GRU WITH SPATIAL PRIOR FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Erting Pan[1], *Yong Ma*[1,2], *Xiaobing Dai*[1,2], *Fan Fan*[1,2,*], *Jun Huang*[1,2], *Xiaoguang Mei*[1,2], *Jiayi Ma*[1,2]

[1]Electronic Information School, Wuhan University, Wuhan, 430072, China
[2]Institute of Aerospace Science and Technology, Wuhan university, Wuhan, 430072, China

ABSTRACT

Neural networks have been successfully used to extract deep features for many hyperspectral tasks. In this study, we propose a tiny effective model based on gate recurrent unit (GRU) with spectral-spatial information for hyperspectral image classification. In our method, the core GRU cell can learn interspectral correlations within an entirely continuous spectrum input, and spatial information is the initial state of this GRU cell as a prior. Experimental results demonstrate that our method can fully utilize spectral and spatial information to obtain competitive performance.

Index Terms— hyperspectral image classification, spectral-spatial, GRU

1. INTRODUCTION

Modern hyperspectral sensors can observe the characteristics of hundreds of continuous observation bands throughout the electromagnetic spectrum with high spectral resolution, making it possible to study the chemical properties of scene materials remotely [1]. Hence, the analysis of hyperspectral imagery has attracted more and more attention in the remote sensing. Hyperspectral images based on abundant spectral and spatial information, have been widely applied in many fields such as agriculture, mining, environmental monitoring, land-cover mapping [2, 3, 4, 5, 6, 7, 8].

A hyperspectral image can be described as a 3D cube, and in its three-dimension structure, two of them belong to the spatial dimension, where we can get spatial characteristics. The other dimension is the spectral dimension, which consists of the reflection values of hundreds of narrow, continuous spectral bands from the visible to the infrared, can be expressed as a continuous curve.

Hyperspectral image classification, which aims to identify each pixel vector into a discrete set of specific classes, is one of the hot topics in the remote sensing community. Many methods have been proposed in the last few decades, for instance, some traditional approaches design for different hand-crafted features, such as support vector machine (SVM) or sparse representation classifier [9, 10, 11]. However, as the increase of spectral channel and spatial variability of spectral signature, these methods cannot extract robust deep feature representations due to their shallow properties.

To address the problem mentioned before, deep learning methods, which seem the most prosperous machine learning methods nowadays, have been proposed with a prominent strategy. Unlike traditional classifiers, these methods exploit feature representation learning exclusively for abundant data. Deep convolutional neural networks (CNN) and deep recurrent neural networks (RNN) have gained great success in a variety of computer vision tasks. Networks with one-dimensional [12], two-dimensional [13], and three-dimensional [14] convolution layers or a combination of CNN and RNN has been developed for hyperspectral image analysis.

The one-dimensional approach takes spectrum as an input and learns to capture only the features of spectral dimension. For spectral feature classification using 1D CNN, the spectral feature of the original image data are directly deployed as input vectors [15]. As for RNN, Mou *et. al* [12] models pixel spectra as a 1D sequence for classification. The two-dimensional methods uses convolutional layers to train on image patches in the principal components of spatial domain. A three-dimensional network that directly learns spatial-spectral features over both spatial and spectral axes is superior to a model based only on spectral or spatial information [13]. Therefore, many spectral-spatial methods have been developed that additionally consider spatial correlation information.

In this paper, we propose a novel structure for hyperspectral image classification, as shown in Fig. 1. The contribution of this work can be summarized as follows. i) We design a novel spatial-spectral deep learning-based method, which is a joint model with spectral and spatial information, and the network can learn features automatically. ii) Taking the hyperspectral spectral data as a 1D sequence, we use the GRU cell to extract spectral features. Considering the high correlation between the neighboring bands, we feed the spectrum data into a GRU cell at one timestep instead of the band-by-band strategy, which greatly simplifies the model. iii) We capture the context dependency of adjacent pixels as a priori information about the model. In this work, spatial neighbor infor-

*Corresponding author (email: fanfan@whu.edu.cn).

This research was funded by the National Natural Science Foundation of China under grant nos. 61805181, 61773295 and 61605146.

