

BAND-WISE MULTI-SCALE CNN ARCHITECTURE FOR REMOTE SENSING IMAGE SCENE CLASSIFICATION

Jian Kang and Begüm Demir

Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin

ABSTRACT

Most of the existing convolutional neural network (CNN) architectures in the framework of image scene classification problems are designed for modeling RGB image bands. Direct application of these architectures to the high-dimensional remote sensing (RS) scene classification can be insufficient to accurately describe the spectral content. To address this issue, we propose a novel CNN architecture for the feature embedding of high-dimensional RS images. The proposed architecture aims at: 1) decoupling the spectral and spatial feature extraction for sufficiently describing the complex information content of images; and 2) taking advantage of multi-scale representations of different land-use and land-cover classes present in the images. To this end, the proposed architecture is mainly composed of: 1) a convolutional layer for band-wise extraction of multi-scale spatial features; 2) a convolutional layer for pixel-wise extraction of spectral features; and 3) standard 2D convolution and residual blocks for further feature learning. Experiments on BigEarthNet validate the effectiveness of the proposed method, when compared to the state-of-the-art CNN architectures.

Index Terms— Deep learning, convolutional neural networks, scene classification, feature extraction, remote sensing

1. INTRODUCTION

With the rapid development of Earth observation missions, remote sensing (RS) scene classification has drawn significant attentions. Supported by large-scale RS archives [1–3], deep learning has been widely applied for RS scene classification, owing to its excellent feature extraction capability. Most of the proposed methods for scene classification are based on RGB RS images, where the pre-trained convolutional neural network (CNN) architectures on the large-scale computer vision archives (e.g., ImageNet) are often adopted. As an example, in [4], the effectiveness of RS scene classification based on the transferred image representations learned from the ImageNet dataset is first investigated. A fine-grained scene classification at a building-instance level based on several pre-trained CNN architectures is proposed in [5]. Due to the different numbers of spectral bands, the pre-trained CNN architectures on the ImageNet dataset cannot be directly applied

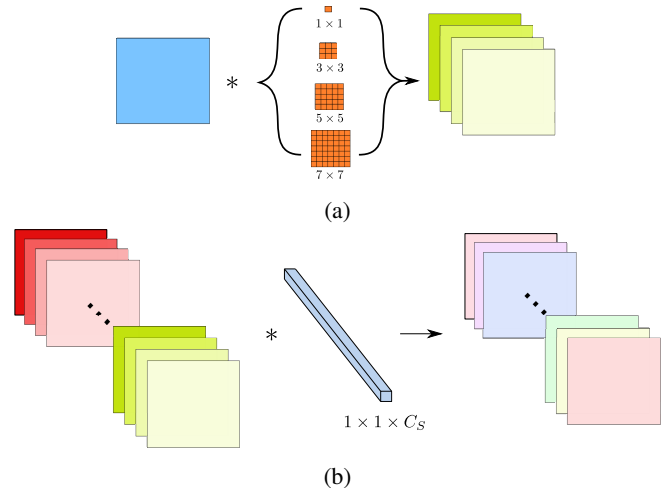


Fig. 1: The spectral-spatial feature extraction module in the proposed band-wise multi-scale CNN architecture: (a) spatial feature extraction based on the band-wise multi-scale convolution; and (b) spectral feature extraction based on the pixel-wise convolution.

on the scene classification with high-dimensional RS images (e.g., multispectral images). To address this issue, one can modify the first convolutional layer to be suitable with the high-dimensional RS images by simply changing the input number of bands and keep the remaining layers initialized by the pre-trained CNN architectures. However, such modification may not sufficiently encode the spectral-spatial information content of high-dimensional RS images.

In order to characterize such information, several methods are proposed in the context of pixel-based image classification. As an example, in [6] a band attention module based on the band characteristic is proposed for hyperspectral image classification. In [7], a novel group CNN architecture for spectral-spatial classification of hyperspectral images is introduced. Although these methods can improve the success of the feature extraction methods for the pixel-level classification problems, they cannot be directly applied to the scene classification.

To address this issue, in this work we propose a novel CNN architecture for accurately learning the features from

Table 1: The network configuration of the proposed BWMS.

