HYPERSPECTRAL IMAGE CLASSIFICATION BASED ON CO-LEARNING THROUGH DUAL-ARCHITECTURE ENSEMBLE

Chen Xiaoyue, Cao Xianghai

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

ABSTRACT

Hyperspectral image classification is a classic topic aiming to assign the category label of each pixel in hyperspectral images. Some deep learning methods have been introduced and achieved good results, such as the CNN-based architecture, which focuses on local and hierarchical feature extraction to obtain visual information from shallow to deep. Recently, Transformer has been applied to the visual field and also used in the hyperspectral image classification task. Some work applied Transformer to process the spectral information but cannot achieve good results. To optimize the results, a new strategy called co-learning is proposed through a dual-architecture ensemble. The samples selected by the dual-architecture network are iteratively added to increase more reliable training samples. CNN and Transformer use completely different methods to extract features from different views and have great diversity. Experimental results show that this method is better than the algorithm using only a single network.

Index Terms— Co-learning, convolutional neural network (CNN), Transformer, dual-architecture ensemble

1. INTRODUCTION

The hyperspectral image (HSI) is a remote sensing image rich in spectral information and can provide effective information for ground object recognition, geological survey, vegetation cover detection, and other fields. The hyperspectral image classification (HSIC) task is essential to practical use, so the accuracy is important. There have been some deep learning methods used for the HSIC task, for example, DBN [1], CNN [2][3], LSTM [4], RNN [5], etc. Some works combine CNN with the attention module to achieve better results [6][7].

The CNN can extract spatial structure information and local context information from HSI, which has performed well in HSIC. However, as a kind of data with hundreds of spectra, the implicit sequence attribute of HSI can not be ignored. There are many categories in an HSI, and the spectral similarity of each category is very high. Because of its structural design, the CNN is difficult to capture sequence attributes, especially medium and long-term dependencies. As a result, a very important part is missing in the feature extraction stage.

The Transformer is good at processing sequence information and can extract medium or long-distance dependencies with global attributes. This feature is complementary to the spatial structure information extracted by CNN. Transformer [8] uses the multi-head attention (MHA) mechanism to get good results in the image classification task because it can extract serialized information. However, Transformer achieves inferior performance to CNNs when trained on an HSI dataset. There are several reasons: first, the simple tokenization of input images fails to utilize spatial information such as edges and lines among neighboring pixels, leading to low training sample efficiency. Second, the insufficient training samples are not enough for the self-attention design of the Transformer, leading to unsatisfactory classification results. Compared with Transformer, CNN can capture local context or semantic features. In addition, it can extract inductive bias, which makes feature extraction easier.

We propose a classification method based on a co-learning strategy through the dual-architecture ensemble. Two networks with completely different structures and principles are selected to increase the diversity of the samples. According to the theory of ensemble learning [9]: the greater the diversity of features, the easier it is to increase the accuracy and avoid the overfitting problem. Co-learning can provide complementary information for two different networks. The specific idea is to fuse the spatial feature from CNN and the spectral feature from Transformer through co-learning in each iteration so that the training samples can be increased for two different methods at the same time. The accuracy, generality, and robustness of classification results can be greatly improved. The principle of the co-learning strategy will be introduced in section 3.

The contribution of this paper can be described below:

1) A dual-architecture ensemble method is proposed, combining the traditional CNN structure and the Vision Transformer structure.
2) A co-learning strategy is proposed and applied to the HSIC task, largely improving the classification accuracy.
3) The classification result of the proposed method shows that the co-learning strategy brings obvious

improvement compared with the single-architecture method. The accuracy is shown in the experiment section.

2. RELATED WORK

2.1. CNN

CNN is the most widely used deep learning network, which can achieve good results in the field of image processing and interpretation. The researches based on CNN have developed rapidly in the past few years and have made great achievements in subfields such as target recognition, image classification, and image segmentation. There are many CNN structures applied to HSI classification tasks and some of them are exquisite and effective. Most importantly, based on optimization experience, CNN has been very mature in both training tricks and network architecture construction [11][12].

