ADAPTIVE SPATIAL-SCALE-AWARE DEEP CONVOLUTIONAL NEURAL NETWORK FOR HIGH-RESOLUTION REMOTE SENSING IMAGERY SCENE CLASSIFICATION

Wei Han[12], Ruyi Feng[12], Lizhe Wang[12] and Lang Gao[12]
[1]School of Computer Science, China University of Geosciences, Wuhan, 430074, P.R.China.
[2] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, P.R.China.

***Abstract*—High-resolution remote sensing (HRRS) scene classification plays an important role in numerous applications. During the past few decades, a lot of remarkable efforts have been made to develop various methods for HRRS scene classification. In this paper, focusing on the problems of complex context relationship and large differences of object scale in HRRS scene images, we propose a deep CNN-based scene classification method, which not only enables to enhance the ability of spatial representation, but adaptively recalibrates channel-wise feature responses to suppress useless feature channels. We evaluated the proposed method on a publicly large-scale dataset with several state-of-the-art convolutional neural network (CNN) models. The experimental results demonstrate that the proposed method is effective to extract high-level category features for HRRS scene classification.**

***Index Terms*—Weight-Adptive, Spatial Coding, Deep Learning, Convolutional Neural Networks (CNNs), High-Resolution Remote Sensing Scene Classification**

I. INTRODUCTION

Remote sensing image scene classification, which plays an important role in earth observation, categorizes scene images into an independent set of semantic-level land use and land cover (LULC) class labels according to image contents [1], [2]. During the past few decades, a lot of remarkable efforts have been made to develop various methods for the task of high resolution remote sensing (HRRS) image scene classification in a wide range of applications, such as LULC determination urban planning, environmental protection, and crop monitoring [3], [4], [5].

Deep-learning-based methods, which achieve significant many improvements over state-of-the-art records in many research fields, have been widely applied in natural image classification, object recognition, natural language, and text processing [6], [7], [8]. Due to their remarkable performance, these methods are used to analyze HRRS images, and have achieved more impressive results than the traditional shallow methods for scene classification [9], [10]. Deep features generated from deep convolutional neural networks (CNNs) have proven to be efficient at the representation of high-level semantic information of object and context relationship in HRRS image.

HRRS scene images contain complex context relationship and large differences of object scale, differently from natural images. Widely used CNN models in remote sensing, including VGG [6], AlexNet [8] and ResNet [7], are based on a simple yet effective strategy of constructing very deep networks: s-tacking building blocks of the same shape. Therefore, this kind of models are not suitable to overcome the problems in HRRS scene images and extract semantic information. Meanwhile, the increment of the depth of deep learning model causes the problems of training difficulty and an incredible amount of parameters. The enhancement of spatial representation and feature dependencies can make the CNNs more efficient [11]. Some works have developed to enhance the capability of spatial coding [12], [13]. And it has been demonstrated to be more effective than going deeper [14]. Differently, SENet [11] focuses on feature channels and utilize a novel architectural unit to adaptively recalibrate channel-wise feature responses. Unlike previous works, we develop the scene classification method by utilizing a new CNN model, which enhances the ability of spatial representation and optimizes feature response weights. It is implemented by combining the squeeze and excitation (SE) module in SENet and the aggregated transformation block in ResNeXt, named SE_ResNeXt. The experimental results show that it is able to extract high-level semantic feature hidden in HRRS image and achieve good performance for large scale HRRS scene classification without pre-training knowledge on ImageNet [15].

The remainder of this paper is organized as follows: In Section 2, we introduce the details of the proposed scene classification method. The experiments and analysis are proposed in Section 3. Conclusions are drawn in Section 4. The acknowledgment section and related references are displayed at the end.

II. SCENE CLASSIFICATION MODEL

A. Enhancement of Spatial Representation

ResNeXt[14] is developed to enhance the spatial representation for the improvement of classification performance. The main innovation is the aggregated transformation block, which is a homogeneous, multi-brach architecture with a

978-1-5386-7150-4/18/$31.00 ©2018 IEEE

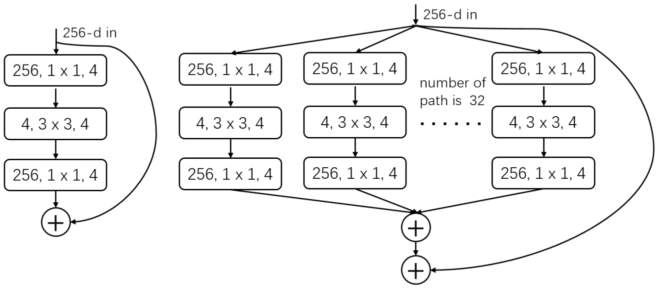


Fig. 1: Left: Illustration of residual block. Right: Illustration of aggregated transformation block combined with residual block.

few hyper-parameters. The special architecture exposes a new dimension, "cardinality", which means the size of the set of transformations and in addition to the dimensions of depth and width.

