# HIGH-ORDER SELF-ATTENTION NETWORK FOR REMOTE SENSING SCENE CLASSIFICATION

*Nanjun He*[1], *Leyuan Fang*[1], *Yi Li*[2], *Antonio Plaza*[3]

[1]College of Electrical and Information Engineering, Hunan University, China
[2]School of Design, Hunan University, China
[3]Hyperspectral Computing Laboratory, University of Extremadura, Caceres, Spain

## ABSTRACT

Convolutional neural networks (CNNs) have recently shown remarkable performance in remote sensing scene image classification. However, long-range dependencies (e.g. non-local similarities) within the scene are often ignored by CNNs. To address this issue, in this paper we develop a new high-order self-attention network (HoSA) for remote sensing scene classification. Specifically, we embed two novel modules, i.e., a self-attention module and a high order pooling module, into off-the-shelf CNN models and then fine-tune the whole network. The advantages of our newly proposed HoSA network are twofold. Firstly, with the self-attention module, the HoSA can capture long-range dependencies within the scenes for high-level semantic feature extraction. Secondly, by means of its high-order pooling mechanism, our newly developed HoSA can further explore high-order information contained in the features. Our experiments with a widely used remote sensing scene data set demonstrate that the proposed HoSA network exhibits better classification performance than the baseline and several well-known methods.

***Index Terms***— Remote sensing scene classification, convolutional neural networks (CNNs), self-attention, high-order pooling.

## 1. INTRODUCTION

Recent advances in remote sensing instruments have allowed for the collection of a considerable number of high resolution remote sensing scenes. It is now critical to perform content-based scene retrieval from large databases for different applications, such as urban mapping or land use classification [1]. Recently, convolutional neural network (CNN) have demonstrated remarkable performance in remote sensing scene classification. Many CNN-based methods have been developed for this purpose, and we can broadly categorize available approaches in two categories:

*1) Pretrained CNN-based feature extraction methods.* These approaches first use the CNN model that pretrained on the Imagenet for feature extraction purposes, and then feed the obtained features into a classifier, e.g., the support vector machine (SVM). In [2], Hu *et al.* investigate different CNN models as feature extractors and integrate them with various feature encoding methods. Their results show that, the CNN model (as a feature extractor) often provides better performance than that obtained by hand-crafted feature-based methods. In [3], Gong *et al.* utilize the bag of visual words (BoVW) model to aggregate the convolutional activation layers. In [4], the last two fully connected layers of a CNN model are combined together to represent the image. In [5], He *et al.* adopt a simple yet effective method (i.e., covariance pooling) to combine the different layers of pretrained CNN models for remote sensing scene classification.

*2) Pretrained CNN-based fine-tuned methods.* These methods generally use a transfer learning scheme to transfer the knowledge learned on Imagenet for fine-tuning purposes. In [6], Castelluccio *et al.* fine-tune two classical prerained CNN models (i.e., Caffenet and GoogLenet) for remote sensing scene classification. In [7], Cheng *et al.* add a new term into the loss function of the aforementioned prerained CNN model to minimize the intra-class distance and maximize the inter-class distance, thus improving the classification performance.

Although the aforementioned methods exhibit promising classification performance, long-range dependencies (e.g., the non-local similarities) within the scenes are often ignored, since the common CNN model generally exploits local connections only for feature extraction. To address this problem, in this paper we introduce a new high-order self-attention network (HoSA) for remote sensing scene classification that embeds a self-attention module and a high-order pooling module into the prerained CNN model. Specifically, we embed the self-attention module in the middle part of the CNN to make use of long-range dependencies within the image. Then, we

This work was supported by the National Natural Science Fund of China for International Cooperation and Exchanges under Grant 61520106001, the Science and Technology Talents Program of Hunan Association for Science and Technology under Grant 2017TJ-Q09, the Fund of Hunan Province for the Science and Technology Plan Project under Grant 2017RS3024, the Science and Technology Plan Project Fund of Hunan Province under Grant CX2018B171,and the Science and Technology Innovation Platform and Talents Program of Hunan Province under Grant 2018TP1013.