2.2. Vision Transformer

Recently, Vision Transformer (ViT) has been increasingly used in visual processing tasks since the work in [8]. ViT attempts to apply the standard Transformer directly to image processing. Some transformer-based methods to solve the task of HSIC have been studied in [13]-[15]. The benefit of ViT is that it can handle the sequential data and model the long-range dependencies. The illustration of the Transformer Encoder is shown in Fig. 1. The original size of the input image is $H \times W \times C$. Assuming that the length and width of each patch are (P, P), then the number of patches would be $N=H \times W/(P \times P)$, and the size of each patch is $P \times P \times C$. Then a patch embedding operation is applied. The positional encoding is introduced to add the location information of the sequence, then the position embedding can be applied. L in the transformer encoder represents L layers in the whole network. Layers are the number of times that encoder blocks are stacked repeatedly in the transformer encoder. Norm refers to the LayerNorm operation and MHA stands for multi-head attention, which is an important part. Multi-head attention allows the model to focus on the information on different representation subspaces from different locations. After the transformer encoder, the classification goal is implemented by adding LayerNorm and fully connected layers, using the GELU activation function.

3. CO-LEARNING STRATEGY AND DUAL-ARCHITECTURE ENSEMBLE FOR HSIC

In this section, the co-learning strategy used in this paper is introduced, along with the dual-architecture ensemble method. We explained our method with the illustration of the architecture (Fig.2.) and the description of the algorithm.

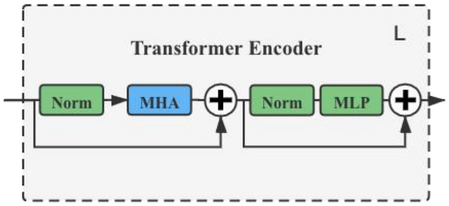


Fig. 1. Illustration of the Transformer Encoder. L means L layers, that is, L transformer encoders in the overall ViT network.

The dual-architecture ensemble method utilizes two different network structures (CNN and Transformer) to select useful test samples, and combine them with the training samples in the other network by iterative operations to improve the accuracy. This method ensures that the samples with high reliability in one network are added to the other network as a supplement with information from a different view. The dual-architecture in this paper was inspired by a method called co-training, which was proposed in [28] for the classification of web pages. The co-training strategy assumes the existence of two separate views of the input feature which are conditionally independent.

The co-learning strategy is used in the dual-architecture ensemble method to add the diversity of training samples

Algorithm 1. The co-learning strategy in the dual-architecture ensemble method for HSI Classification

```
Network:
    CNN & Transformer (with different parameters w1
    & w2)
Parameter setting:
    D1: initial training set for CNN
    D2: initial training set for Transformer (D2 = D1)
    D1t & D2t: test set for CNN & Transformer
    D1_top & D2_top: the first n% samples, considering
    the reliability (n=2-10)
    N: iteration number
Procedure:
for m = 1, 2, . . ., N:
    Train CNN with D1 and train Transformer with D2;
    With trained CNN, classify the test set D1t and
    select n% reliable samples (D1_top), add them into
    the training set D2;
    With trained Transformer, classify the test set D2t
    and select n% reliable samples (D2_top), add them
    into the training set D1;
    Get updated w1 & w2;
end
Output the classification results of CNN and
Transformer by fully connected layer.
```

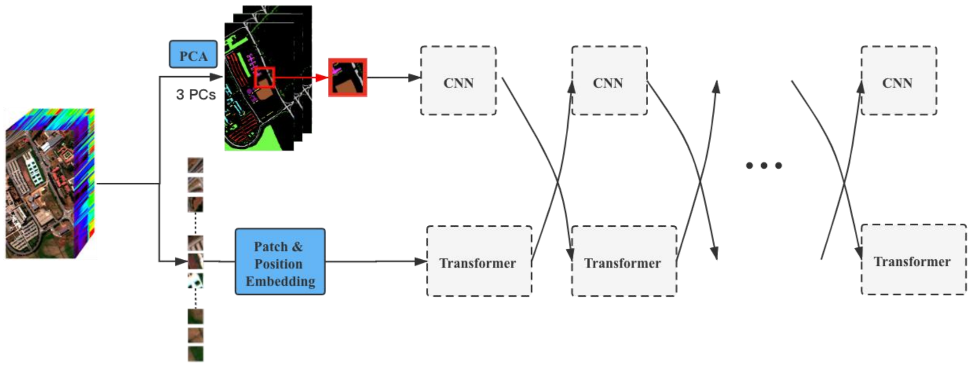


Fig. 2. Illustration of the dual-architecture ensemble method with co-learning strategy.

