LAND USE CLASSIFICATION BASED ON MULTI-STRUCTURE CONVOLUTION NEURAL NETWORK FEATURES CASCADING

J. Men[1], L. Fang[2], Y. Liu[1], Y. Sun[1]

[1]China University of Geoscience, Wuhan, China - yueyanliu@cug.edu.cn, (1032764332,1120167513)@qq.com
[2] Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Science Quanzhou, China -fangli@fjirsm.ac.cn

ICWG II/III: Pattern Analysis in Remote Sensing

KEY WORDS: Remote Sensing; High Resolution Image; Land Use; Classification; Convolutional Neural Network

ABSTRACT:

Learning efficient image representations is at the core of the classification task of remote sensing imagery. The existing methods for solving image classification task, based on either feature coding approaches extracted from convolution neural networks(CNNs) or training new CNNs, can only generate image features with limited representative ability, which essentially prevents them from achieving better performance. In this paper, we investigate how to transfer features from these successfully pre-trained CNNs for classification. We propose a scenario for generating image features via cascading features extracted from different CNNs. First, pre-trained CNNs, like CaffeNet, VGG-S and VGG-F, are used as feature extractor since their different structures help extract richer information of images. Then the fully-connected layers of the pre-trained CNNs are fine-tuned with UC Merced land use dataset. Finally, the image features generating from cascading the outputs of three networks above, are fed into multi-class Optimal Margin Distribution Machine (mcODM) to obtain the final classification results. Extensive experiments on public land use classification dataset demonstrates that the image features obtained by the proposed scenario can result in remarkable performance and improve the state-of-the-art by a significant margin. The results reveal that the features from pre-trained CNNs generalize well to land use dataset and are more expressive than features from single CNN.

1. INTRODUCTION

The division of land use types has complex natural and social attributes, which makes meeting the user's need for classification of land use become a hot and difficult problem in the field of land resource management. With the progress of remote sensing technology, remote sensing images can provide more abundant property information, which provides the possibility of solving the above-mentioned problems (Castelluccio et al., 2015; Ban et al., 2015).

The core of land use classification lies in the effective expression of land use type information by using expressive features descriptor (Cheng et al., 2017). According to the dimension of feature extraction, features can be roughly divided into three categories: low-level features, middle-level features and high-level features (Xia et al., 2017). Low-level features are based on visual attributes (texture, structure, spatial information, etc.), such as scale invariant feature transform (SIFT) (Lowe et al., 2004). Although it can achieve good classification results for general classification tasks, the limitations of its poor generalization ability are exposed for classification tasks with many kinds of scenes and high complexity. By coding low-level features, middle-level features can improve the semantic expression ability of the model and obtain better classification results (Yang et al., 2010; Lazebnik et al., 2006). However, middle-level features depend extensively on low-level manual features, and there are constraints on the expressiveness of features and the learning ability of models. In recent years, deep learning has become a research hotspot in many fields. As one of the most successful deep learning models, convolutional neural networks (CNNs) (Krizhevsky et al., 2012) has made important breakthroughs in scene classification by virtue of its outstanding learning performance. At present, several ways of using convolutional neural network are as follows: (1) Using Pre-trained CNNs as Feature Extractor. Some researchers combine pre-trained CNNs as feature extractors with traditional coding methods, such as bag of visual words(BOVW)，improved Fisher vector(IFK), which improves the classification accuracy to a certain extent. (2) Fine-tuning pre-training CNNs. In this method, specific datasets are used to train only some layers in CNNs, and in this way, information can be added to the model, while other network layers are frozen. And many experimental results show that the method of fine-tuning can effectively improve the results of classification. (3) Training a new CNN. A CNN architecture is adopted to initialize the parameters randomly, and then a large number of training datasets are used to train the whole model. This training method can effectively avoid the problem of over-fitting, but there are some problems in the application, such as long training cycle and large demand for training data.

Based on the above research, this paper proposes a method of multi-structure convolutional neural network features cascading (MCNNFC) for land use classification. First, we select CaffeNet, VGG-S and VGG-F three pre-trained models as feature extractors. Then the first two full-connection layers parameters of the above networks are fine-tuned using UC Merced land-use dataset. Thus, we removed the SoftMax classifier from CNNs and replaced it with Multi-Class Optimal Margin Distribution Machine (mcODM), and the CNN features extracted from second full-connect layer are cascaded as input of mcODM classifier (as shown in Figure 1).