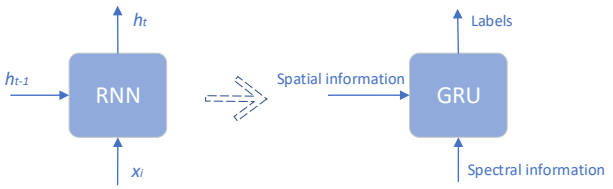


Fig. 1: The left part shows a traditional RNN structure, which contains many sub-units, the right part shows our framework for hyperspectral image classification, both spectral information and spectral information are training in one GRU cell.

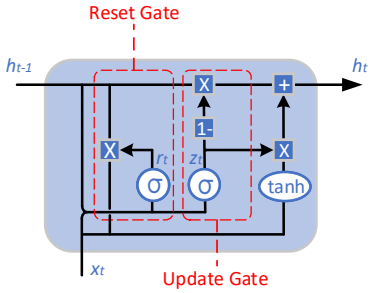


Fig. 2: Illustration of GRU cell.

mation participates in training as the initial state of the GRU unit. By adding spatial features, our model is built to be more robust.

2. RELATED WORKS

RNNs are much of concern for modeling sequence data. Unlike feedforward neural networks, RNN is called recurrent because of its recurrent hidden state, and its activation at each step depends on the previous computations. RNN has a memory function that remembers the information about what has been calculated so far.

A simple RNN unit is shown in the left of Fig. 1, the input x_t and previous hidden state h_{t-1} are combined to form a vector, which contains information on the current input and previous inputs. And the output of this RNN unit is the new hidden state, or the memory of the network. In other words, h serves two purposes: the hidden state for the previous sequence data as well as making a prediction.

The most commonly used RNN types are Long Short-Term Memory (LSTM) or GRU architectures, which are explicitly designed to deal with vanishing gradients and efficiently capture long-term dependencies. These two architectures have no fundamentally difference from RNN, but they use a different function to compute the hidden state.

LSTMs were first proposed in 1997 [16] and are the perhaps most widely used models in NLP today. The memory in LSTMs are called cells and and can be regarded as black boxes that take the previous state h_{t-1} and current x_t as input. Internally these cells decide what to keep in (and what to erase from) memory. They use three gates to combine the previous state, the current memory and the input to control what information will be passed through. It turns out that these types of units are very efficient at capturing long-term dependencies. GRUs (see Fig. 2), first proposed in 2014 [17], are simplified versions of LSTMs. Compare with LSTM, GRU does not maintain a cell state *C* and uses two gates instead of three. GRUs have fewer parameters and thus may train a bit faster or need less data to generalize.

A GRU has two gates, *i.e.*, a reset gate r_t and an update gate z_t:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \tag{1}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]). \tag{2}$$

Intuitively, the reset gate determines how to combine the new input with the previous memory, and it acts similar to the forget and input gate of an LSTM. It decides what information to throw away and what new information to add. The update gate defines how much of the previous memory to keep around. If we set the reset gate to all 1 and update gate to all 0 we again arrive at our plain RNN model. The new hidden state is compute as:

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t, \tag{3}$$

$$\widetilde{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t]). \tag{4}$$

where $\sigma(\cdot)$ denotes a logistic sigmoid function, and $tanh(\cdot)$ is the hyperbolic tangent function.

3. METHODOLOGY

Our proposed methodology is illustrated in Fig. 3. It is a tiny but effective network.

Clearly, the core member in our model is the GRU cell. For every single pixel in the original hyperspectral data, the spectrum data actually is a continuous curve. From the point of sequential view, a direct way is considering each channel as a time step and input the GRUs channel by channel. But this way would make the whole network become too deep. Our strategy is to input the whole spectrum data in one GRU cell directly. Considering the indispensable spatial information, we put the spatial characteristics of adjacent pixels as the initial state of GRU, and it is equivalent to priori of the classification problem. Therefore, we combine spatial and spectral information of the hyperspectral image, train them at the same time and get a sensational performance.

As we mentioned before, the value of each spectral channel in the spectrum is correlated. That is the reason why RNN is cascaded by multiple GRUs to learn spectral features automatically. For the same reason, we put forward a new way, that is to input the whole spectrum directly to one GRU cell. In a manner, a GRU cell is one kind of deformation of fully connected layer, and the difference is that GRU can customize the initial state and it can filter information internally with the reset gate and update gate.