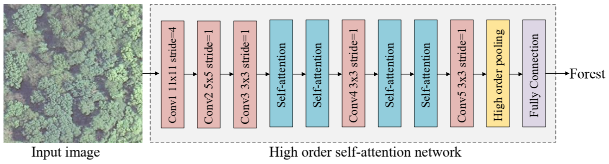

**Fig. 1**. Architecture of the proposed HoSA network. Alexnet is used as the baseline network. For the sake of simplicity, the ReLU, max-pooling and softmax layers are omitted.

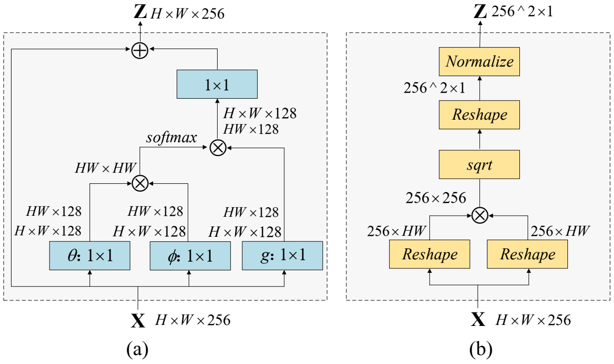

**Fig. 2**. Graphical illustration of the two novel modules used in this paper. (a) Self-attention module, where $\mathbf{X}$ is the input feature map with $H \times W \times 256$ dimensions, $\mathbf{Z}$ is the output feature map, with $H \times W \times 256$ dimensions. The blue squares represent the convolutional kernel. $\bigotimes$ denotes the matrix multiplication operation, and $\bigoplus$ denotes an element-wise sum. The *softmax* operation is performed on each row. (b) High-order pooling module, where $\mathbf{X}$ is the input feature map with $H \times W \times 256$ dimensions, $\mathbf{Z}$ is the output feature map, with $256^2 \times 1$ dimensions. The *sqrt* operation is conducted on each element. $\bigotimes$ denotes the matrix multiplication operation.

append a high-order pooling module on top of the last convolutional layers of the CNN. Finally, we fine-tune the whole system for classification purposes.

The remainder of this paper is organized as follows. Section II details the architecture of our newly proposed HoSA network. In section III, several experiments are conducted on a widely used remote sensing scene data set to verify the effectiveness of the proposed approach HoSA. Finally, section IV concludes the paper with some remarks and hints at plausible future research lines.

## 2. PROPOSED METHOD

Fig. 1 shows a graphical description of the proposed HoSA network. The classic Alexnet is used as the baseline network. Four self-attention modules are inserted into the Alexnet (two of them are inserted before *Conv4* and the rest of them are inserted before *Conv5*). The high-order pooling layer is added on the top of *Conv5* and followed by a fully connected layer. In the following, we detail the main two modules (i.e., self-attention and high-order pooling) of the HoSA architecture. Note that the Alexnet can be easily replaced by another deep network (e.g., the Resnet).

### 2.1. Self-Attention Module

A graphical illustration of our self-attention module is shown in Fig. 2(a). Before introducing our self-attention module, we first briefly describe the local connected convolutional operations (conv). Given a input feature $\mathbf{x}$, the conv is performed as shown below, where $\mathbf{y}$ is the output, *w* is a processing window, $|w|$ is the number of elements within the window *w*, and $W_{i,j} = \frac{1}{|w|}$.

$$\mathbf{y}_i = \sum_{j \in w} W_{i,j} \mathbf{x}_j. \tag{1}$$

Instead of considering only the local information into account, our self-attention module uses all the elements within the input to calculate the output. Moreover, the weight $W_{i,j}$ is not constant but a measure of the similarity measurement between the element $i$ and the element $j$. As a result, the self-attention module can capture the long-range dependencies within the input for feature learning purposes. The formulation of the self-attention module is given below:

$$\mathbf{y}_i = \sum_{j \in x} W_{i,j} \mathbf{x}_j, \tag{2}$$

where $W = softmax(\mathbf{x}^T W_\theta W_\phi \mathbf{x})$. Here, $W_\theta$ and $W_\phi$ are two affine matrices. In practice, $W_\theta$ and $W_\phi$ are implemented by a $1 \times 1$ convolutional operation on $\mathbf{x}$, denoted by $\theta(\cdot)$ and $\phi(\cdot)$, respectively. Moreover, in order to reduce the number of parameters, the input $\mathbf{x}$ is also processed by a $1 \times 1$ convolutional operation, $g(\cdot)$. With these considerations in mind, the practical formulation of the self-attention module is given below:

$$\mathbf{z} = softmax(\mathbf{x}^T W_\theta W_\phi \mathbf{x}) g(\mathbf{x}) + \mathbf{x}, \tag{3}$$

where "$+\mathbf{x}$" is a residual connection, which makes sure that the newly added self-attention module does not break the original behavior of the prerained CNN. For additional details, please refer to [8, 9]

