CONVOLUTION ENHANCED SPATIAL-SPECTRAL UNIFIED TRANSFORMER NETWORK FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Ziqi Xin [1], *Zhongwei Li* [1,*], *Mingming Xu* [1], *Leiquan Wang* [2] *and Xue Zhu* [2]

[1] College of Oceanography and Space Informatics, China University of Petroleum(East China), Qingdao, China
[2] College of Computer Science and Technology, China University of Petroleum(East China), Qingdao, China

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9884124

ABSTRACT

Convolutional neural network has achieved great success in hyperspectral image classification for its excellent local context modeling capabilities. However, the convolution operation with fixed-size local receptive fields is difficult to establish long-distance dependence in hyperspectral image. To address this problem, we propose a spatial-spectral unified transformer network, which utilizes self-attention mechanisms to extract global spatial and spectral features. In addition, in order to introduce local spatial and spectral information, the convolution operation is integrated into the network. Specifically, spatial and spectral convolutional embedding layers are designed to generate embeddings of spatial patches and spectral bands. Besides, depth-wise convolution is exploited in the locally-enhanced feed-forward layer to bring locality into transformer. Experimental results on two datasets demonstrate that our proposed network has greatly improved compared with other state-of-the-art methods.

Index Terms— hyperspectral image classification, spatial-spectral unified network, transformer, convolution

1. INTRODUCTION

Each pixel in the hyperspectral image (HSI) is composed of hundreds of continuous spectral bands, carrying rich spectral information. Meanwhile, homogeneous areas in HSI are of high spatial correlation, which can provide spatial contextual information. HSI classification is aimed at assigning a pixel to a certain land cover type based on spectral and spatial features. Because of its wide application in various fields, HSI classification has attracted extensive attention in research.

In recent years, deep learning methods represented by convolutional neural networks (CNN) have made great progress in computer vision field. More and more researchers are applying CNN models to HSI classification tasks. Spectral-spatial unified network(SSUN) [1] employs multiscale 2DCNN and long short-term memory (LSTM) network to obtain spatial and spectral features respectively, and then merge them. CNN can well capture local context information through local connectivity pattern and shared weights, however, it will encounter bottlenecks in establishing long-distance dependencies.

The emergence of transformer effectively solves the above mentioned problem. Transformer [2] was first proposed in the natural language processing(NLP) field and has made significant achievement, later it gradually expanded to the field of computer vision tasks, represented by Vision Transformer(ViT) [3]. It exploits self-attention mechanism to calculate the correlation of each location, thereby establishing the long-range dependencies along either spatial or channel dimension. Because of its advantage in capturing global information, Hong et al. [4] proposed a SpectralFormer network to globally characterize spectral sequential properties thereby improving HSI classification performance. However, it pays more attention to spectral domain and does not make full use of spatial information.

Recently, some studies have tried to combine the locality of convolution with the global connectivity of transformer by integrating convolution into the Vision Transformer network. Wu et al. [5] designed Convolutional Token Embedding Layer and Convolutional Projection Layer to introduce convolutions into Vision Transformer architecture. Yuan et al. [6] proposed a convolution-enhanced image Transformer (CeiT) which designed three modules including Image-to-Tokens Module, Locally-enhanced Feed-Forward Module and Layer-wise Class token Attention Module to strength locality while retaining the advantage of Vision Transformer in modeling long-range dependencies.

Inspired by the above methods, a convolution enhanced spatial-spectral unified transformer (CESSUT) is proposed for HSI classification. The main contributions of the proposed method are as follows: 1) An end-to-end transformer based framework is proposed for HSI classification which consists of spatial transformer branch and spectral transformer branch to extract discriminate features. 2) Spatial and spectral convolutional embedding layers are designed to generate embeddings of spatial patches and spectral bands as the input of transformer. 3) Spatial and spectral locally-enhanced feed-forward layers are exploited to enhance the correlation among neighboring tokens in the spatial and spec-

This research was funded by the National Natural Science Foundation of China (62071491,U1906217).

tral dimension. 4) The experimental results compared with other methods show that our method can further improve the accuracy of HSI classification.

2. METHODOLOGY

The proposed framework of CESSUT is shown in Fig. 1. CESSUT consists of spatial transformer branch and spectral transformer branch extracting spatial and spectral features respectively, and they are finally concatenated for classification.

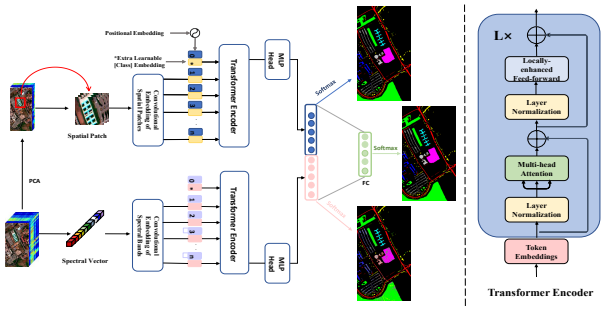

Fig. 1. The framework of CESSUT.