Output size	BWMS
60×60	band-wise conv, $\{1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7\}$, 40, stride $\{1, 2\}$, concat
	pixel-wise conv, 1×1 , 32, stride 1
	conv, 3×3 , 32, stride 1
	conv, 3×3 , 64, stride 1
30×30	max pool, 3×3 , stride 2
	$\begin{bmatrix} \text{conv}, 3 \times 3, 64 \\ \text{conv}, 3 \times 3, 64 \end{bmatrix} \times 2$
15×15	$\begin{bmatrix} \text{conv}, 3 \times 3, 128 \\ \text{conv}, 3 \times 3, 128 \end{bmatrix} \times 2$
8×8	$\begin{bmatrix} \text{conv}, 3 \times 3, 256 \\ \text{conv}, 3 \times 3, 256 \end{bmatrix} \times 2$
4×4	$\begin{bmatrix} \text{conv}, 3 \times 3, 512 \\ \text{conv}, 3 \times 3, 512 \end{bmatrix} \times 2$
1×1	global average pooling, 128-d fc, 19-d fc, sigmoid

high-dimensional RS scene images. The proposed architecture consists of three modules: 1) spatial feature extraction based on band-wise multi-scale convolution; 2) spectral feature extraction based on pixel-wise convolution and 3) standard 2D convolutional and residual blocks for further feature learning.

2. BAND-WISE MULTI-SCALE CNN ARCHITECTURE

Let $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a RS archive that consists of N images, where \mathbf{x}_i is the i -th image represented as $\mathbf{x}_i = \{\mathbf{x}_i^1, \dots, \mathbf{x}_i^C\}$. \mathbf{x}_i^c is the c -th image band, and C is the total number of bands. Each image is associated with labels from a label set $\mathcal{L} = \{l_1, \dots, l_P\}$. The labels of \mathbf{x}_i can be represented by a label vector $\mathbf{y}_i \in \{0, 1\}^P$, where the p -th element is set to 1 if l_p is associated with \mathbf{x}_i and 0 otherwise. We aim to learn a mapping $\mathcal{F}(\cdot)$ based on a CNN model that assigns class labels to image \mathbf{x}_i , where the label prediction accuracy is as high as possible.

The standard 2D convolutional layer for encoding the spectral-spatial features of \mathbf{x}_i can be described as:

$$\mathbf{u}_m = \sigma(\mathbf{w}_m * \mathbf{x}_i) = \sigma\left(\sum_{c=1}^C \mathbf{w}_m^c * \mathbf{x}_i^c\right), \quad (1)$$

where $\mathbf{w}_m = \{\mathbf{w}_m^1, \dots, \mathbf{w}_m^C\}$ denotes the parameters of the m -th filter, $\mathbf{w}_m^c \in \mathbb{R}^{K \times K}$ is the c -th 2D spatial filter, $\sigma(\cdot)$ represents the activation function and $*$ is the convolutional operation. The bias term is omitted here for the simplicity. It can be observed that the feature maps are produced by the summation of all the band-wise convolution results based on the 2D spatial filter and the associated band. However, through

this operation, the spectral features may not be optimally extracted, since the process for the spectral feature extraction is entangled within the summation of the spatial convolution results. Moreover, for spatial information encoding, K is usually set to a fixed number, (e.g., $K = 3$ or $K = 7$), in the convolutional layer. However, the convolutional layer with the fixed size of 2D spatial filter may not optimally extract the spatial features in that case. Different land-use and land-cover classes usually cover the scene with different spatial sizes. For instance, *Arable land* may cover the whole area of a scene, while only a small part of a scene belongs to *Industrial units*.

To overcome these limitations, we enforce the first convolutional layer to extract the spectral and spatial features, separately. To this end, we propose a Band-Wise Multi-Scale (BWMS) CNN architecture composed of: 1) spatial feature extraction based on a band-wise multi-scale convolution, which can sufficiently learn band-wise multi-scale spatial features; 2) spectral feature extraction based on a pixel-wise convolution, which is mainly designed for spectral information fusion in a pixel-wise manner; and 3) standard 2D convolutional and residual blocks for further feature learning. The spatial feature extraction is conducted by convolving multi-scale 2D spatial filters with each band of the input image \mathbf{x}_i . Specifically, such operation can be formulated as:

$$\hat{\mathbf{x}}_i^{c_s} = \mathbf{w}^{c_s} * \mathbf{x}_i^c, \quad (2)$$

where $\hat{\mathbf{x}}_i = \{\hat{\mathbf{x}}_i^1, \dots, \hat{\mathbf{x}}_i^{C_S}\}$ denotes the produced multi-scale feature maps, $\hat{\mathbf{x}}_i^{c_s}$ is the s -th scale feature map of image band \mathbf{x}_i^c , \mathbf{w}^{c_s} is the 2D spatial filter at the s -th scale of the c -th band, and S is the total number of scales. As an illustration, the operation of (2) is demonstrated in Figure 1 (a). Each image band is convolved with four 2D spatial filters with the sizes of 1×1 , 3×3 , 5×5 and 7×7 , and results in four feature maps at different scales. Then, the spectral feature extraction is carried out on the produced multi-scale feature maps with the 1×1 2D convolutional filters. This process can be described as:

$$u_m(i, j) = \sigma\left(\sum_{c_s} \hat{w}_m^{c_s} \hat{\mathbf{x}}_i^{c_s}(i, j)\right), \quad (3)$$

where (i, j) is the pixel index, $\hat{\mathbf{w}}_m = \{\hat{w}_m^1, \dots, \hat{w}_m^{C_S}\}$ denotes the weights of the spectral filters, and $\hat{w}_m^{c_s}$ is c_s -th element. Differently from (1), the features extracted by (3) are obtained through a weighted summation of the multi-scale feature maps in a pixel-wise manner. These weights are learned to address the different contributions of the spectral bands.