2. RELATED WORK

2.1 Convolutional Neural Network

The typical architecture of a CNN is composed of multiple functional layers, including convolutional layers, pooling layers, fully-connected(FC) layers and classifier. The convolution layers are responsible for extracting simple features such as texture and geometry. Since different convolution kernels have different sensitivity to features, we select three CNNs (CaffeNet, VGG-S, VGG-F) with different sizes of convolution kernels as experimental models.

CaffeNet: CaffeNet model has five convolution layers and two full connection layers. The size of convolution core is large, and it runs faster in classification and training.

VGG-S: VGG-S not only increases the number of convolution cores but also reduces the size, which makes it more capable of expressing the feature details extracted, but also reduces the operational efficiency.

VGG-F: The network structure of VGG-F is similar to that of CaffeNet, but it runs faster.

Model	Size of Convolution Core				
	COV1	COV2	COV3	COV4	COV5
CaffeNet	11×3×96	5×48×256	3×256×384	3×192×384	3×192×256
VGG-S	7×3×96	5×96×256	3×256×512	3×512×512	3×512×512
VGG-F	11×3×64	5×64×256	3×256×256	3×256×256	3×256×256

Table 1. Convolution core size of CNNs

2.2 Fine-tune

Compared with training a new CNN model, fine-tuning can train the model pertinently and is more effective. In the process of feature extraction, convolution layers are responsible for transforming low-level features into high-level features, and forming the global expression of images in FC layers. So the features from FC layers have a better expression than them from convolution layers.

Previous studies have shown that (Cheng et al., 2017; Nogueira et al., 2017; Cortes et al., 1995), full-connection layer has better trainability. In this paper, we fine-tune the parameters of FC layer and keep the convolution layer unchanged. Fine-tuning of CNNs using stochastic gradient descent (SGD) algorithm, calculating the minimum square sum of the error between the actual output and the theoretical output, and updating the weights and thresholds of the network. The calculation equation is as follows.

$$E^N = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\left(t_k^n - y_k^n\right)^2 \tag{1}$$

where E^N = calculation error
t_k^n = the k-dimension of the label corresponding to the nth sample
y_k^n = the k-dimension of the actual output value of the nth Sample
c = number of categories of samples
N = number of training samples

2.3 Feature Descriptors

The output of each convolution layer can be used as a feature descriptor， and the feature descriptors of different convolution layers have different expression effects. The convolution calculation is shown in Equation (2).

$$F_{ij} = f(b_1 + \sum_{i=1}^{n}\sum_{j=1}^{n} k_{ij} \times x_{ij}) \tag{2}$$

where $i \in R$， $j \in R$
K_{ij} = the value of column j of row i in the convolution kernel
X_{ij} = the value of input item corresponding to K_{ij}
b_1 = bias
$f(\cdot)$= Relu($\cdot$)

After obtaining convolution feature, dimension reduction is carried out by pooling layers (equation (3)) to prevent feature dimension from being too high.

$$x_j^n = f(b_j^n + \beta_j^n \times down(x_j^{n-1})) \tag{3}$$

where $down(\bullet)$ = down sampling function
β = multiplicative bias
b = additive bias

This process can avoid the "dimension disaster" and over-fitting problems.

The convolution layers encode the features continuously through Equation (2) and (3), and gets the feature maps which are input to the FC layer. The output of second FC layers in CNNs (CaffeNet, VGG-S, VGG-F) are cascaded through Equation (4) to get a new feature descriptor.

$$F_{\text{MCNNFC}} = \sum_{i=1}^{n} F_i \tag{4}$$

where F = feature descriptor

In terms of classifiers, mcODM can effectively solve the problem of boundary optimization in multi-category classification. Compared with support vector machine (SVM), mcODM shows general advantages in classification accuracy and running time. Therefore, F_{MCNNFC} is input into mcODM to obtain the classification results.

2.4 Work Flow of Proposed Method

An overview of the method used can be seen in Figure 1. The pre-training models are trained by the ImageNet dataset and fine-tuned by the UC Merced land use dataset in its FC layers. The fine-tuned models are used to extract features, and then the outputs of the second FC layer of CNNs as the feature expression of the image are cascaded into a new feature descriptor and input into the mcODM classifier to get the classification results.

