GRU WITH SPATIAL PRIOR FOR HYPERSPECTRAL IMAGE CLASSIFICATION

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

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

ABSTRACT

Deep learning methods 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 priori. Experimental results demonstrate that our method can fully utilize spectral and spatial information to obtain competitive performance.

***Index Terms*—** hyperspectral image classification, deep learning, RNN

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].

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 mainly include low-level features, such as location and distribution information, texture features. The other dimension is the spectral dimension, and its data, which consist of the reflection values of hundreds of narrow, contiguous spectral bands from visible to middle infrared wavelength ranges, can be expressed as a continuous curve, which represents the chemical composition of this pixel.

Hyperspectral image classification, which aims to identify each pixel vector into a discrete set of specific classes, is one of hotspot topics in the remote sensing community. Many methods have been proposed in the last few decades, for instance, some traditional approaches designed for different hand-crafted features, such as support vector machine (SVM) or sparse representation classifier [3, 4]. 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 s-trategy. Unlike traditional classifiers, these methods exploit feature representation learned 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 [5], two dimensional [6], and three dimensional [7] convolution layers or combination of CNN and RNN have been developed for hyperspectral image analysis.

One dimensional approaches take spectra as input and learn features that capture only spectral information. For spectral feature classification with 1D CNNs, the spectral feature of the original image data is directly deployed as an input vector [8], as for RNN, from a sequential point of view, Mou *et. al* [5] modeled pixel spectra as a 1D sequences for classification. Methods that use two dimensional convolutional layers are trained on the principle component bands of image patches around the pixel in spatial domain, and three dimensional networks which directly learn spatial-spectral features over both spatial and spectral axes outperform ways only based on spectral or spatial information [6]. Therefore, many spectral-spatial methods have been developed that additionally consider spatial correlation information.

In this paper, we proposed a novel structure for hyperspectral image classification, as shown in Fig. 1. The contribution of this work can be summarized as follows:

- We design a novel spatial-spectral deep learning-based method, which is a joint framework with spectral and spatial information, and the network can learn features automatically.

- Take the hyperspectral spectral data as a 1D sequence, we use gate recurrent unit (GRU) to extract spectral features. Instead of the band by band strategy, considering the high correlations between the reflectance of the neighboring bands, we feed the whole spectrum data

This work was supported by the National Natural Science Foundation of China under Grant no. 61605146.

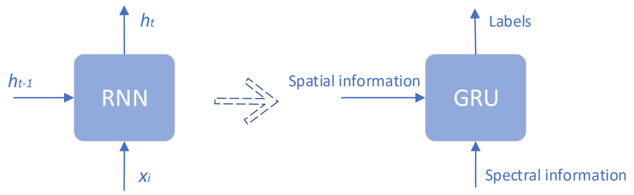

Fig. 1: Our framework for hyperspectral image classification.

into a GRU cell at one time.

- We capture the contexture dependency of adjacent pixels as the priori input of classification. In this work, spatial neighboring information is the initial state of an GRU unit. Our model can be exploited be build more robust with spatial features.

2. RELATED WORKS

RNNs are much of concern for modeling sequential data. Unlike feedforward neural networks, RNN are called recurrent because of its recurrent hidden state, whose activation at each step depends on the previous computations. RNNs have a memory function, which can remember 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 type of RNNs are Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) architectures, which are explicitly designed to deal with vanishing gradients and efficiently at capturing long-term dependencies. These two architectures do not have a fundamentally different architecture from RNNs, but they use a different function to compute the hidden state.

LSTMs were first proposed in 1997 [9] 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 [10], 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.

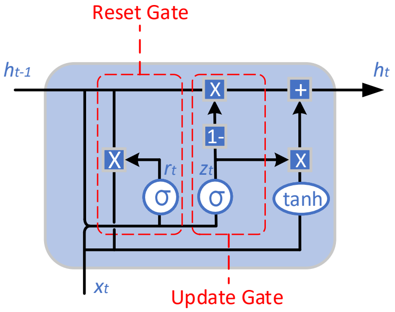

Fig. 2: Illustration of GRU cell.