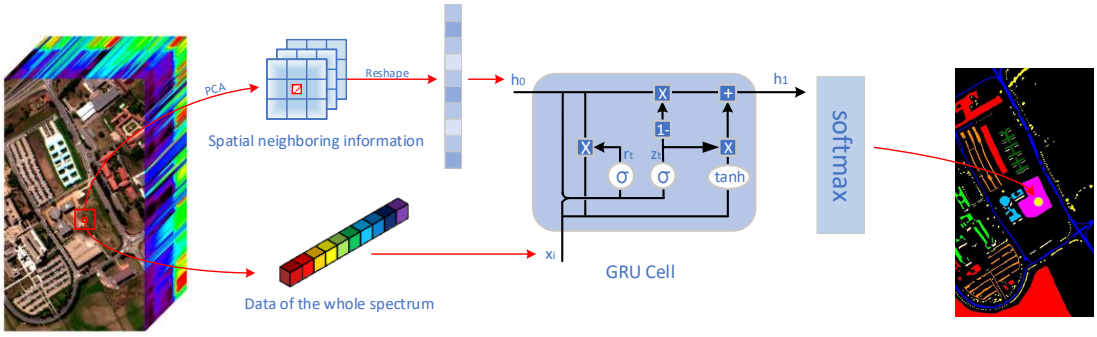


Fig. 3: A illustration of the proposed method, the entire spectral vector is input the GRU cell at one timestep, the spatial neighbor information after PCA and reshape is used as the initial state to participate in the training.

Spatial feature is a valuable complement to the spectral signatures. Similar to the correlations across spectral dimension, there are also spatial dependencies between the neighboring pixels in a hyperspectral image. For a certain pixel in the original hyperspectral image, it is natural to consider its neighboring pixels to extract the spatial feature representation.

With hundreds of spectral bands, it is necessary to reduce the spectral feature dimensionality before the spatial feature representation. PCA is commonly executed in the first step to map the data to an acceptable scale with a low information loss. After PCA, for instance, the first three components of the Pavia University dataset are reserved because they have almost 99.3% information. Then, in the second step, the spatial information is collected by the use of a $k \times k \times 3$ neighboring region of every certain pixel in the original image. In 2D CNN, a common way is to choose a larger patch around the target pixel and sliding window with a 3×3 or 5×5 kernel. With the same way, our method selects a neighbor region with an appropriate size which contains almost all relevant spatial information. In order to meet the requirement of the GRU initial state input, we transform the selected spatial information into one-dimensional data. Training the initial state as a variable can improve model performance.

4. EXPERIMENTAL RESULTS AND ANALYSIS

We train and test our method on two public hyperspectral image classification datasets, namely, the Pavia University dataset and the Pavia Center dataset. Both of them contain 9 land cover classes in urban areas. To overcome the class imbalance problem, We split these datasets into training, validation, and test sets, and select 200 samples for training, 100 for validation of each labeled class randomly.

We compare our method with four state-of-the-art classification methods, such as SVM with radial basis function kernel, deep learning method 2DCNN, and RNN with different input types like input band-by-band or entirely. 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 (AA), and the kappa coefficient are used as the evaluation measurements for the compared methods. The experimental results of the Pavia University dataset are shown in Table 1, and the results of the Pavia Center dataset are presented in Table 2. The classification results of both datasets show that our proposed method, GRU with spatial prior (SPGRU), exhibits the best performance among all compared methods in all scenarios.

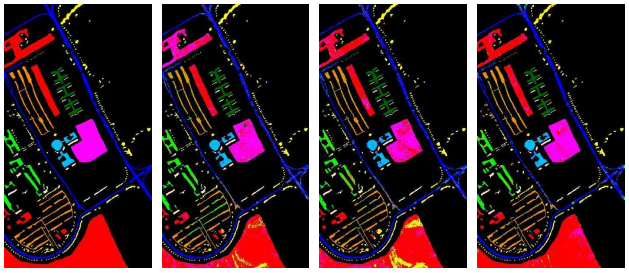
Table 1: Classification performance of different methods for the Pavia University dataset. Bold indicates the best result.