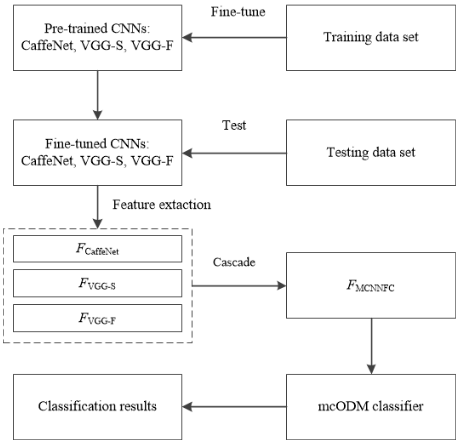


Figure 1. The workflow of the proposed method

3. EXPERIMENTS AND ANALYSIS

3.1 Experimental setup

We evaluate the effectiveness of proposed method on the following publicly UC Merced dataset, as shown in Figure 2.

The UC Merced dataset: manually collected from large aerial orthoimagery, contains 21 distinctive scene categories. Each class consists of 100 images with a size of 256×256 pixels. Each image has a pixel resolution of one foot. Figure 2 shows some examples of each category included in this dataset. Note that this dataset shows very small inter-class diversity among some categories that share a few similar objects or texture (e.g., dense residential and medium residential), which makes the UC Merced dataset a challenging one.

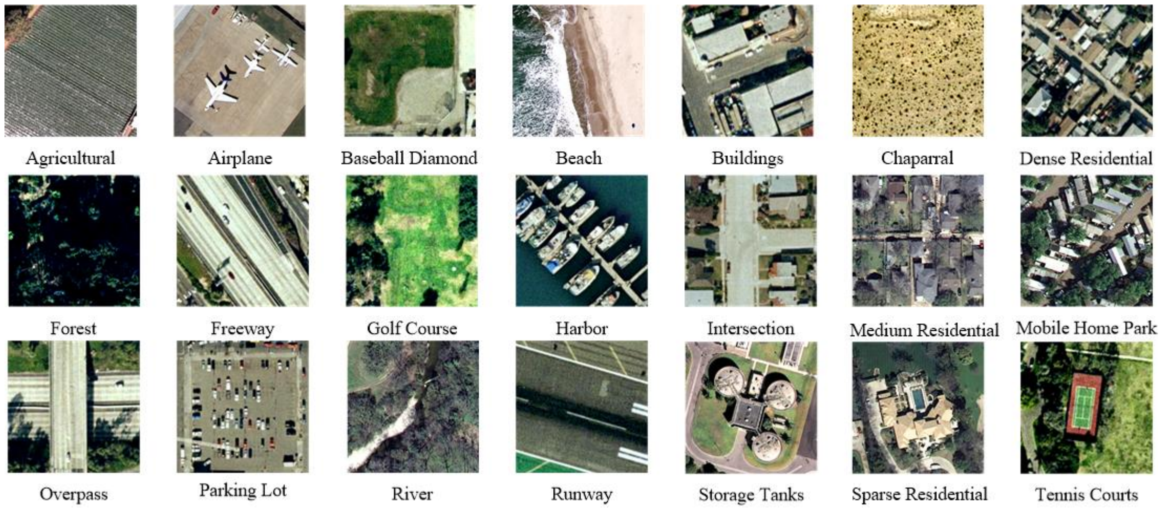


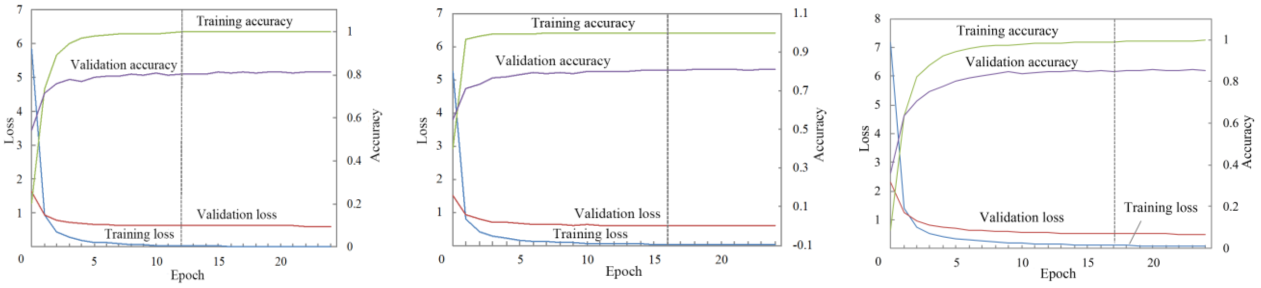
Figure 2. Examples of UC Merced land-use dataset