Concretely, for a pixel p_i to be classified, its spectral vector $p_i \in R^{1\times1\times B}$ and spatial patch $H_i \in R^{q\times q\times b}$ are split from the original 3-D hyperspectral image. The spatial transformer branch takes the patch H_i as input, and it is composed of convolutional embedding layer, transformer encoder and MLP head. The proposed convolutional embedding layer aims to map the patch H_i into embeddings which serve as the input of transformer encoder. The transformer encoder mainly consists of multi-head self-attention and locally-enhanced feed-forward layer, respectively extracting global and local spatial information. MLP head aims to assign the most probable category to the pixel based on the extracted features. Similarly, spectral transformer branch takes the spectral vector p_i as input, also consists of convolutional embedding layer and transformer encoder to extract spectral features. It is worth noting that the convolutional embedding layer and locally-enhanced feed-forward layer is designed separately based on the characteristics of spatial patch and spectral vector. At last, spatial and spectral features are fused for HSI classification. The details of convolutional embedding layer, transformer encoder and feature fusion are introduced as follows.

2.1. Convolutional Embedding Layer

ViT [3] divides the image into patches, flatten the patches and linearly maps into 1D vectors. However, some low-level information (such as edges) will inevitably be lost in this way. To alleviate this problem, convolutional embedding of spatial patches and spectral bands are designed separately. Fig. 2 illustrates the process of convolutional embedding layer, for a spatial patch $H_i \in R^{q\times q\times b}$, a 2D convolution operation of kernel size $k_1 \times k_1$ and s_1 stride is utilized to map H_i into a feature map $H_{i+1} \in R^{m_1\times m_1\times d}$, where $m_1 = \lfloor (q - k_1/s_1) + 1 \rfloor$ denoting the width of H_{i+1} and d represents the channel size of token embeddings. H_{i+1} is then flattened into size $m_1 \times m_1 \times d$ which is entered as input to the transformer encoder, and there is a layer normalization layer to normalize H_{i+1} before flatten operation. For a spectral vector $p_i \in R^{1\times1\times B}$, we adopt a 1D convolution of kernel size k_2 and s_2 stride to map multiple neighboring bands into an embedding with a channel size d. Convolutional embedding layer takes advantage of the convolution's ability to extract low-level features, so that each generated token embedding contains local spatial or local spectral information.

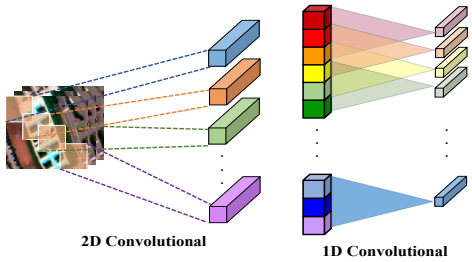

Fig. 2. The convolutional embedding layer.

In addition, like ViT [3], we also add a learnable embedding named class embedding before all generated token embeddings, which represents the output of the transformer encoder and is utilized for the final classification. At the same time, position embeddings are added to each token embedding to introduce position information.

2.2. Transformer Encoder

The transformer encoder consists of alternating layers of multi-head self-attention layer, locally-enhanced feed-forward layer and layer normalization which is utilized before the two layers mentioned above. Moreover, residual connections are adopted after layers. The detailed structure of multi-head self-attention layer and locally-enhanced feed-forward layer are explained as follows.

2.2.1. Multi-head Self-attention Layer

The purpose of the self-attention mechanism is to calculate the correlation between each token embedding and all other token embeddings so as to obtain global information. Each token embedding is multiplied by three different matrices to get queries (i.e.,Q), keys(i.e.,K), and values(i.e.,V) three matrices. Each query calculates the similarity with all keys through dot product, then a Softmax operation is employed to get the attention weight on the values. The whole calculation process can be formulated by:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V. \quad (1)$$

Multi-head self attention linearly projects the input h times with different linear projections to multiple feature subspaces and processes them with independent self-attention mechanism in parallel. The results are concatenated and projected again to obtain the final ouput. The process of MHSA is shown in the following formual:

$$\begin{aligned} &Q_i = XW^{Q_i}, K_i = XW^{K_i}, V_i = XW^{V_i}, \\ &Z_i = Attention(Q_i, K_i, V_i), i = 1, 2...h, \\ &MultiHead(Q, K, V) = Concat(Z_1, Z_2...Z_h)W^O. \end{aligned} \tag{2}$$

2.2.2. *Locally-Enhanced Feed-forward Layer*

In ViT [3], feed-forward layer is composed of two fully connected layers, which ignore the relationship of neighboring token embeddings. Ceit [6] first proposed locally-enhanced feed-forward layer by using depth-wise convolution to promote the correlation among neighboring token embeddings. Based on this, we have designed locally-enhanced feed-forward layer for spatial and spectral features respectively.