The simplest neurons in artificial neural networks perform inner product (weighted sum), which is the elementary transformation done by fully-connected and convolutional layers. The operation of aggregated transformation block is recast as a combination of splitting, transforming, and aggregating, as shown in Fig. 1. The aggregated transformation is formulated as:

$$f(x) = \sum_{i=0}^{c} T_i(x) \tag{1}$$

where $T_i(x)$ can be an arbitrary function. Analogous to a simple neuron, T_i should project x into an (optionally low-dimensional) embedding and then transform it. C is cardinality, meaning the size of the set of transformations to be aggregated.

In Fig. 1, the aggregated transformation in Eqn.2 serves as the residual function [7]:

$$y = x + \sum_{i=0}^{c} T_i(x) \tag{2}$$

where y is the output. The special architecture has a more effective ability of spatial representation feature and keeps the equivalent amount of parameters with residual block [14]. Additionally, increasing cardinality is proved to be more effective than increasing network depth.

B. Recalibration of Feature Response

Since the output of a convolutional operation is produced by a summation through all channels, the channel dependencies are implicitly embedded in weights of convolutional layers. But these dependencies are entangled with the spatial correlation captured by the filters. A special unit, squeeze and excitation (SE) module [11] is developed to ensure that the network is able to increase its sensitivity to informative features so that they can be exploited by subsequent transformations, and to suppress useless ones. It achieves the function by explicitly modeling channel inter dependencies and recalibrating filter responses in two steps: squeeze and excitation, before they are fed into next convolutional layer.

In the stage of squeeze, a channel descriptor is conducted by using global average pooling to generate channel-wise statistics compressing global spatial information. Formally, a statistic $z \in R^A$ is generated by shrinking U through spatial dimensions $W \times H$, where for the a_{th} channel response, the a_{th} element of z is calculated by:

$$z_a = F_{sq}(u_a) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_a(i,j). \tag{3}$$

Then, in order to capture the channel-wise dependencies, a gating mechanism with a sigmoid activation is

$$s_a = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{4}$$

where δ is the ReLU [16] function, $W_1 \in R^{(\frac{A}{r} \times A)}$ and $W_2 \in R^{(A \times \frac{A}{r})}$. A is the total number of feature channels in the current convolutional layer. To limit model complexity and aid generalization, the SE module is constructed by forming a bottleneck with two fully-connected (FC) layers around the non-linearity, i.e. a dimensionality-reduction layer with parameters W_1 with the reduction ratio r (the reduction ratio is set to be 16 as the original work [11]) , following by a ReLU and a dimensionality-increasing layer with parameters W_2. The final output of the block is obtained by rescaling the transformation output U with the activations:

$$\tilde{x_a} = F_{scale}(u_a, s_a) = s_a \cdot u_a \tag{5}$$

where $\tilde{X} = [\tilde{x_1}, \tilde{x_2}, \tilde{x_3}, ..., \tilde{x_A}]$ and $F_{scale}(u_a, s_a)$ refers to channel-wise multiplication between the feature map $u_a \in R^{(W \times H)}$ and the scalar s_a.

C. Descriptions of SE_ResNeXt

We utilize the neural network, SE_ResNeXt, to complete the task of HRRS scene classification, which is formed by a new neural unit. The unit is formed by an aggregated transformation block [14] and a following feature recalibration [11] for enhancing useful feature channels and suppressing useless feature channels. The SE-ResNeXt has fewer parameters than the early CNN models, including AleNeXt and VGG. For example, the parameter number of a SE-ResNeXt with 50 layers only account for 44% of AlexNet with 8-layer structure. Meanwhile, SE module causes 5% additional paramters compared with ResNeXt without SE block. The SE-ResNeXt has two main advantages:

1) This method is based on feature aggregation, which strengthens the spatial encoding ability. At the same time, the features are reactivated adaptively, which can significantly enhance the ability of feature representation, and is suitable for remote sensing images with complex various scales and complex relationships of scene image context.
2) Based on this module, it can be very deep and is very suitable for extracting high-level semantic information. Therefore, it can be used for scene classification on large-scale datasets. We tested its performance in a large scale scene dataset in the experimental section.

