A SCALABLE PRETRAINED MODEL FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-based Hyperspectral image classification methods have achieved fantastic performance due to their superior ability to represent features at the cost that these methods are complex, inflexible and weak to generalize. Thus, we propose a simple and scalable pretrained model which can greatly accelerate convergence rate and improve the performance in downstream classification task. Experiments shows the performance of our method achieved state-of-the-art performance with limited training samples.

1 INTRODUCTION

Hyperspectral image(HSI) is a kind of image differs greatly from RGB image, which has hundreds of channels and contains rich spectral information. Analyzing HSI is generally done at the pixel level by sample patches. With the convolution network, the structure and depth of network are limited by the size of sampled patch, and weak to generalize. Recently, some studies used transformer-based models has achieved excellent performance in HSI classification tasks(He et al. (2019),Sun et al. (2022)). Due to the scarcity and complexity of HSI data, there is still no work on pretrained models. Inspired by BEiT(Bao et al. (2021)) and MAE(He et al. (2022)) which have achieved SOTA performance in computer vision. We designed a self-supervised leaner for HSI, which can generalize to inputs of arbitrary size and spectral resolution and accelerate the downstream training process, as shown in Figure 1.

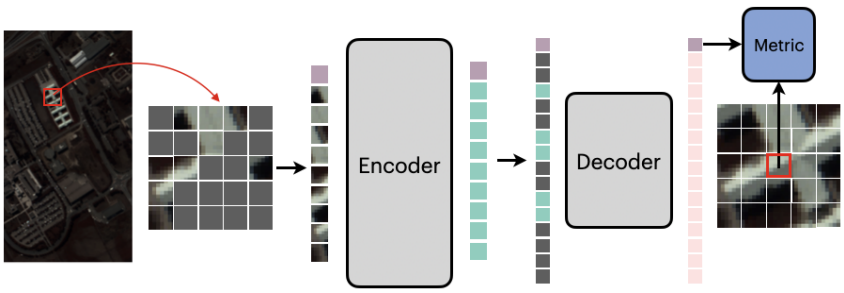

Figure 1: The structure of self-supervised leaner (Masked on pixelwise)

2 METHODOLOGY

In our method, we augment domain knowledge by adding an auxiliary task to the model, to overcome the defect of transformer-based model which lack inductive bias(Vaswani et al. (2017)). Specifically, we flatten the patch $\boldsymbol{X} \in \mathbb{R}^{n\times n\times c}$ and contact a random token $\boldsymbol{x}_t$ into $\mathbb{X} = \{\boldsymbol{x}_t, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n^2} | \boldsymbol{x} \in \mathbb{R}^c\}$, where n is the size of the patch, c is the channel of hyperspectral image. The loss function of the pretraining section is

$$l = \frac{1}{n^2}\sum_{i=1}^{n^2} ||\boldsymbol{x}_i - \boldsymbol{x}_i^{'}|| + ||\boldsymbol{W}_1\boldsymbol{x}_t^{'} - \boldsymbol{W}_2\boldsymbol{x}_c|| \tag{1}$$

where $\boldsymbol{x}_c$ represent the center pixel of patch, $\boldsymbol{x}_t^{'}$ represent the output of random token,$\boldsymbol{x}_i^{'}$ represent the reconstruction of each pixel in patch, $\boldsymbol{W}_1$, $\boldsymbol{W}_1$represent the linear transform.

In downstream task, in order to make full use of the information learned by the network and domain knowledge,the tokens output from the encoder are combined in a learnable combination and the output is fed to the classifier as the final logit for training. We donote the output of encoder is $\mathbb{Z} = \{\boldsymbol{z}_t, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, , \boldsymbol{z}_n^2 \,|\boldsymbol{z} \in \mathbb{R}^d\}$, the final logit $\boldsymbol{v}$ can compute as follow equations:

$$\boldsymbol{M} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n^2}]^T, \boldsymbol{M} \in \mathbb{R}^{n^2 \times d} \tag{2}$$

$$\boldsymbol{M}^{'} = [f(\boldsymbol{z}_1), f(\boldsymbol{z}_2), \ldots, f(\boldsymbol{z}_{n^2})]^T \tag{3}$$

$$\boldsymbol{v} = \boldsymbol{M}^T \boldsymbol{M}^{'} g(\boldsymbol{z}_t) + \boldsymbol{z}_t, \boldsymbol{v} \in \mathbb{R}^d \tag{4}$$

where f and g represent MLP mapping, d represent the embedding dimension of encoder.