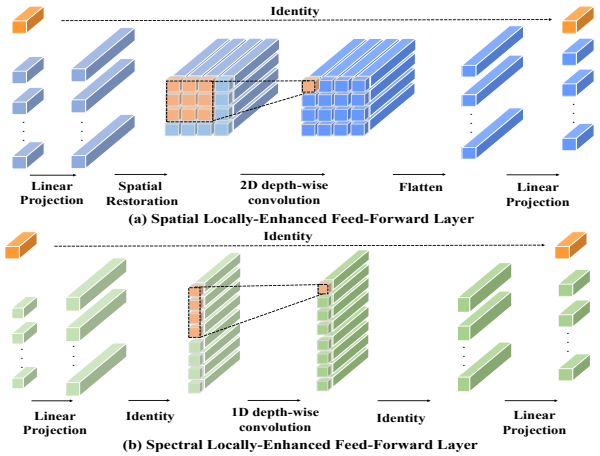

Fig. 3. The details of locally-enhanced feed-forward layer.

For spatial features, as shown in Fig. 3(a), the locally-enhanced feed-forward layer can be performed according to the following procedures. First, given the feature embeddings $z^h \in R^{(N+1)\times d}$ generated by the previous multi-head self-attention layer, class embedding $z_c^h \in R^{1\times d}$ and other feature embeddings $z_f^d \in R^{N\times d}$ are split,and then feature embeddings are projected to higher dimensions $z_f^{l_1} \in R^{N\times(d\times e)}$,where e is a scale factor. Second, the feature embeddings are rearranged in the spatial dimension based on the position in the original input spatial patch, so a feature map $z_f^r \in R^{\sqrt{N}\times\sqrt{N}\times(d\times e)}$ is obtained. Third, a 2D depth-wise convolution of kernel size k_{d_1} , s_{d_1}stride and padding p_{d_1} is applied on the feature map z_f^r to extract spatial information in local neighborhoods, yielding a locally-enhanced feature map $z_f^d \in R^{\sqrt{N}\times\sqrt{N}\times(d\times e)}$. Fourth, the feature map is flattened into $z_f^f \in R^{N\times(d\times e)}$ in the order of the original feature embeddings sequence. Fifth, another linear projection is utilized to reduce the dimensions of feature embeddings to d to get $z_f^{l_2} \in R^{N\times d}$. Last, $z_f^{l_2}$ and class embedding z_c^h are concatenated together again, resulting in $z^h \in R^{(N+1)\times d}$.

For spectral features, as shown in Fig. 3(b), because the input spectrum is originally one-dimensional, there is no need to reconstruct it into the form of a 2D image. In addition, 1D depth-wise convolution is utilized to learn the spectral information in the local neighborhood.

2.3. Feature Fusion and Classification

The MLP head consists of a normalization layer and a fully connected layer which aims to project the dimension of class embedding to the number of categories. And then Softmax operation is applied to calculate the probability of each category. In order to integrate spatial transformer, spectral transformer and classifier into a unified network, the last fully connected layer in two branches are concatenated, then another fully connected layer and a softmax function are used to calculate the final classification probability based on joint spatial-spectral features. The loss of proposed model is :$L = L^{joint} + L^{spatial} + L^{spectral}$. where L^{joint} is joint classification loss, $L^{spatial}$ and $L^{spectral}$ are spatial and spectral classification loss respectively. All three losses employ the cross-entropy loss function.

3. EXPERIMENTS

3.1. Datasets and Experimental settings

Two public hyperspectral datasets are selected for experimentation, i.e., University of Pavia (UP) and Salinas Valley(SV). The UP dataset was collected by ROSIS sensor, containing 610×340 spatial pixels and 103 spectral bands. There are 9 categories to be distinguished in the UP dataset. The SV dataset was gathered by AVIRIS sensor and comprises 512×217 pixels and 204 bands. SV dataset has 16 classes for classification.

Adam optimizer is utilized in our experiments, and the batch size is set to 16. In addition, the learning rate is $5e-4$ and the number of training epochs is 60. Regarding some parameters in the model, a 64-length embedding generated by convolutional embedding layer is input into transformer encoder which comprises 8 cascaded blocks. And the head number of MHSA is set to 8.