Label	SVM	CNN (2D)	RNN (band by band)	RNN (all)	SPGRU
OA	84.43	89.20	91.68	97.24	**98.38**
AA	88.59	92.20	86.68	88.51	**93.82**
Kappa	79.94	85.91	88.84	92.34	**95.49**

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

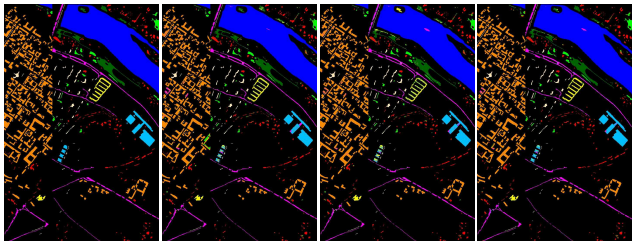
Label	SVM	CNN (2D)	RNN (band by band)	RNN (all)	SPGRU
OA	84.48	86.20	96.83	97.24	**99.73**
AA	84.88	91.20	91.12	91.73	**95.89**
Kappa	83.0	68.91	95.81	96.09	**99.18**

The results indicate that the proposed model is effective in hyperspectral image classification. The traditional SVM demonstrates poor performance. Deep learning methods, such as CNN and RNN, are effective because of their discriminative features. A comparison of two input ways of RNN indicates that our strategy performs better when it comes to inner spectral 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 network combines features from spatial and spectral domain and acquires well-balanced results. We



(a) (b) (c) (d)

Fig. 4: Visual results on the Pavia University dataset. (a) gt, (b) RNN (band by band), (c) RNN (all), (d) SPGRU.



(a) (b) (c) (d)

Fig. 5: Visual results on the Pavia Center dataset. (a) gt, (b) RNN (band by band), (c) RNN (all), (d) SPGRU.

show the classification maps in Figs. 4 and 5.

5. CONCLUSION

In this study, a tiny effective model is proposed to extract spectral-spatial features based on a GRU cell for hyperspectral image classification. By adding spatial information as the trainable initial state with an entire spectra data input, we can learn spatial contextual features in spatial dimensions and numerous inner spectral correlations in the continuous spectrum domain. Analysis of experimental results on two datasets shows that our method not only outperforms other state-of-the-art methods but also extracts more homogeneous discriminative feature representations. We will generalize our method for other remote sensing applications, such as unmixing and change detection, in the future.

6. 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," *Inf. Sci.*, 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 Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, 2016.

[3] Jiayi Ma, Chang Li, Yong Ma, and Zhongyuan Wang, "Hyperspectral image denoising based on low-rank representation and superpixel segmentation," in *Proc. ICIP*, 2016, pp. 3086–3090.

[4] Haobo Lyu, Hui Lu, and Lichao Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, pp. 506, 2016.

[5] Jiayi Ma, Huabing Zhou, Ji Zhao, Yuan Gao, Junjun Jiang, and Jinwen Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, 2015.

[6] Junjun Jiang, Jiayi Ma, Chen Chen, Zhongyuan Wang, Zhihua Cai, and Lizhe Wang, "Superpca: A superpixelwise pca approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, 2018.

[7] Yong Ma, Chang Li, Xiaoguang Mei, Chengyin Liu, and Jiayi Ma, "Robust sparse hyperspectral unmixing with $\ell_{2,1}$ norm," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1227–1239, 2017.

[8] Jiayi Ma, Junjun Jiang, Huabing Zhou, Ji Zhao, and Xiaojie Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, 2018.

[9] Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu, "Deep learning-based classification of hyperspectral data," *IEEE JSTARS*, vol. 7, no. 6, pp. 2094–2107, 2014.

[10] Chang Li, Yong Ma, Xiaoguang Mei, Chengyin Liu, and Jiayi Ma, "Hyperspectral image classification with robust sparse representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 641–645, 2016.

[11] Junjun Jiang, Jiayi Ma, Zheng Wang, Chen Chen, and Xianming Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 851–865, 2019.

[12] Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, 2017.

[13] 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 Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[14] Yonghao Xu, Liangpei Zhang, Bo Du, and Fan Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, 2018.

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

[16] Sepp Hochreiter and Jurgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.