FUSING 3D-CNN AND LIGHTWEIGHT SWIN TRANSFORMER NETWORKS FOR HSI

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently deep learning has occupied an important position in hyperspectral image (HSI) classification. In this study, we explore the advantages of using convolutional neural networks (CNN) for feature extraction and fusing an advanced shift-window (swin) transformer network based on the transformer model for HSI classification. The swin transformer network attention perception, capable of learning local and global features, can avoid the dependence on single features during HSI classification. The experiments show that our proposed model outperforms traditional machine learning models, and achieves competitive results with advanced models.

1 INTRODUCTION

Hyperspectral image classification are remote sensing images that contain both image information and spectral information. The CNN have been one of the hottest research hotspots in pattern recognition for a long time, and the HSI classification field is no exception (Lee & Kwon (2017)). Swin Transformer is a Vision Transformer (VIT) network based on the self-attentive mechanism (Liu et al. (2021)). The transformer-based architecture processes patch through fixed-length contexts and are limited when applied to HSI classification tasks (He et al. (2021)). And the simple CNN structure tends to lose the global information in different dimensional feature maps because of the fixed size of the convolutional kernel setting.This letter proposes a new framework that fuses CNN and swin transformer. The structure of swin transformer network is improved to make it more suitable for classification processing of hyperspectral data, and Compress the model complexity without affecting the classification accuracy.

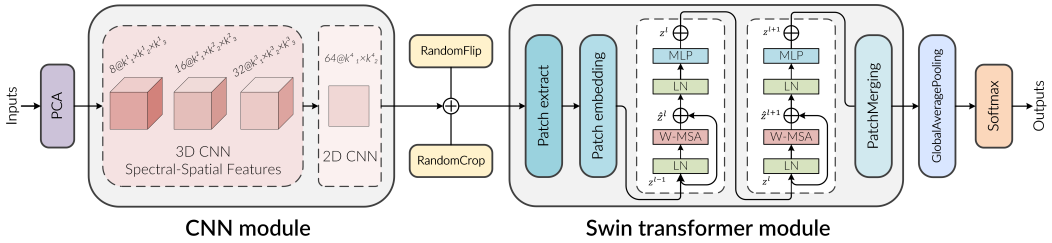

Figure 1: The proposed classification model fusing CNN and Swin transformer structure.

2 PROPOSED APPROACH

Fig 1 shows the structural diagram of the model proposed in this letter. As shown in the figure, this model consists of two parts. The first part is a feature extraction CNN module consisting of three 3D-CNN layer and one 2D-CNN layer. PCA is a dimensionality reduction for HSI. The function of the 3D-CNN module is to extract spatial-spectral features. The 2D-CNN module processes it to conform to the dimensions of the swin transformer module (Roy et al. (2019)).

The swin transformer is a computer vision algorithm. It introduces an innovative sliding window mechanism that restricts the attention computation to fixed window size, allowing the model to

learn features that span the window. The hierarchical design of the subsampling layer enables the model's self-attentive mechanism to take into account both local and global features of the data. Compared with the normal swin transformer structure, we reduce some redundant liner embedding and Patch Merging modules to improve computing efficiency, simplify the structure and adapt to HSI classification. The random-crop is the most commonly used data augmentation method in DL, which can enhance the model's accuracy and stability. And random-flip is a data augmentation method that randomly flips the matrix during training. The patch extract is a further extraction of the enhanced features of the data. The Patch Embedding part is to crop the feature data to the settings window size and embed it in each patch.

Two successive swin transformer blocks are the core of the swin transformer. Shifted Windows Multi-head Self-Attention (W-MSA) can send self-attention calculations across the four Windows in the window layer. The computation of two successive Swin Transformer Blocks can be expressed as equation (1):

$$\begin{gathered} \hat{z}^l = W - MSA(LN(z^{l-1})) + MSA(LN(Z^{l-1})) + z^{l-1} \\ z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \qquad z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \\ \hat{z}^{l+1} = SW - MSA(LN(z^l)) + MSA(LN(Z^l)) + Z^l \end{gathered} \tag{1}$$

The function of the Patch Merging module is to subsample before the beginning of each Stage, to narrow the resolution and adjust the number of channels, and thus form a hierarchical design. Finally, the global average pooling layer will get the output features with a softmax fully connected layer corresponding to different classes.

