Global Descriptors of Convolution Neural Networks for Remote Scene Images Classification

Q. Wang[1], Qian Ning[1,2(✉)], X. Yang[1], Bingcai Chen[3,4], Yinjie Lei[1], C. Zhao[1], T. Tang[1,2,3,4], and R. Hu[1,2,3,4]

[1] College of Electrical & Information Engineering,
Sichuan University, Chengdu 610065, China
ningq@scu.edu.cn

[2] School of Physics & Electronics,
Xinjiang Normal University, Urumqi 830054, China

[3] School of Computer Science & Technology at Dalian University
of Technology, Dalian 116024, China

[4] School of Computer Science & Technology at Xinjiang Normal University,
Urumqi 830054, China

Abstract. Nowadays, the deep learning-based methods have been widely used in the scene-level-based image classification. However, the features automatically obtained from the last fully connected (FC) layer of single CNN without any process have little effect because of high dimensionality. In this paper, we propose a simple enhancing scene-level feature description method for remote sensing scene classification. Firstly, the principal component analysis (PCA) transformation is adopted in our research for reducing redundant dimensionality. Secondly, a new method is used to fuse features obtained by PCA transformation. Finally, the random forest classifier applying to classification makes a significant effect on compressing the training procedure. The results of experiments on the public dataset describe that feature fusion with PCA transformation performs great classification effect. Moreover, compared with the classifier softmax, the random forest classifier outperforms the softmax classifier in the training procedure.

Keywords: Remote sensing image (RSI) · Global feature descriptors · Feature fusion · Scene classification

1 Introduction

More recently, in the field of RSI investigation, RSI scene classification [1] is one of the most important processes. For RSI classification, image semantic understanding is generally reflected by feature descriptors. Therefore, the key to classifying is features. In our research, we focus on the investigation of RSI feature

Q. Liang et al. (Eds.): *Artificial Intelligence in China*, LNEE 572, pp. 1–11, 2020.
https://doi.org/10.1007/978-981-15-0187-6_1

descriptors, premeditating the feature extraction as a critical step to obtain a great classification performance. Existing RSI features extraction research methods are primarily divided into local features extracting methods and global features extracting methods. For the former, there exist two types of methods, corner feature extraction methods and edge feature extraction methods. In the past, corner feature descriptor was the main job in feature extraction, such as features from accelerated segment test (FAST) [2], oriented fast and rotated brief (ORB) [2]. However, the curves of edge are discontinuous. Hence, they have limitations on the description of scene semantic information.

Numerous algorithms have been introduced for edge feature descriptor extraction. Signature of Histograms of Orientations (SHOT) [3] and Raster Operations Units (ROPS) [4] are based on the histogram statistics for local descriptors extraction, but they are low expression for semantic information in most of the experiments. In summarize, local descriptors extracting methods are considerably depending on the experiences of researchers, and higher requirements for extracted feature descriptors are necessary for these methods.

Among all kinds of methods to extract features, comprehensive feature descriptors can be obtained by deep learning-based methods, which help to enhance the ability of image description greatly. In 2012, Krizhevsky et al. [5] proposed a sensational deep learning neural network for image classification, which picks up the 2012 image recognition contest champion. From then on, the CNN architecture detonated the application boom of neural networks. Simultaneously, more deep CNN architectures were proposed after AlexNet [5], such as VGGNet [6], ResNet [7]. Due to the convenience of extracting features and the better result of classification, they are applied widely in many aspects of image recognition. Liu et al. [8] proposed to concatenate features extracted from convolutional layers of CaffeNet and VGG-VD16 to deep descriptors.

In this paper, we propose a method to accelerate the speed of RSI classification model training with significant performance increase. Specifically, we first obtain features from the last FC layer of two deep CNN models. Then we reduce the feature vector dimension utilizing PCA transformation. For accelerating the training speed and exploring a better effect on classification result, we propose an optimize and simple way to generate feature vector by concatenating the features from two types of different CNN models.

2 Proposed Method

The details of the proposed method are shown in Fig. 1. The main content consists of the following three parts: global descriptors extracting, feature fusion, and reducing dimensions. The procedure is organized as follows. The first part extracts global feature descriptors from the last fully connected layers, including a sample of the training set and the used pre-train CNN. Two types of pre-train CNN, VGGNet-16 and ResNet-50, are introduced to extract global feature descriptors. The second part is the details of the proposed method.

For the first step (Fig. 1a), the training set and testing set are used as the input of a CNN mode to extract the global feature descriptors from the last fully connected layer.

For the third step (Fig. 1b), the features of every input data obtained from the first part are concatenated to form the complete global feature descriptors. With a PCA transformation to reduce the dimension, the last classification results are given by the random forest classifier.