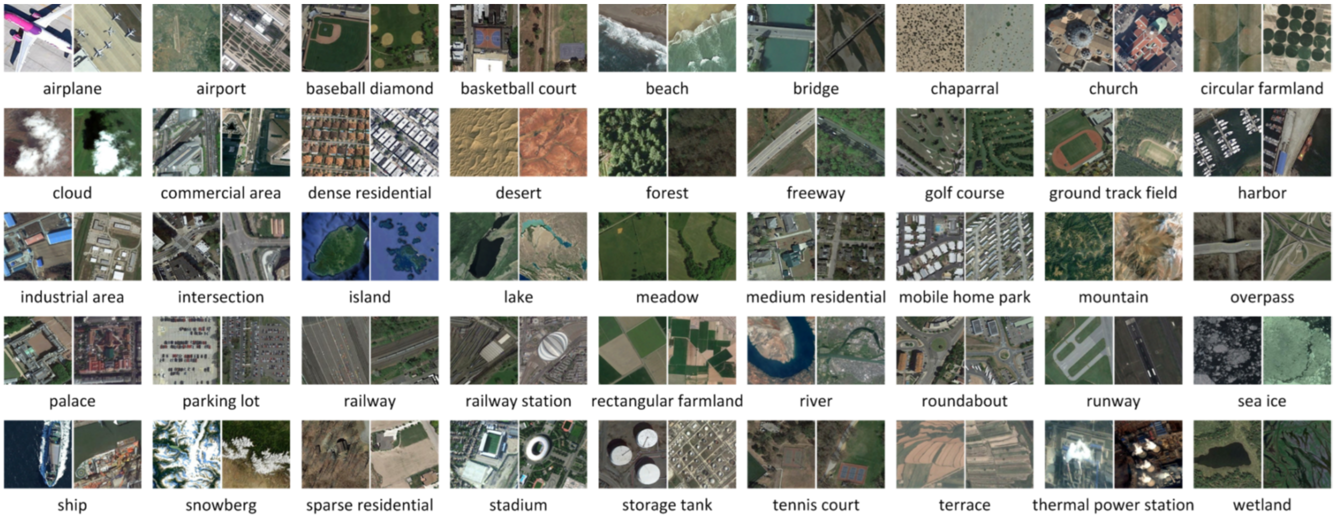
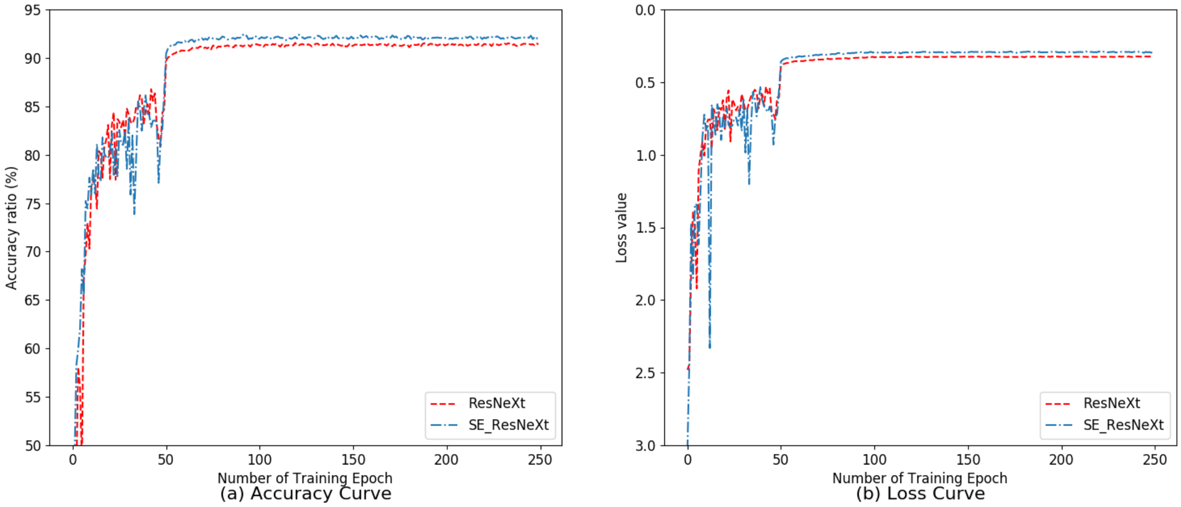


Fig. 2: Some instances from NWPU-RESISC45 dataset. Each class shows two images.