The proposed spectral-spatial feature extraction formulas given in (2) and (3) decouples the spectral-spatial feature learning procedure through the standard 2D convolutional operation given in (1). The parameterized filters are individually learned to extract the corresponding spatial and spectral features. In this way, the spectral features can be more accurately

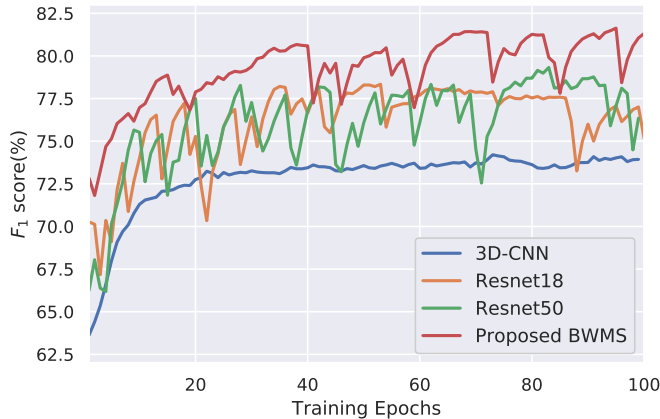


Fig. 2: Learning curves of all the CNN architectures.

Table 2: Classification performances (%) under four different metrics and the numbers of parameters (#para) for 3D-CNN, ResNet18, ResNet50 and the proposed BWMS.

Architectures	F_1	Acc	HL	RL	#para
3D-CNN	74.67	64.73	7.74	4.51	1.75M
ResNet18	78.68	69.38	6.74	3.52	11.2M
ResNet50	81.05	72.13	6.18	2.87	23.6M
Proposed BWMS	81.84	73.07	5.97	2.70	11.3M

extracted compared to the approach that exploits the mixed operation given in (1). Moreover, the multi-scale spatial filters are conducive for learning the multi-scale representations of different classes, which facilitates the follow-up feature transformation. After projecting the original high-dimensional RS images into the spectral-spatial features through (2) and (3), we adopt standard convolutional and residual blocks [8] for further feature learning. The remaining layers of the proposed BWMS are set as the same of the corresponding layers in ResNet18 [8]. As an example, the network configuration of the proposed BWMS is described in Table 1.

3. EXPERIMENTAL RESULTS

To evaluate the proposed CNN architecture, experiments were conducted on BigEarthNet¹, which is a large-scale benchmark archive composed of 590,326 Sentinel-2 images [1]. Each image is composed of: 1) 10m bands with 120×120 pixels; 2) 20m bands with 60×60 pixels; and 3) 60m bands with 20×20 pixels. The archive is split into train, validation and test sets with 269,695, 123,723, and 125,866 images, respectively. These numbers are obtained after eliminating images that are fully covered by seasonal snow, cloud and cloud shadow. For the details on BigEarthNet, the readers are referred to [1].

The proposed method is implemented in PyTorch. The

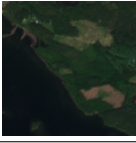

class probabilities are obtained by applying *sigmoid* activation function on the last layer. To train the network, we exploit the binary cross-entropy loss optimized by Adam with the initial learning rate of 10^{-3} and the weight decay of 10^{-4} . The mini-batch size is 500 and the epoch number is 100. The learning rate is reduced by a factor of 0.5 every 30 epochs. We use *Leaky ReLU* as the activation function in the proposed architecture. ResNet18, ResNet50 and 3D-CNN, which is also exploited to extract spectral-spatial features from high-dimensional RS images, are selected as baseline methods for the comparison.

The evaluation was conducted based on several metrics for multi-label classification: 1) F_1 score (which is an integrated metric of sample precision and recall); 2) Accuracy (Acc) (which reflects the degree of sample-wise correctness); 3) Hamming Loss (HL) (which evaluates the fraction of misclassified labels); and 4) Ranking Loss (RL) (which evaluates the fraction of reversely ordered label pairs).

