectral information into consideration. In constrast

This work was supported by the National Natural Science Foundation of China under Grant 61871306, by the National Natural Science Foundation of Shaanxi Province under Grant 2019JM-194, by the Fundamental Research Funds for the Central Universities under Grant JBX181707, by the Joint Fund of Equipment Research of Ministry of Education under Grant 6141A020333, by the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences, under Grant LSIT201803D.

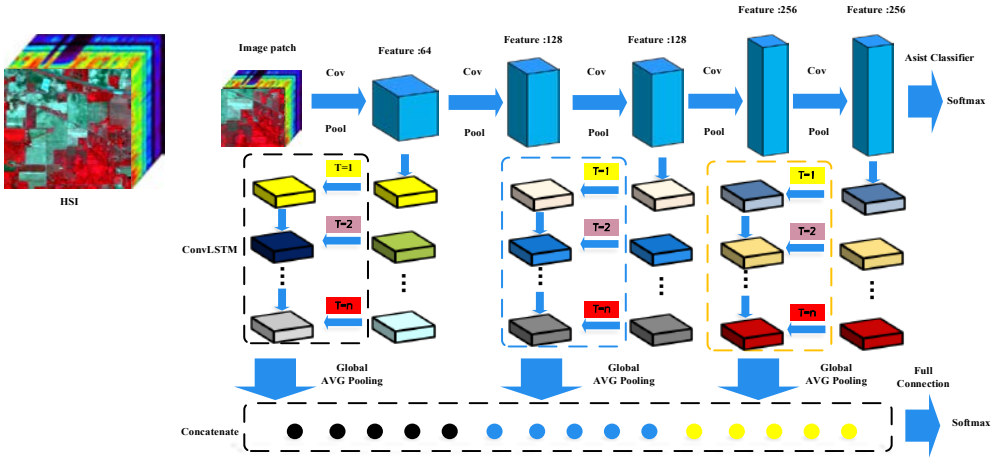

Fig. 1 The flowchart of the proposed MCNN-ConvLSTM method

with CRNN, the proposed method uses both deep and shallow features.

2. THE PROPOSED CLASSIFICATION METHOD

Fig.1 shows the flowchart of the proposed MCNN-ConvLSTM method. MCNN-ConvLSTM is introduced from three aspects.: ConvLSTM in MCNN-ConvLSTM, the classification of the network and the structure of the network.

2.1 ConvLSTM in MCNN-ConvLSTM

The CNN is constructed by stacking several convolution layers and pooling layers. In the convolution layer, a spatial window is chosen whose height and width are P, Q with its center at point (x, y). A neuron v_{ij}^{xy} of the j th feature map in the i th layer is calculated as follow:

$$v_{ij}^{xy} = g\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \quad (1)$$

$$g(x) = ReLu(x) = max(x, 0)$$

where m represents the feature map in the $i-1$ th layer connected to the current feature map, w_{ij}^{xy} indicates the weights connected to the m th feature map of position (p, q), b_{ij} is a bias and $g(\cdot)$ is activation function.

In CNNs, deep architecture potentially extracts hierarchical features layer by layer and the features can be divided into shallow and deep features. The shallow features contain more detailed and boundary information, while deep features consist of more semantic information. The features extracted by CNN only contain spatial information. In order to make full use of spectral information, CovLSTM is applied to extract joint spatial-spectral information, which is benefits for pixel-wise classification.

The outputs of CNN are considered as the input of ConvLSTM. Given feature maps, ConvLSTM is adopted to extract joint spatial-spectral information. The feature maps obtained by CNN are divided into several blocks in order and one block represents a time node. Every block is a input of ConvLSTM at a time step. For example, if the amount of feature maps is 100 and the number of time steps is 20, 100 feature maps are divided into 5 blocks. For a moment T, 20 feature maps are fed into ConvLSTM. The involved computation is given as follows:

$$\begin{aligned}
F_{ij}^k &= \sigma\left(W_{hf} * h_{ij}^{k-1} + W_{xf} * v_{ij}^k + b_f\right) \\
I_{ij}^k &= \sigma\left(W_{hi} * h_{ij}^{k-1} + W_{xi} * v_{ij}^k + b_i\right) \\
C_{ij}^k &= F_{ij}^k \circ C_{ij}^{k-1} + I_{ij}^k \circ tanh\left(W_{hc} * h_{ij}^{k-1} + W_{xc} * v_{ij}^k + b_c\right) \\
O_{ij}^k &= \sigma\left(W_{xo} * h_{ij}^{k-1} + W_{xo} * v_{ij}^k + b_o\right) \\
h_{ij}^k &= O_{ij}^k \circ \tanh(C_{ij}^k) \qquad (2)
\end{aligned}$$

Where v_{ij}^k is the k th feature map, and I , F ,O indicate the key components of ConvLSTM, input gate, forget gate and output gate, respectively. H is the hidden state. C presents the output of cell. W is the weight of ConvLSTM and $*$ represents the convolutional operator. .

The convolutional operator in (2) is applied to capture the dependencies among spectral bands. The output of the current moment is calculated by the hidden state H of the previous moment and the input of the current moment. In other words, the spectral information of each moment is used for joint spatial-spectral information of the next moment. In general , while spectral features are obtained by LSTM model, the spatial features are also futher utilized because of the convolutional operator.

The input gate, forget gate and output gate are the key components of ConvLSTM. Due to these components, the ability of long data sequential modeling of ConvLSTM is guaranteed. In the proposed method, ConvLSTM captures global dependencies of joint spatial-spectral information among all the feature maps.

2.2 MCNN-ConvLSTM Classification

In MCNN-ConvLSTM, shallow spatial features with detailed information and deep spatial features with semantic

TABLE II
CLASSIFICATION RESULTS OF RBF-SVM, DC-CNN, SSUN, CRNN, MCNN-CONVLSTM ON THE INDIAN PINES DATASET

Class	RBF-SVM[11]	DC-CNN[12]	SSUN[13]	CRNN[7]	MCNN-CONVLSTM
1	6.1±11.2	86.82±7.3	98.64± 3.2	72.27±15.5	96.82±3.6
2	72.9±3.6	89.06±2.4	91.73± 2.0	81.49±7.2	94.38±1.9
3	58.0±3.6	79.14±5.4	91.40± 10.9	84.80±9.2	91.85±2.8
4	39.0±15.0	88.71±2.9	94.22±7.6	80.27±3.8	97.60±2.5
5	87.0±4.5	88.19±1.9	92.59±6.8	72.90±6.4	97.30±1.9
6	92.4±2.0	98.56±1.2	96.63±7.0	95.62±5.9	99.45±0.9
7	0±0	53.33±5.2	93.33±19.3	74.81±41.5	86.67±23.7
8	98.1±1.4	98.33±2.3	99.52±1.1	99.69±1.0	99.82±0.7
9	0±0	82.11±24.2	69.47±32.6	35.79±14.7	94.74±5.3
10	65.8±3.7	92.44±3.7	91.55±1.7	70.25±31.8	95.36±0.7
11	85.3±2.9	98.28±0.9	96.56±3.9	94.07±2.2	99.34±0.2
12	69.6±6.5	91.37±5.0	93.89±6.3	78.69±12.6	95.84±1.0
13	92.3±4.1	94.97±2.2	98.46±1.5	92.41±3.2	98.67±1.2
14	96.6±1.0	99.00±0.4	98.37±2.2	95.47±8.1	99.15±0.6
15	41.7±7.0	88.94±10.2	90.25±12.9	78.31±6.6	94.77±2.9
16	75.2±9.0	85.45±4.8	93.64±10.7	85.00±11.1	83.64±6.4
OA（%）	77.8±0.8	93.17±1.1	94.69±0.5	86.67±2.4	96.95±0.5
AA（%）	61.3±1.4	88.42±2.5	93.14±3.4	80.74±4.6	95.34±1.9
Kappa(%）	74.5±1.0	92.20±1.3	93.94±0.6	84.71±2.9	96.52±0.5

information are aquired by CNN. Then, ConvLSTM works as shown in 2.1 to extract joint spatial and spectral information for shallow and deep features.