(a) Accuracy Curve

(b) Loss Curve

Fig. 3: Accuracy and loss curves for SE_ResNext and ResNext. (a) is the accuracy curve, (b) is the loss curve.

III. Experiments and Analysis

A. Experimental Setting

In the experiments, we comprehensively evaluate the performance of the introduced deep CNN-based scene classification method on a publicly available NWPU-RESISC45. This dataset keeps 31,500 remote sensing images divided into 45 scene classes and each scene class contains 700 images, each of which is set at 256×256 pixels in the RGB color space. The dataset was extracted by the specialists of the remote sensing field from Google Earth with spatial resolution ranges from about 30m to 0.2m per pixel.The NWPU-RESISC45 is a large scale dataset of scene classes. It maintains a lot of variations in translation, viewpoint, object pose, spatial resolution, illumination, background, and occlusion. Accordingly, it contains high sufficiency between class similarity and within-class diversity.

All experiments were carried out on a work bench equipped with an Intel CPU i7 5820 k, an NVIDIA Geforce Pascal Titan X GPU, and 32 GB DDR4 memory. Operation system is Ubuntu 16.10. The deep learning framework used is Pytorch. The comparison CNNs include AlexNet, VGG-16, ResNet and ResNeXt. During the limitation of input image size, all

TABLE I: **Classification results of all CNN models on the NWPU-RESISC45 dataset**

CNN Name	**top-1 acc(%)**	**best top-1 acc(%)**	**top-5 acc (%)**
AlexNet	68.38	68.75	90.93
VGG-16	77.16	77.85	95.23
ResNet	89.87	90.54	98.72
ResNext	91.39	91.55	99.02
SE_ResNeXt	92.18	92.31	99.05

image samples are pre-resized to meet the input of each CNN (224×224 or 227×227). The dataset is randomly split into 90% for training and 10% for test. All CNN models are trained by 250 epoch on the training set. The initial learning rate is set to 0.1 and multiplied 0.1 per 50 epochs. All experiments were performed through five times, by averaging over all results. Data augmentation was not used in each method. Overall accuracy, a widely-available standard evaluation metrics in image classification, is set as the only method here. Overall accuracy is calculated as the number of correctly classified samples divided by the total number of samples.

B. Analysis of Experimental Results

We have collected all the classification results, including final top 1 accuracy, best top 1 accuracy during training process, and top 5 results in TABLE I. It is important to note that all of our models here are not pre-trained on ImageNet [15], which means there are no prior knowledge from other fields. In TABLE I, AlexNet and VGG achieve the worst performance. The remaining three networks have achieved good performance far beyond AlexNet and VGG. It shows that CNNs have significant development and the recent and advanced models absolutely outperform the early networks. SE_ResNeXt gets the best performance in both the top1 and top5 results. This shows that its network architecture is very efficient to extract the category features of the HRRS scene image compared to the comparison algorithm. Since the SE_ResNeXt are developed based on the ResNeXt, we collect data of the training process and show it in Fig. 3. In 0∼50 epochs, the convergence of the two methods is faster. After 50 epoch, SE_ResNeXt is superior to resnext on accuracy and loss. And the curves of accuracy and loss keep stable.

IV. CONCLUSIONS

In this paper, we introduced a deep CNN-based scene classification method, which not only enables to enhance the ability of spatial representation, but adaptively recalibrates channel-wise feature responses to suppress useless feature channels. Comparison with state-of-the-arts CNN models in remote sensing on a large-scale HRRS image dataset demonstrated the effectiveness of the proposed method. In the future work, we would test the method on more publicly available HRRS image datasets; and apply prior knowledge in remote sensing to improve the performance of the proposed method.

V. Acknowledgment

We gratefully acknowledged the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research as well.

References

[1] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 99, pp. 1–19, 2017.

[2] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[3] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[4] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2015.

[5] G. Cheng, C. Ma, P. Zhou, X. Yao, and J. Han, "Scene classification of high resolution remote sensing images using convolutional neural networks," in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*. IEEE, 2016, pp. 767–770.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *In Proceedings of the International Conference on Learning Representations*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," . Available online: http://arxiv.org/abs/1508.00092 (accessed on 14 August 2015).

[10] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2017.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, 2017, pp. 4278–4284.

[14] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5987–5995.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.