For the experimental setup, we first evaluated the effect of CNNs fine-tuning on classification. Judging the effect of CNNs fine-tuned by training and validation accuracy curve and loss curve. Second, we verified the validity of features cascading method. Several original pre-trained CNNs full connection layer features are cascaded to obtain the optimal classification results, and the performance of mcODM and SVM classifiers were compared. Finally, we compared proposed method with other methods, such as SPM (Lazebnik et al., 2006), BOVW+SCK (Yang et al., 2010), VGG-S (Chatfield et al., 2014), VGG-M (Chatfield et al., 2014), VGG-F(Hu et al., 2015), VGG-19(Simonyan et al., 2014), VLAD+VGG-VD16(Hu et al., 2015), IFK+VGG-M (Hu et al., 2015), Using OverFeat (Penatti et al., 2015), CaffeNet + SVM(Castelluccio et al., 2015), etc. And 70% randomly selected as training set, 30% as test set. The configurations of experiments are as follows: Intel(R) Core (TM) i7-3770 CPU @ 3.40 GHz 3.40 GHz, 16GB RAM and MATLAB 2014b.

3.2 Experimental results and analysis

3.2.1 Fine-tune

The UC Merced land use dataset is used to train the parameters of CNN two full connection layers. And the learning rate was set as 0.01, the batch size was set as 100, the weight decay rate was set as 0.002, and the number of training set and validation set were set as 1449 and 651 separately. The following figures show the loss value and classification accuracy of training dataset and validation dataset after each iteration.



(a)CaffeNet (b)VGG-F （c）VGG-S

Figure 3. Cross entropy loss and classification accuracy in fine-tuning

Figure 3 shows the training process of three models. With the improvement of training accuracy, the validation accuracy shows a trend of rapid growth first and then gradually stable, which indicates that the fine-tuned models have good generalization ability. At the same time, the curves of training loss and validation loss decrease rapidly in the previous iterations. With the increase of iterations, VGG-S has a steady trend. Compared with the other two models, VGG-S has a larger initial training loss value and shows stronger learning ability, because after parameter training, its validation accuracy improves the most. As the loss value decreases, the learning ability of the model decreases gradually. When the number of iterations (dotted line) are 12, 16 and 17, the curves of loss and accuracy become stable, training process was stopped. In Figure 4, the overall changes of classification accuracy before and after fine-tuning are compared. From Figure 4, it can be seen that the overall classification accuracy of each method has been improved obviously. But due to the difference of network architecture, the range of increase is different, among which CaffeNet has the highest increase, 5.71%, VGG-F has increased 5.1%, VGG-S has the lowest increase, 2.72%; MCNNFC still has the highest classification accuracy after fine-tuning, reaching 97.55%.

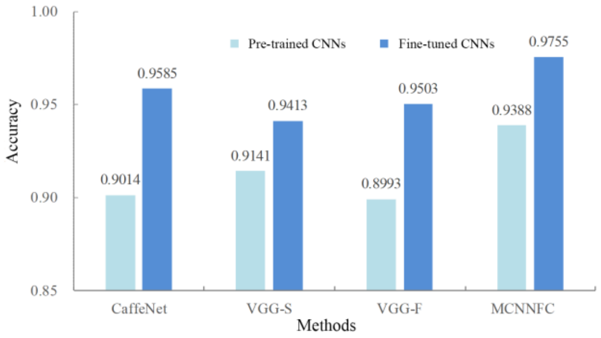


Figure 4. Result of CNN overall classification accuracy after fine-tuning

3.2.2 Feature cascading

In order to exclude the influence of training on the models, we cascaded the outputs of the second FC layers of pre-trained CNNs to study the expressive ability of features. According to the number of cascaded CNNs, the cascade methods are divided into two networks cascade and three networks cascade, and combined the new feature descriptors with SVM and mcODM respectively. The experimental results are shown in Table 2.