After processing by ConvLSTM, global average pooling is adopted to regularize the structure of the entire network to prevent overfitting. Then the features obtained form several ConvLSTM units are concatenated and sent to softmax layer, which is an end-to-end classification. The combination of high-level semantic features and low-level detailed features provide complementary information for a better classification results.

2.3 The Architecture of MCNN-ConvLSTM

For CNN in MCNN-ConvLSTM, five convolutional layers are stacked one by one to extract hierarchical abstract features. Each convolutional layer is followed by a max pooling layer. In addition, batch normalization (BN) is used to accelerate the training process. Dropout is adopted in the fourth and fifth convolutional layers to alleviate over-fitting. The set of our kernel size is listed in TABLE I.

TABLE I THE SET OF OUR KERNEL SIZE

Layer Name	C1	C2	C3	C4	C5
Kernal Size	4×4	3×3	3×3	3×3	3×3
Feature Maps	64	128	128	256	256

For ConvLSTM, only features extracted form C1,C3 and C5 are fed into three ConvLSTMs. The time step of ConvLSTM is set to be 8 and the numbers of hidden layer node are 32, 64 and 128. The sizes of convolutional kenels of ConvLSTM are set as 5×5, 3×3, 3×3 respectively.

As shown in Fig. 1, the network consists of two softmax layers S1 and S2. S1 is connected with CNN and S2 is connected with ConvLSTM. In the training phase, the loss function consists of two cross entropy produced in two softmax layers and the loss is the sum of these two cross entropies. S1 is a auxiliary classifier, which helps ConvLSTM to improve the ability of feature extraction. S1 works during the training process and S2 is used during both training and testing phases.

3. EXPERIMENTAL RESULTS

In this section, we investigate the performance of proposed MCNN-ConvLSTM on the Indian Pines HSI dataset . The classification performances are measured by kappa coefficient (Kappa), average accuracy (AA) and overall accuracy (OA).

3.1 Data Description

The Indian Pines dataset is a mixed vegetation site over the Indian Pines test area in Northwestern Indian. It was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor, with the size of 145×145 pixels. There are 220 spectral bands in the wavelenth range of 0.4-2.5μm in the visible and infrared spectrum. In the experiments, 200 spectral bands are preserved after 20 lower signal-to-noise ratio bands being diacarded. The dataset contains 16 different land-cover classes.

3.2 Experimental Setting

The performance of the proposed method is compared with some state-of-the-art method for HSI classification, which includes CRNN [7], SVM [11], SSUN [12], DC-CNN [13]. The dataset is divided into 5% training set and 95% test set randomly. All the experimental results are obtained by averaging 20 independent runs. The experiments are carried out by using Python language based on TensorFlow library on NVIDIA 2080Ti graphics card.

For RBF-SVM, multi-classification is dealt with one-against-all strategy and the penalty and gamma parameters in RBF-SVM are determined by five-fold cross validation. For DC-CNN, the spatial window of 1D-CNN is 3×3 and 2-DCNN is 37×37. In SSUN, the size of spatial window is set

as 14×14×4. The numbers of node of two hidden layers in CRNN are set as 128 and 256 respectively. In the proposed method, the spatial window size is set as 27×27×200.

3.3 Classification Results of the Indian Pines Hyperspectral Dataset

The classification results on the Indian Pines dataset are listed in TABLE II. The best classification results of every row are stressed in gray regions. Compared with RBF-SVM, other deep learning method are achieved better classification performance thanks to hierarchical nonlinear feature extraction. The reson why CRNN gets unsatisfing results is that only spectral information is used. In contrast with the SSUN and DC-CNN, we take global spectral information into consideration and fuse shallow and deep features, so our method outperforms them. Among the five methods, the MCNN-ConvLSTM achieves the best classification performance among most class. MCNN-ConvLSTM improves the classification performance than the best baseline by 2.26% in the OA index, 2.2% in the AA index and 2.58% in the Kappa index.