3 EXPERIMENTAL

We use the Indian Pine (IP) dataset, the Salinas (SA) and the Pavia University (PU) datasets. In this study, adam optimizer trains our proposed model of hybrid CNN and swin transformer. For the IP dataset, the initial learning rate is set to 0.001. The test set division ratio is 70%. In the swin transformer model, we use a patch size of 2×2, a dropout rate of 0.03, a number of attention heads of 8, an embedding dimension of 64, a number of multilayer perceptrons of 256, a window size of 2 and shift window step of 1. The model under comparison uses open source projects as its source of code (Audebert et al. (2019)).

Table 1 shows the comparison of the metrics of the models under the average accuracy (AA), overall accuracy (OA) and Kappa evaluation systems. It can be seen that the proposed model has the same excellent performance as the current mainstream models in terms of both classification reality map and accuracy, and is better than the traditional machine learning models (Lee & Kwon (2016))(Li et al. (2017)).

Table 1: Comparison results of the classification of the three open data sets in each model (%).

	Indian Pines			Salinas			PaviaU		
Methods	AA	OA	Kappa	AA	OA	Kappa	AA	OA	Kappa
3D FCN	0.545	0.702	0.658	0.904	0.939	0.932	0.867	0.976	0.968
S-S 3D CNN	0.744	0.779	0.749	0.908	0.947	0.941	0.876	0.969	0.959
Ours	0.988	0.997	0.9974	0.998	0.999	0.999	0.996	0.998	0.998

4 DISCUSSION

In this letter, a new network with deep CNN and state-of-the-art swin transformer fusion is proposed. The excellent feature extraction capability of CNN is used to extract complex spatial spectral joint features in hyperspectrum. And these features are sent to the lightweight swin transformer global attention container for training.

ACKNOWLEDGEMENTS

The work was supported by the Natural Science Foundation of Heilongjiang Province for Key projects, China (Grant no. ZD2021F004), the Postdoctoral Scientific Research Developmental Fund of Heilongjiang Province, China (Grant no. LBH-Q18110)

URM STATEMENT

Author Yuanjia Liu meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE geoscience and remote sensing magazine*, 7(2): 159–173, 2019.

Xin He, Yushi Chen, and Zhouhan Lin. Spatial-spectral transformer for hyperspectral image classification. *Remote Sensing*, 13(3):498, 2021.

Hyungtae Lee and Heesung Kwon. Contextual deep cnn based hyperspectral classification. In *2016 IEEE international geoscience and remote sensing symposium (IGARSS)*, pp. 3322–3325. IEEE, 2016.

Hyungtae Lee and Heesung Kwon. Going deeper with contextual cnn for hyperspectral image classification. *IEEE Transactions on Image Processing*, 26(10):4843–4855, 2017.

Ying Li, Haokui Zhang, and Qiang Shen. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sensing*, 9(1):67, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Swalpa Kumar Roy, Gopal Krishna, Shiv Ram Dubey, and Bidyut B Chaudhuri. Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 17(2):277–281, 2019.

A APPENDIX

Figure2, 3 and 4 shows the classification chart of the IP dataset for the classification results in our proposed model. Our proposed model achieves the best visual performance on the classification graph. Not only does it perform well on the IP dataset, but also on the SA and PU datasets.

Figure 5 depicts a graphical comparison of the classification accuracy of various models in the IP, SA and PU datasets as training calendar time increases. In comparison to other models, our model achieves higher accuracy in shorter training epochs and has a more stable accuracy curve.