between two classifiers. The description of the co-learning strategy is listed in Algorithm 1. In the dual-architecture ensemble method, the overall procedure contains several iterations. The co-learning strategy means that in each iteration, we get some samples with high reliability and add them into the other network. In several iterations, the reliable test samples will be used between two structures. The diversity and the number of samples will naturally increase. Samples full of the spatial feature extracted from the CNN structure are combined with samples full of the spectral feature extracted from Vision Transformer to ensure that the performance can be greatly improved by the dual-architecture ensemble method even when the amount of training samples is insufficient. Another advantage is that the dual-architecture ensemble method is reasonable and easy to design. The structure can be the combination of CNNs or Vision Transformers, or the CNN and the Vision Transformer. It only takes a few to a dozen iterations to see that the result is significantly improved compared with the classification results trained by a single network structure.

In general, in the dual-architecture method, one network can utilize samples of high reliability from the other network. By utilizing the high-reliability training samples, the diversity of the training sets and the number of the training samples increase so that the results of classification will be better. The process is iterated about 20 times in our experiments, showing great improvement in the accuracy of the HSIC task, compared with using a single network. Parameter Settings are explained in the next section.

4. EXPERIMENTS AND RESULTS

In this section, we apply a dual-architecture ensemble method with the co-learning strategy for the HSIC task. The result of the proposed method is compared with the results from single networks and some classic methods. To ensure the quality of training data and further improve the accuracy, a data augmentation strategy is used. Compared with CNN, Vision Transformer, and other existing methods, the proposed method shows greater effectiveness.

Table 1. The network setting of CNN in different datasets

Dataset	Layer	Convolution kernel	Relu	Pooling
Pavia University	1	4×4×8	√	2×2
	2	5×5×28	√	2×2
	3	3×3×36	√	No
Indian Pines	1	4×4×24	√	2×2
	2	5×5×48	√	2×2
	3	3×3×128	√	No

4.1. Dataset

The methods were evaluated on Indian Pines (IP) and Pavia University (PU). The size of IP is 145 × 145 (pixels) with 220 bands. IP contains 16 classes with 10249 pixels. The size of the PU is 610 × 340 (pixels) containing 103 spectral bands. It contains 9 classes with 42776 labeled pixels. The false-color composite image and the ground-truth image of Pavia University are shown in Fig. 3.

4.2. Implement detail

For each dataset, the training data is selected randomly in the percentage of 10% from the whole dataset, while the rest are preserved for testing. 30 principal components are selected using PCA to remove the spectral redundancy. The input size is 7 × 7 × 30 for IP and 9 × 9 × 30 for PU. The data augmentation method is adopted and the number of the generated virtual samples is three times the original ones. This will improve the accuracy. The network architecture setting of CNN in different datasets is shown in Table 1. As for the ViT model, the patch size is 4 and the number of heads in MHA is 16, with 5 transformer encoders. Overall accuracy (OA), average accuracy (AA), and Kappa coefficient were measured and the results are in Table 2. OA means the percentage of correctly classified pixels. AA is the average value of the percentage of correctly classified pixels for each class. Kappa coefficient is the ratio of error reduction to completely random classification. The final

Table 2. Results of classification in different methods. 'Cl' means the dual-architecture ensemble method with the co-learning strategy (20 iterations). 'Cl_Aug' means Cl with the data augmentation strategy.

Dataset	Metric	SVM	EMP [29]	EPF [30]	MSTV [22]	CRNN [23]	ViT [8]	CNN	Cl	Cl_Aug
Pavia University	OA (%)	82.73	88.45	91.94	88.82	91.60	76.99	85.33	92.88	95.68
	AA (%)	83.67	87.43	93.76	90.06	92.28	80.22	88.26	93.28	96.15
	Kappa	0.82	0.87	0.93	0.87	0.91	0.70	0.83	0.92	0.96
Indian Pines	OA (%)	80.02	88.02	93.04	86.43	92.32	71.86	77.23	91.46	93.19
	AA (%)	75.21	86.16	92.13	88.64	93.64	78.97	83.16	92.14	94.40
	Kappa	0.77	0.86	0.9	0.84	0.91	0.68	0.76	0.91	0.93