Methods	Classification accuracy（%）	
	SVM	mcODM
CaffeNet	89.63 ± 0.95	**90.14 ± 0.63**
VGG-S	91.11 ± 0.81	**91.41 ± 0.78**
VGG-F	89.71 ± 1.02	**89.93 ± 0.74**
CaffeNet + VGG-S	92.45 ± 0.71	**92.72 ± 0.34**
VGG-S + VGG-F	**92.59 ± 0.60**	92.36 ± 0.58
CaffeNet + VGG-F	**91.65 ± 0.52**	91.59 ± 0.55
CaffeNet + VGG-S + VGG-F	92.81 ± 0.86	**93.88 ± 0.72**

Table 2. Results of different cascading ways

When cascading two kinds of CNNs, the CaffeNet + VGG-S + mcODM method has the best result, and the classification accuracy has been improved by 2.58% and 1.31% respectively, and the maximum improvements of the other two methods (VGG-S + VGG-F+SVM, CaffeNet + VGG-F+SVM) are 1.94% and 2.02% respectively. The method of combining CaffeNet + VGG-S + VGG-F with mcODM classifier achieves 93.88%. In addition, compared with SVM, mcODM has better overall performance and more stability. So, the method of CaffeNet + VGG-S + VGG-F feature cascade combined with mcODM has the most significant effect on improving classification accuracy.

3.2.3 Comparison with other classification methods

In this section, we compared F_{MCNNFC} + SVM and F_{MCNNFC} + mcODM with other methods, F_{MCNNFC} is obtained by cascading fine-tuned CaffeNet, VGG-F and VGG-S FC layer features. The classification accuracy is shown in Table 3.

Methods	Accuracy（%）
SPM (Lazebnik et al., 2006)	74
BOVW+SCK (Lazebnik et al., 2006)	77.71
Using OverFeat(Perronnin et al., 2015)	90.91 ± 1.19
CaffeNet+SVM (Perronnin et al., 2015)	93.42 ± 1.00
VLAD+VGG-VD16 (Hu et al., 2015)	95.16
IFK+VGG-M (Hu et al., 2015)	96.9
GoogLeNet with fine-tuning (Castelluccio et al,2015)	97.1
VGG-S (Chatfield et al, 2014)	87.76 ± 0.66
VGG-F (Hu et al., 2015)	94.35
VGG-M (Chatfield et al, 2014)	87.15 ± 0.80
VGG-19 (Simonyan et al, 2014)	91.8
F_{MCNNFC} + SVM	96.71 ± 0.42
F_{MCNNFC} + mcODM [This paper]	**97.55 ± 0.74**

Table 3. Accuracy of proposed method and state-of-the-art methods over UC Merced land use dataset

It can be seen that SPM, BOVW+SCK and other high-level feature coding methods improve the expression ability of low-level features, but they are obviously disadvantaged compared with high-level features such as CafeNet + SVM. And high-level features have brought great improvement to classification accuracy. At the same time, the classification accuracy of cascade features between SVM and mcODM is compared. Experiments show that the classification accuracy of mcODM is 97.55% higher than that of SVM.

Figure 5. compares the classification accuracy of each classification method for a single land use type. In class I, MCNNFC achieves the highest classification accuracy over a single CNN; in class II, MCNNFC and a single CNN achieve the highest classification accuracy; in class III, MCNNFC is lower than the highest classification accuracy of a single CNN.