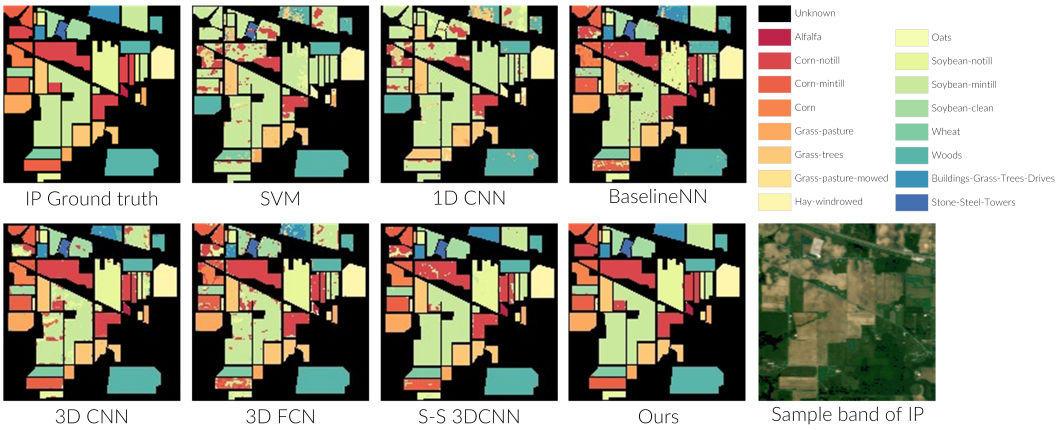

Figure 2: Comparison of IP dataset classification chart with other models.

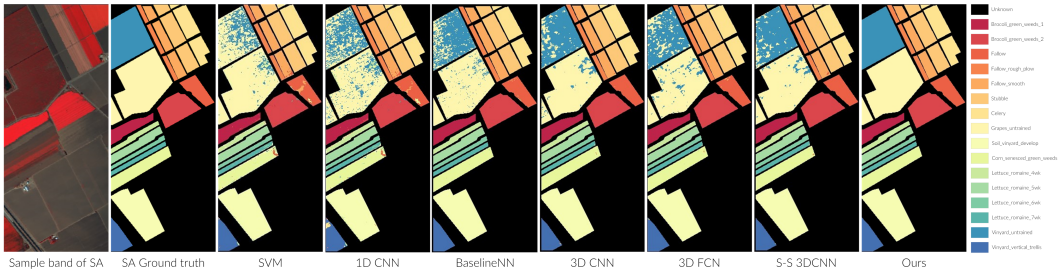

Figure 3: Comparison of SA dataset classification chart with other models.

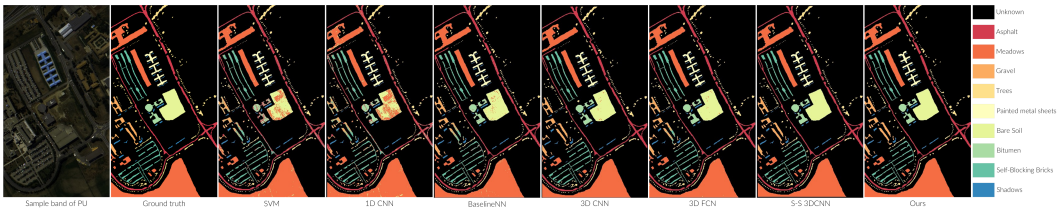

Figure 4: Comparison of PU dataset classification chart with other models.

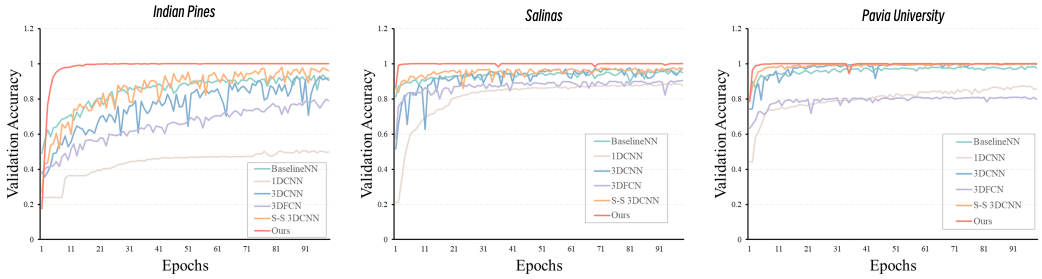

Figure 5: Comparison of model validation accuracy in the IP, SA and PU datasets.