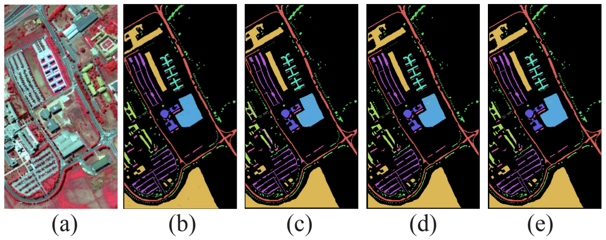


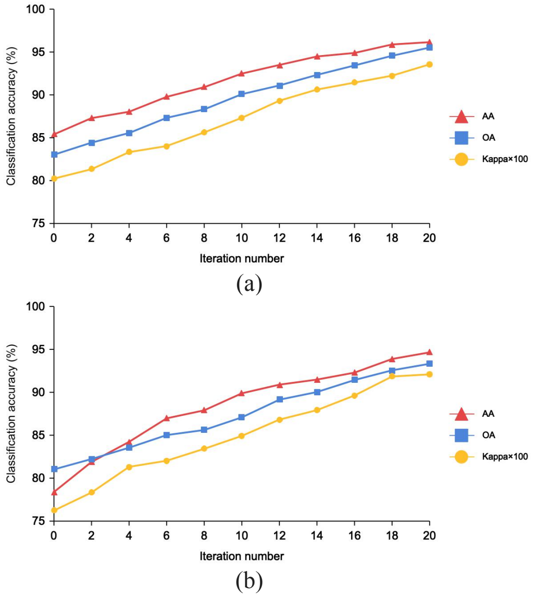
Fig. 3. Illustration of Pavia University and classification results through co-learning strategy. (a) False-color composite image; (b) Groundtruth; (c)-(e) Results after 10, 15 and 20 iterations.

number of iterations is set to 20 for the co-learning strategy. The data augmentation strategy is used to expand the data set by generating new samples by linearly combining two existing samples from the same class.

Five representative classification methods are listed for comparison, which are reproduced with the same setting. SVM is applied as a traditional classifier. EMP [29] utilizes the mathematical morphology and EPF [30] introduces the edge-preserving filter. MSTV [22] extracts multi-scale features and applies KPCA. CRNN [23] considers the spectral signature as a sequence and uses cascade RNN. The original Transformer and CNN are considered as controlled groups to evaluate the effect.

4.3. Results analysis

The data from Table 2 shows that our proposed method can achieve better classification results and the results in PU are better than that in IP. Overall, the OA, AA, and Kappa with the proposed method improved compared with those with the single network. The best result is from co-learning with data augmentation strategy, its OA, AA, and Kappa are 95.68%, 96.15%, and 0.96 for PU and 93.19%, 94.40%, and 0.93 for IP. This result is 4.08%, 3.87%, and 0.05 in OA, AA, and Kappa higher than CRNN, which apply RNN in a cascade way, for the PU dataset. Owing to the powerful learning ability of the dual-architecture method, classic backbone networks like CNN or Transformer can extract reliable samples, add diversity for each other and improve accuracy.



(a)

(b)

Fig. 4. Classification accuracy in different iterations for two datasets. (a) Pavia University. (b) Indian Pines.

The results in Fig. 4 show the trend of classification accuracy in different iteration numbers in the co-learning strategy. With the increase of iteration numbers, the classification accuracy is gradually improved from 82.68% to 95.68% in OA and from 85.06% to 96.15% in AA for PU, from 81.02% to 93.19% in OA and from 78.39% to 94.40% in AA for IP.

5. CONCLUSION

Transformer, as a network different from CNN structure, has been explored. However, in practical applications, a single Transformer network cannot achieve satisfactory results. We consider building a dual-architecture ensemble structure, using the co-learning strategy to add training samples extracted from the two networks to the next iteration. Experiments have proved that this method is of great significance to improve the classification effect. In future research, more ways to merge the two networks should be explored.

6. REFERENCES

[1] G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[2] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, “Deep supervised learning for hyperspectral data classification through convolutional neural networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4959–4962.

[3] Y. Li, H. Zhang, and Q. Shen, “Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network,” *Remote Sens*., vol. 9, no. 1, p. 67, 2017.