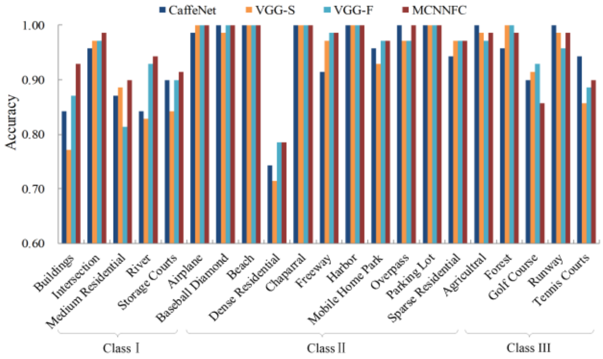


Figure 5. Classification accuracy of single land use by different methods

The classification accuracy of each classification method in class I is relatively low (below 90%), while MCNNFC greatly improves the classification accuracy of this category; in the land use types of class II and class III, the overall classification accuracy is more than 95%, and the classification results of each method are small difference. Some land use types, such as dense residential, medium residential and sparse residential are similar in structure and texture, resulting in low classification accuracy. Therefore, CNNs features cascade generally improves the classification accuracy of a single land use type, especially when the classification accuracy is low. However, for individual land use types, the classification accuracy decreases, indicating that there is also a problem of information hiding among cascade feature individuals. Therefore, method of cascading CNNs features fine-tuned can effectively improve the overall classification accuracy in small sample classification task.

4. CONCLUSION

For land use classification of high-resolution remote sensing images, the classification method of MCNNFC proposed in this paper draws the following conclusions through theoretical analysis and experimental comparison: Multi-structure convolution neural network features cascade classification method can compensate for the deficiencies of single convolution neural network information extraction. In the experiment, the overall classification accuracy of proposed method reaches 97.55%, which is 2%~5% higher than that of single convolution neural network. Fine-tuning the parameters of CNN FC layers can effectively improve the classification accuracy. The experimental results of fine-tuning CaffeNet, VGG-S and VGG-F show that the classification accuracy is improved by 3%~5%. Finally, in the selection of classifiers, mcODM has better classification effect than SVM classifier in small sample datasets, and the classification accuracy fluctuation is smaller in repeated classification. However, the proposed method also has some drawbacks, such as long running time, higher overall accuracy but lower accuracy of individual categories, which need further improvement in the follow-up study.

REFERENCES

Ban, Y., Jacob, A., Gamba, P., 2015. Spaceborne SAR data for global urban mapping at 30 m resolution using a robusturban extractor. *ISPRS J. Photogramm. Remote Sens.*, 103, 28–37.

Belward, A., Skøien, J.O., 2015. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS J. Photogramm. Remote Sens.*, 103, 115–128.

Castelluccio, M., Poggi, G., Sansone, C., 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. arXiv:1508.00092.

Chatfield, K. , Simonyan, K. , Vedaldi, A. , Zisserman, A., 2014. Return of the devil in the details: delving deep into convolutional nets. arXiv:1405.3531.

Chen, J., et al., 2015. Global land cover mapping at 30 m resolution: a POK-based operational approach. *ISPRS J. Photogramm. Remote Sens.*, 103, 7–27.

Cheng, G., Han, J., Lu, X.,2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. In *Computer Vision and Pattern Recognition,* 105(10), 1865-1883.

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Kluwer Academic Publishers.

Grauman, K., Darrell, T., 2005. Pyramid Match Kernels: Discriminative Classification with Sets of Image Features. In *International Conference on Computer Vision,* 1458-1465.

Hu, F., Xia, G.S., Hu, J., Zhang, L., 2015. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.*,7(11), 14680-14707.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675-678.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: *International Conference on Neural Information Processing Systems*, 84-90.

Lazebnik, S. , Schmid, C., Ponce, J., 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: *Computer Vision and Pattern Recognition Workshops*, 2169-2178.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.*, 60(2), 91-110.

Nogueira, K., Penatti, O.A.B., Santos, J.A.D., 2017. Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification. Pattern Recogn., 61, 539-556.

Penatti, O.A.B., Nogueira, K., Santos, J.A.D., 2015. Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains. In: *Computer Visionand Pattern Recognition Workshops*, 44-51.

Perronnin, F., Mensink, T., 2010. Improving the Fisher Kernel for Large-scale Image Classification. In *European Conference on Computer Vision*, 143-156.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

Vedaldi, A., Lenc, K., 2014. MatConvNet: Convolutional Neural Networks for MATLAB. 689-692.

Wang, J.J., Yang, J.C., Yu, K., et al., 2010. Locality-constrained Linear Coding for Image Classification. In C*onference on Computer Vision and Pattern Recognition*, 3360-3367.

Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., et al., 2017. Aid: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.*, 55(7) 1-17.

Yang, Y. , Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification.In *Sigspatial International Conference on Advances in Geographic Information Systems*, 270-279.

Zhang, T., Zhou, Z.H., 2016. Optimal Margin Distribution Machine.