### 2.2. High-Order Pooling Module

A graphical illustration of our high-order pooling module is shown in Fig. 2(b). Different from the max- or average-pooling strategies, which only take the first order information into consideration, our high-order pooling module can exploit the global high order information for feature aggregation. Hence, it is expected that the high-order pooling module can lean more discriminative features. Given an input

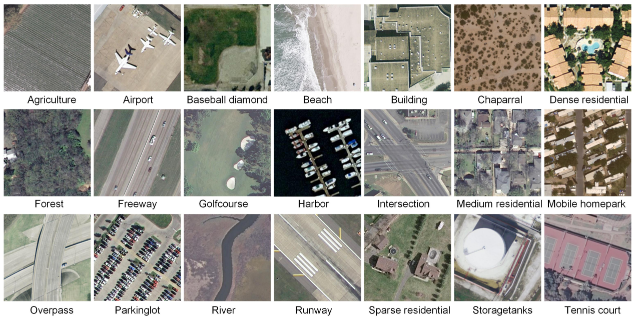

**Fig. 3**. Examples of the UC Merced land-use data set.

$\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, we first reshape $\mathbf{x}$ into $\hat{\mathbf{x}} \in \mathbb{R}^{HW\times C}$. Then, the $\hat{\mathbf{x}}$ is multiplied by its transpose. The obtained matrix is normalized by the $sqrt$ and column normalization operations, sequentially. Here, the $sqrt$ operation is conducted on each element of the input matrix.

## 3. EXPERIMENTAL RESULTS

### 3.1. UC Merced Land-Use Data Set

To evaluate the performance of the proposed method, we conducted our experiments on a widely used land-use data set i.e., the UC Merced data set [10]. This data set is composed of 21 land-use scene classes. Each class consists of 100 aerial images of $256 \times 256$ pixels, with a spatial resolution of 0.3 meters per pixel in the RGB color space. Some examples of this data set are shown in Fig. 3.

### 3.2. Implementation

Our newly developed HoSA network has been implemented using PyTorch, and the prerained Alexnet was downloaded from torchvision. A two-stage training strategy has been adopted. In the first stage, all the layers (except the last FC layer) are frozen, so that only the last FC is trained. The SGD optimizer is used for training, with learning rate of 1.0, momentum of 0.9, batch size of 64, and weight decay of 5e-4. In the second stage, we unfroze all the layers and set the learning rate of all the layers to be 1e-2, while keeping all other experimental settings identical to the ones adopted in stage one. All the input images are resized to $224 \times 224$ pixels before feeding them to our HoSA. The horizontal flip with probability equal to 0.5 is used for data augmentation purposes (no other data augmentation methods are adopted).

### 3.3. Comparison with Baseline and State-of-the Art Methods

To validate the performance of our newly developed HoSA framework, the following methods are used for comparison:

- First, the Alexnet (with fine-tuning strategy) is set as the baseline method, in which the last FC layers are re-initialized to match the number of classes. The SGD optimizer, with learning rate of 1e-3, momentum of 0.9, batch size of 64, and weight decay of 5e-4, is used for training. The horizontal flip method is also adopted for data augmentation purposes.
- Then, we compared the proposed method with the one in [7], in which the author added a new term into the loss function of the prerained Alexnet model to minimize the intra-class distance and maximize the inter-class distance.
- Next, we compared the proposed method with two prerained CNN feature extraction based methods, i.e., [4] and [5]. Both methods use the prerained Alexnet as the feature extractor.
- Finally, we compared the proposed method with two popular deep frameworks [6], i.e., CaffeNet and GoogleNet. These networks are trained from scratch and by means of a fine-tuning strategy, respectively.