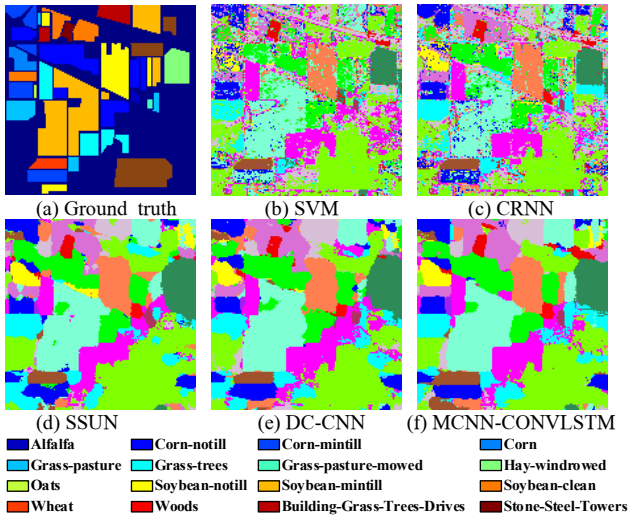

Fig. 2. (a) Ground truth and (b)-(f) Classification visual maps of the Indian Pines dataset by RBF-SVM, SVM, CRNN, SSUN, DC-CNN, MCNN-CONVLSTM respectively.

The classification maps of all the algorithms for the Indian Pines dataset are showed in Fig. 2. It is obvious that there is massive noisy scattered points in SVM and CRNN. Compared with two methods, SSUN and DC-CNN, MCNN-CONVLSTM improves the region uniformity significantly and obtains better boundary localization of the Corn-notill and Grass-pasture classes.

4. CONCLUSION

In this letter, a joint multilayer spatial-spectral classification based on CNN and ConvLSTM for hyperspectral image method is proposed. The ConvLSTM we used is able to capture long dependencies in the input sequence data instead of local dependencies,which makes best use of spectral information. In addition, several ConvLSTM units are applied to take shallow, middle and deep into consideration in order to make full use of complementary among different layers. The experimental results demonstrate that the algorithm outperformes the existing methods based on combing CNN and CRNN.

5. REFERENCES

[1] C. I. Chang, Hyperspectral Data Exploitation: Theory and Applications. Wiley-Interscience, Hoboken, NJ, USA, 2007.

[2] A. O. B. Özdemir, B. E. Gedik and C. Y. Y. Çetin, "Hyperspectral classification using stacked autoencoders with deep learning," *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Lausanne, pp. 1-4, 2014,.

[3] T. Li, J. Zhang and Y. Zhang, "Classification of hyperspectral image based on deep belief networks," *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, pp. 5132-5136, 2014.

[4] Wei H , Yangyu H , Li W , et al. "Deep Convolutional Neural Networks for Hyperspectral Image Classification[J]." *Journal of Sensors*, 2015:1-12, 2015.

[5] Zhang, H., Li, Y., Zhang, Y., Shen, Q., "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network, " *Remote Sens*. Lett. 8 (5), pp. 438–447, 2017.

[6] Y. S. Chen, H. L. Jiang, and C. Y. Li, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens*., vol. 54, no. 10, pp. 6232-6251, Feb. 2016.

[7] Wu H, Prasad S, "Convolutional recurrent neural networks forhyperspectral data classification," *Remote Sensing*, pp. 9(3): 298, 2017.

[8] A. Graves. "Generating sequences with recurrent neural networks, " arXiv preprint arXiv:1308.0850, 2013

[9] Schmidhuber, Juergen . Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies.Wiley-IEEE Press, 2001.

[10] Mou L, Ghamisi P, Zhu X X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens*, 2017, 55(7): 3639-3655.

[11] F. Melgani, L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens*., vol. 42, no. 8, pp. 1778-1790, Aug. 2004.

[12] Y. Xu, L. Zhang, B. Du and F. Zhang, "Spectral–Spatial Unified Networks for Hyperspectral Image Classification, " *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5893-5909, Oct. 2018.

[13] Zhang H , Li Y , Zhang Y , et al. "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sensing Letters*, 8(5):10, 2017.