3 EXPERIMENTAL

Dataset. We select hyperspectral images from various scenes which captured by GF-5 satellite with 297 channels to make pretraining dataset, and divide them into patches of 4 different sizes, we use 90% to train, 10% to test. We fine-tuning the pretrained model on Salinas(SA), Indian Pines(IP) and PaviaU(PU). We use we use 10% of the samples to train, 90% to test.

Hyper-parameters. We use MAE as the baseline, the mask ratio, embedding dimension, depth, heads of encoder are set to 0.5, 256, 4, 8 respectively, the decoder is half of it. We use AdamW as optimizer, the learning rate is set to 8×10^{-4} in pretrain stage. In downstream task, the learning rate of encoder is set to 10^{-5}, of classifier is set to 10^{-3}.

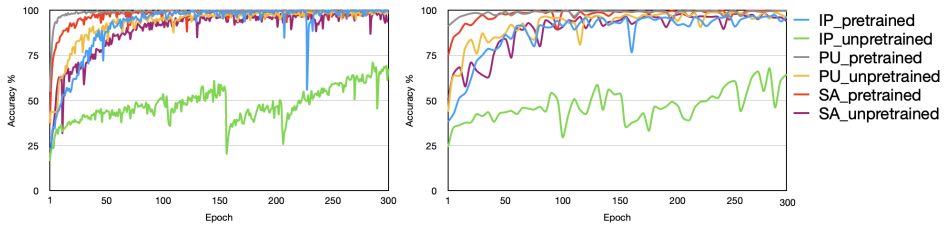

Figure 2: Accuracy curve in training process(Left figure is the accuracy curve on training set, right figure is the accuracy curve on testing set)

Quantitative Evaluation. The average PSNR and SSIM of the reconstructions is about 48dB and 0.98. In downstream task, Figure 2 shows the curve of accuracy on pretrained model and random initialization model. It indicates that our method can greatly improve the performance and speed up the convergence rate especially when the training data is relatively small. In addition, our method have achieved SOTA performance on above datasets with different scenes, spatial resolutions and spectral resolutions. That means our method can effectively extract the general feature of HSI and has strong generalization ability. We used OA, AA, Kappa to evaluate the classification performance and compared our method with other SOAT methods as shown in Table 1 in the appendix.

Influence of auxiliary task and learnable combination. We employ an auxiliary task to overcome the problem of large data requirements due to the lack of inductive bias in tansformer-based model. In the ablation experiments, with auxiliary task and learnable combination the classification performance has improved $2\% \sim 3\%$. Consequently, we consider by adding appropriate auxiliary tasks, it can supplement domain knowledge in some data scarce fields, thus reducing the amount of data required 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

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Ji He, Lina Zhao, Hongwei Yang, Mengmeng Zhang, and Wei Li. Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):165–178, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

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

Yuhao Qing, Quanzhen Huang, Liuyan Feng, Yueyan Qi, and Wenyi Liu. Multiscale feature fusion network incorporating 3d self-attention for hyperspectral image classification. *Remote Sensing*, 14(3):742, 2022.

Hao Sun, Xiangtao Zheng, Xiaoqiang Lu, and Siyuan Wu. Spectral–spatial attention network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (5):3232–3245, 2019.

Le Sun, Guangrui Zhao, Yuhui Zheng, and Zebin Wu. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A APPENDIX

We campare our method with SSAN(Sun et al., 2019), 3DSA-MFN(Qing et al., 2022), SST-FA(He et al., 2021).

Table 1: Comparison of other methods

Method	Salinas			Indian Pains			PaviaU		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
SST-FA	94.94	93.05	94.32	88.98	68.15	86.7	93.37	85.01	97.49
SSAN	96.81	98.33	96.54	95.49	94.17	94.85	98.02	96.90	97.37
3DSA-MFN	99.72	99.32	99.13	96.02	93.89	94.78	98.96	96.32	97.49
Ours	99.85	99.73	99.75	96.55	93.12	96.10	99.73	99.69	99.46