In all our experiments, we randomly extract 80 samples for each class in the UC data set for training purposes, while the remaining samples are used for testing. Table 1 shows the average classification results obtained by all the tested methods after conducting five repeated experiments. From Table 1, we can observe that the HoSA method achieves better performance than the baseline fine-tuned Alexnet, with quantitative accuracy improvements of 1.4%. This is because of the following reasons. First, by inserting our self-attention module into the prerained Alexnet, the newly developed HoSA can capture non-local similarities and spatial layouts within the images for high-level semantic feature extraction. Second, by means of the high-order pooling method, the high-order information carried out by these features can be further exploited. In addition, the proposed HoSA method achieves higher classification accuracy than methods in [4–7]. For example, the classification accuracies obtained in the same context by [7] and [5] are 96.67% and 96.76%, respectively, while the HoSA achieves 97.29% accuracy. It is also worth noting that our method achieves higher classification accuracies than the fine-tuned GoogLenet, which comprises a much deeper architecture than the proposed HoSA.

### 3.4. Ablation Experiments

In this part, in order to demonstrate the effectiveness of the self-attention module, we conduct the ablation experiment on the proposed HoSA. Specifically, we remove all the self-attention (SA) modules in the HoSA, while keep the high order pooling module. We term it by HoSA without SA. The HoSA without SA is implemented by the same configuration of HoSA, which is shown Section in 3.2. Fig. 4 shows the

**Table 1**. Average accuracy (%) after five repeated experiments conducted using different methods.

| Method | Alexnet | CaffeNet [6] |  | GoogleNet [6] |  | [7] | [5] | [4] | HoSA |
|---|---|---|---|---|---|---|---|---|---|
|  |  | From scratch | Fine-tuned | From scratch | Fine-tuned |  |  |  |  |
| Backbone | Alexnet | Alexnet | Alexnet | Alexnet | CaffeNet | CaffeNet | GoogleNet | GoogleNet | AlexNet |
| Accuracy | 95.90 | 96.67 | 96.76 | 96.90 | 85.71 | 95.48 | 92.86 | 97.10 | **97.29** |

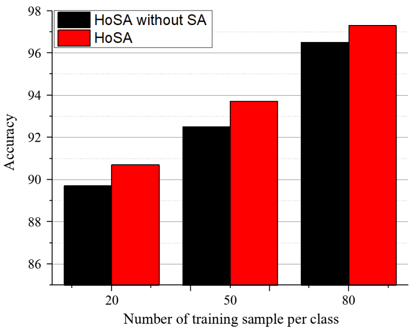

**Fig. 4**. Ablation experiments on our self-attention module using different numbers of training samples.

experimental results comparison between the HoSA and the HoSA without SA on different number of training samples. As can be seen, the HoSA can consistently outperform the HoSA without SA, which verifies that the self-attention module can indeed capture the nonlocal information, and thus improve the classification performance.

## 4. CONCLUSION AND FUTURE LINES

In this paper, we proposed a new CNN, called high-order self-attention (HoSA) network for remote sensing scene classification purposes. The proposed network introduces two new modules, i,e., the self-attention module and the high-order pooling module. With these two new modules, HoSA can not only exploit long-range dependencies within the scene for high-level semantic feature extraction, but also fully exploit the high-order information among the features for more discriminative feature learning. Our experiments, conducted on the well-known UC Merced land-use data set, demonstrated that the proposed HoSA network exhibits better classification performance than the baseline method, and also outperforms several state-of-the-art approaches for remote sensing scene classification. In the future, we combine deep CNN models and context-based encoding modules to further improve the classification performance.

## 5. REFERENCES

[1] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, “Extinction profiles fusion for hyperspectral images classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.

[2] Fan Hu, Gui-Song Xia, Jingwen Hu, and Liangpei Zhang, “Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery,” *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.

[3] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, “Remote sensing image scene classification using bag of convolutional features,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[4] S. Chaib, H. Liu, Y. Gu, and H. Yao, “Deep feature fusion for VHR remote sensing scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[5] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, “Remote sensing scene classification using multilayer stacked covariance pooling,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Jul. 2018.

[6] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, “Land use classification in remote sensing images by convolutional neural networks,” *arXiv*, vol. 1508, pp. 1–11, Aug. 2015.

[7] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[8] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. CVPR*, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. NIPS. 2017.*

[10] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proc. ICAGIS, 2010.*