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 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 makes 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 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,

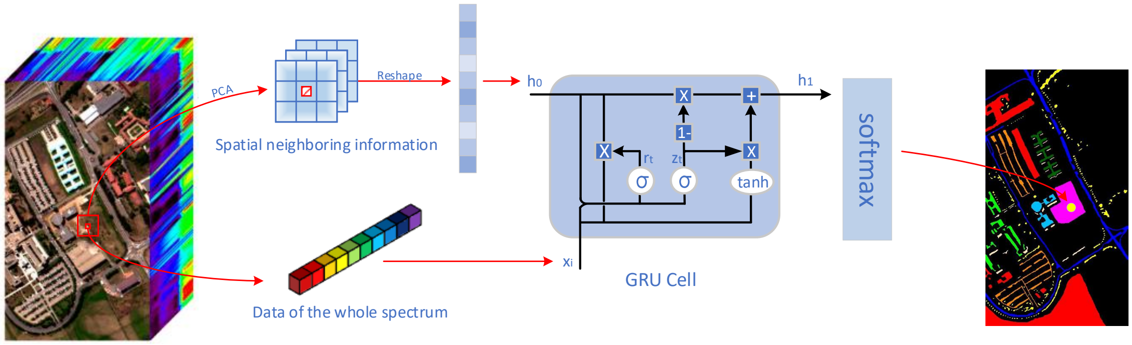

Fig. 3: A illustration of the proposed method.

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 internal with the reset gate and update gate.

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. This is due to the fact that the material properties in a natural scene vary smoothly in space and the presence of a material can increase or decrease the likelihood of the occurrence of another material in its vicinity. 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 spectra 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. However, our method overleaps this and selects an 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. It has been proved that training the initial state as a variable can improve the 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 nine land cover classes of 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.

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

Label	SVM	CNN	RNN	RNN	SPGRU
		(2D)	(band by band)	(all)	
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	RNN	RNN	SPGRU
		(2D)	(band by band)	(all)	
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**

We compare our method with four state-of-the-art classification methods, such as linear 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 priori (SPGRU), exhibits the best performance among all compared methods in all scenarios.

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 dis-

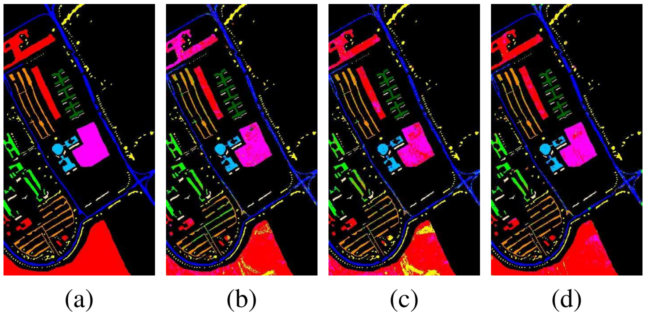

Fig. 4: Visual results on the Pavia University dataset. (a) gt, (b) RNN (band by band), (c) RNN (all), (d) SPGRU.

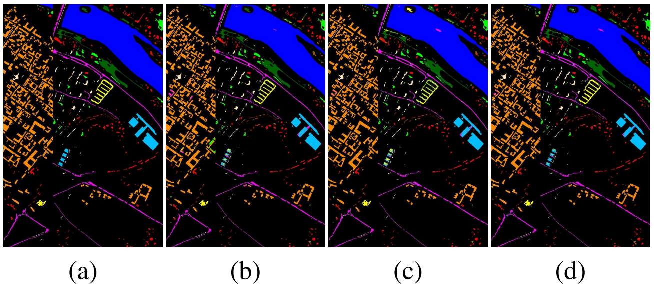

Fig. 5: Visual results on the Pavia Center dataset. (a) gt, (b) RNN (band by band), (c) RNN (all), (d) SPGRU.

criminative features. A comparison of two types of RNN indicates that the our strategy perform better when it comes to 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 network combines spatial and spectral dimensions and acquires well-balanced results. We show the classification maps for our proposed method 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 entirely spectra data input, we can learn spatial contextual features in spatial dimensions and numerous interspectral 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," *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] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.