3.2. Comparison With Other Approaches

To verify the performance of the method, we compare it with other classic methods, including conventional classifier SVM, CNN-related methods 2DCNN, 3DCNN and SSUN [1], and transformer-based method SpectralFormer [4]. Moreover, the results of spatial branch and spectral branch also included in the comparison. 3% of the samples are randomly selected for

training. The classification results and maps are set out in Table 1- 2 and Fig. 4.

Table 1. Classification results of the UP dataset.

Number	SVM	2DCNN	3DCNN	SSUN	Spectral Former	Spatial Branch	Spectral Branch	CESSUT
OA	92.10	95.51	95.84	98.38	84.09	98.63	88.94	**99.13**
AA	91.44	93.27	93.45	96.86	74.66	97.81	85.23	**98.72**
Kappa	89.45	94.02	94.47	97.86	78.40	98.18	85.32	**98.85**

Table 2. Classification results of the SV dataset.

Number	SVM	2DCNN	3DCNN	SSUN	Spectral Former	Spatial Branch	Spectral Branch	CESSUT
OA	91.60	94.17	94.45	98.53	87.41	98.39	90.18	**98.98**
AA	95.02	96.39	97.07	99.01	92.23	99.11	94.73	**99.44**
Kappa	90.63	93.50	93.82	98.36	85.95	98.21	89.02	**98.86**

As shown in Table 1, 3DCNN surpasses 2DCNN because it can extract spatial and spectral features simultaneous. SSUN, which uses LSTM and CNN to extract spectral and spatial features respectively, achieves good classification results. Since SpectralFormer only focuses on spectral information, the accuracy is not ideal. CESSUT achieves the highest OA, AA and Kappa among all methods, because it combines the advantages of transformer extracting global information and convolution extracting local information. Besides, CESSUT is superior to separate spatial branch and spectral branch, which proves that it achieves effective fusion of spatial and spectral information. When scaled to fewer training samples such as 1%, our method still achieves the best classification results, OA, AA and Kappa are 95.08%, 92.17%, 93.46%, which are at least 1.58%, 0.44%, 2.10% higher than other methods. In SV dataset, CESSUT also performed best on the three evaluation indicators.

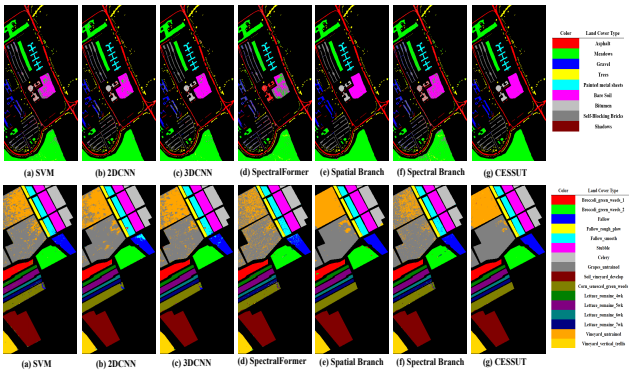

Fig. 4. The classification maps of two datasets.

In the comparison of classification maps, it is obvious that CESSUT has the least misclassified points. In addition, CESSUT performs better in some details such as edges and textures. This also reflects the effectiveness of the method from a qualitative point of view.

3.3. Ablation Studies

To evaluate the validity of two layers proposed, ablation experiments are conducted on UP dataset and results are presented in Table 3. The experimental results show that the introduction of the convolutional embedding(CE) layer and locally-enhanced feed-forward(LEFF) layer improves the accuracy to varying degrees, no matter in the spatial branch or spectral branch, and then the effect of unified network will also be improved. Best results are obtained when two layers are applied at the same time. This shows that the use of two layers brings locality to transformer model, thereby obtaining state-of-the-art performances.

Table 3. Ablation Studies on CE and LEFF.

CE	LEFF	Spatial Branch(OA)	Spectral Branch(OA)	CESSUT(OA)
		95.61	85.96	97.22
√		97.89	86.86	98.22
	√	97.34	88.02	97.64
√	√	**98.63**	**88.94**	**99.13**

4. CONCLUSIONS

This paper proposes a convolution enhanced spatial-spectral unified transformer network for hyperspectral image classification. Firstly, a transformer-based network which consists of spatial branch and spectral branch is proposed to extract global spatial and spectral features. Secondly, spatial and spectral convolutional embedding layers are designed to generate embeddings of spatial patches and spectral bands as the input of transformer. Thirdly, locally-enhanced feed-forward layer employs depth-wise convolution to introduce locality into the network to enhance representation ability of features. The experimental results prove that the superiority of the proposed method, and the ablation experiment verifies the effectiveness of the two layers.

5. REFERENCES

[1] 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.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[5] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.

[6] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu, "Incorporating convolution designs into visual transformers," *arXiv preprint arXiv:2103.11816*, 2021.