[4] Q. Liu, F. Zhou, R. Hang, and X. Yuan, “Bidirectional-convolutional lstm based spectral-spatial feature learning for hyperspectral image classification,” *Remote Sens*., vol. 9, no. 12, p. 1330, 2017.

[5] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, “Cascaded recurrent neural networks for hyperspectral image classification,” IEEE Trans. Geosci. *Remote Sens*., vol. 57, no. 8, pp. 5384–5394, 2019.

[6] R Hang, et al, “Hyperspectral Image Classification with Attention Aided CNNs,” *IEEE Transactions on Geoscience and Remote Sensing* PP.99(2020):1-13.

[7] Zhang, Xiangdong, T. Wang, and Y. Yang, “Hyperspectral image classification based on multi-scale residual network with attention mechanism,” *Geosciences Journal* (2020).

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.

[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging, boosting, and hybrid-based approaches,” *IEEE Trans. Syst*., Man, Cybern. C, Appl. Rev., vol. 42, no. 4, pp. 463–484, Jul. 2012.

[10] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, “Random forest ensembles and extended multiextinction profiles for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, Jan. 2018.

[11] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE CVPR*, Boston, MA, USA, 2015, pp. 1–9.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR*, 2015, pp. 1–14.

[13] Hong D, Han Z, Yao J, et al, “SpectralFormer: rethinking hyperspectral image classification with transformers[J],” 2021.

[14] Hu X, Yang W, Wen H, et al, “A lightweight 1-D convolution augmented transformer with metric learning for hyperspectral image classification[J],” *Sensors*, 2021, 21(5):1751.

[15] He Xin, Chen Yushi, Lin Zhouhan, “Spatial-spectral transformer for hyperspectral image classification,” *Remote Sensing*. 13.10.3390/rs13030498, 2021.

[16] R. O. Duda and P. E. Hart, “Pattern classification and scene analysis,” *Wiley New York*. 1973, vol. 3.

[17] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, “Classification of hyperspectral images with regularized linear discriminant analysis,” *IEEE Trans. Geos. Remote Sens*., vol. 47, no. 3, pp. 862–873, 2009.

[18] M. Sugiyama, “Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis,” *J. Mach. Learn. Res*., vol. 8, no. May, pp. 1027–1061, 2007.

[19] C. Lee and D. Landgrebe, “Feature extraction based on decision boundaries,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 4, pp. 388–400, 1993.

[20] W. Zhao and S. Du, “Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach,” *IEEE Trans. Geosci. Remote Sens*., vol. 54, no. 8, pp. 4544– 4554, 2016.

[21] Y. Chen, H. Jiang, C. Li, X. Jia, and P. 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.

[22] P. Duan, X. Kang, S. Li, and P. Ghamisi, “Noise-robust hyperspectral image classification via multi-scale total variation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 12, no. 6, pp. 1948–1962, Jun. 2019.

[23] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, “Cascaded recurrent neural networks for hyperspectral image classification,” I*EEE Trans. Geosci. Remote Sens*., vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[24] Velasco-Forero, S., and V.Manian, “Improving Hyperspectral Image Classification Using Spatial Preprocessing,” *IEEE Geoscience & Remote Sensing Letters* 6.2(2009):297-301.

[25] Li, W., et al, “Data Augmentation for Hyperspectral Image Classification With Deep CNN,” *IEEE Geoscience and Remote Sensing Letters* (2018):1-5.

[26] Huang, L., and Y. Chen, “Dual-Path Siamese CNN for Hyperspectral Image Classification With Limited Training Samples,” *IEEE Geoscience and Remote Sensing Letters* PP.99:1-5.

[27] Xiao, Z., J. Jiang, and C. Ni, “Spectral-Spatial Classification of Hyperspectral Image Based on Self-Adaptive Deep Residual 3D Convolutional Neural Network,” *Journal of Computer-Aided Design & Computer Graphics* (2019).

[28] Samiappan, S., and R. J. Moorhead, “Semi-supervised co-training and active learning framework for hyperspectral image classification,” *IEEE IEEE*, 2015.

[29] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, “Classification of hyperspectral data from urban areas based on extended morphological profiles,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[30] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, “PCA-based edge preserving features for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens*., vol. 55, no. 12, pp. 7140–7151, Dec. 2017.