Table 2 demonstrates the classification results on the test set based on all the CNN architectures under four metrics. Moreover, the numbers of parameters in the CNN architectures are presented in the table. As shown in Table 1, most of the parts of BWMS are the same as ResNet18, except the proposed spectral-spatial feature extraction part. Compared with ResNet18, the performances of F_1 score and Acc based on BWMS can achieve the improvement of around 3% and 4%, respectively. In terms of HL and RL, BWMS can provide an improvement about 1% compared to ResNet18. However, the numbers of parameters of ResNet18 and BWMS are similar. Furthermore, by increasing the CNN depth, ResNet50 still cannot achieve the same classification performance as BWMS. All the metrics of ResNet50 are around 1% lower than BWMS. However, the number of parameters of ResNet50 is more than two times of BWMS. In other words, BWMS can reach the promising classification accuracy with significantly reduced computational cost, which is much lower than the existing deep CNN architectures. Thus, BWMS is more suitable for the large-scale scene classification. Table 3 shows two test images associated with their true multi-labels and the labels assigned by ResNet18, ResNet50 and BWMS. Note that the performance of 3D-CNN is lower than the other considered methods (see Table 2). Thus, we do not include its results in this figure. For the first image, all the methods can correctly predict *Coniferous forest*, *Mixed forest* and *Inland waters*. However, ResNet18 wrongly predicts *Broad-leaved forest* present in the scene. It indicates that the distinction between the class of *Broad-leaved forest* and *Coniferous forest* or *Mixed forest* may not be correctly captured by ResNet18. For the second image, all the methods can accurately categorize *Coniferous forest* and *Pastures*. However, in terms of *Arable land*, ResNet50 creates a false positive. From the two images, it can be observed that the proposed architecture can sufficiently capture the discriminative features to represent different classes. Figure 2 demonstrates

¹The BigEarthNet is available at <http://bigearth.net>.

Table 3: Images associated with their true multi-labels and the multi-labels assigned by ResNet18, ResNet50 and the proposed BWMS.

Image	Multi-Labels	ResNet18	ResNet50	Proposed BWMS
	Coniferous forest Mixed forest Inland waters	Broad-leaved forest Coniferous forest Mixed forest Inland waters	Coniferous forest Mixed forest Inland waters	Coniferous forest Mixed forest Inland waters
	Coniferous forest Pastures	Coniferous forest Pastures	Arable land Coniferous forest Pastures	Coniferous forest Pastures

the F_1 score calculated on the validation set by different CNN architectures with respect to the number of training epochs. From the figure, it can be seen that the proposed CNN architecture can preserve the higher classification accuracy compared to 3D-CNN, ResNet18 and ResNet50 during the training process. From the experimental results, one can conclude that the proposed spectral-spatial feature extraction modules can sufficiently capture the inherent spectral-spatial information content of the high-dimensional RS images. Moreover, the significant improvement of the classification results based on the proposed CNN architecture demonstrate the prominent role of the early layers on the feature encoding for high-dimensional RS images. It is worth noting that the learning capability of a CNN architecture for high-dimensional RS images not only depends on its depth, but also highly relies on the configuration of early layers for spectral-spatial feature extraction.

4. CONCLUSION

In this paper, we have proposed a novel CNN architecture for accurately capturing the spectral-spatial information content present in high-dimensional RS images. To this end, the proposed CNN architecture is mainly composed of: 1) a convolutional layer for extracting band-wise multi-scale spatial features; 2) a convolutional layer for extracting pixel-wise spectral features; 3) standard 2D convolutional and residual blocks for mapping the spectral-spatial features into the associated classes. Experiments on BigEarthNet validate the effectiveness of the proposed CNN architecture compared to the several state-of-the-art CNN architectures. In particular, experimental results point out that the learning capability of a CNN architecture for high-dimensional RS images relies on its configuration of early layers for spectral-spatial feature extraction. A CNN architecture with a suitable spectral-spatial feature extraction layer can improve the classification performance and can also significantly reduce the associated number of parameters (and thus the computational cost). As a future work, we will investigate the effectiveness of the pro-

posed CNN architecture on large-scale image retrieval problems.

5. ACKNOWLEDGEMENT

This work was supported by the European Research Council under the ERC Starting Grant BigEarth-759764.

6. REFERENCES

- [1] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, “Bigearth-net: A large-scale benchmark archive for remote sensing image understanding,” in *IEEE International Geoscience and Remote Sensing Symposium*, 2019.
- [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] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, “Sen12ms—a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion,” *arXiv preprint arXiv:1906.07789*, 2019.
- [4] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, “Deep learning earth observation classification using imagenet pretrained networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, 2015.
- [5] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, “Building instance classification using street view images,” *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 44–59, 2018.
- [6] H. Dong, L. Zhang, and B. Zou, “Band attention convolutional networks for hyperspectral image classification,” *arXiv preprint arXiv:1906.04379*, 2019.
- [7] X. Li, M. Ding, and A. Pižurica, “Group convolutional neural networks for hyperspectral image classification,” in *IEEE International Conference on Image Processing*, 2019, pp. 639–643.
- [8] 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.