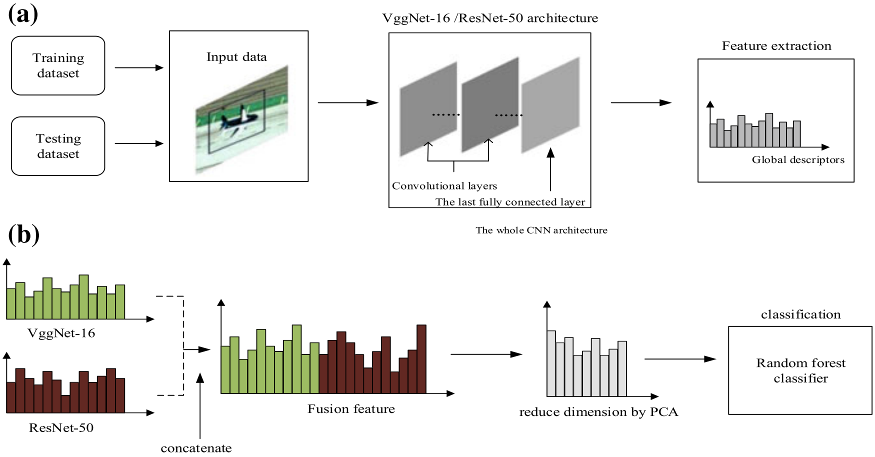

Fig. 1. Framework of proposed method. **a** Is the process of feature extraction. **b** Proposed method: concatenating features followed by PCA transformation.

2.1 Global Descriptors Extracting

In recent years, many CNN architectures have been proposed. Most of them perform a great effect on a large testing set. Such as VGGNet performs better on classification than AlexNet or CaffeNet. ResNet can obtain significant accuracy with deeper architecture and fewer parameters.At the first step, RSIs enter into VGGNet-16 and ResNet-50, respectively, for feature extracting, and the result is the global descriptor, which refers to the relationship between the extracted features and the entire image.

Feature extraction: A pre-train CNN mode serves as the feature extractor in many researches. As using the CNN as feature extractor, the minimal image shift has no effect on the last feature vectors because of the properties of convolution and pooling calculation. Hence, the obtained features have powerful fit abilities and make no influence on the classification result. In addition, because of this stability, it fits to every kind of image for feature extraction. When we apply a CNN for feature extraction, a popular feature extraction strategy is extracting an activation vector from the last fully connected layer (including the classifier layer) [9].

2.2 Feature Fusion and the Dimension Reduction

Existing deep learning pre-trained CNN models are used as the feature extractor to extract the feature from the final FC layers (include the last classification layer). Although the feature can be utilized to train the classifier directly, effective features extracted by only a single CNN pre-trained model are not enough, which will lead to a disappointing effect. Hence, an efficacious way to solve this question, feature fusion, is proposed.

Feature fusion: Deep feature fusion is a new solution to handle complex data. In 2014, two MIT engineers developed deep feature synthesis [10]. Most prediction decisions rely on the features descriptors based on input images in the vision classification tasks. Hence, it is necessary to overcome the obstacles of data dependence. Feature fusion refers to concatenating global features descriptors extracted from several different pre-trained CNN models. Vectors obtained by these models expand the dimension of the final vector by vector-spliced which is an efficient method to mitigate data dependence. In addition, the key advantage of feature fusion is new features obtained in this process, which can improve the performance of classification.

Reduce dimension: Dimensionality reduction, as the name implies, means feature selection and feature extraction. Since principal component analysis [11] was successfully proposed, applying PCA transformation to reduce dimension becomes a mainstream tendency. In the proposed method, PCA, as an important processing technology, is introduced for feature descriptors distinguishing and dimension reducing. With PCA more consummate, the field of data it can handle becomes wider. It is also the main technology for compressing the time of training process without losing the quality of a model. The main process for PCA is to transform the original data onto a set of linearly independent representations of each dimension through linear transformation, which reduces the dimension of input data set while maintaining the feature of the largest contribution of the data set in the data set. The final result is the key feature components of all the features.

3 Experiments and Analyses

3.1 UC Merced Land Use Dataset Description

In this section, we investigate the performance of the proposed methods on the "UC Merced Land Use" dataset[1] [12] extracted from large images from the US Geological Survey National Map Urban Area Imagery collection for various urban areas around the country. This dataset contains 21 classes. Each class includes 100 images with the size of 256 × 256 pixels in the color space of red–green–blue with different space structure, color distribute, region cover, and object cover. Every image is operated by rotating.

[1] http://vision.Ucmerced.edu/datasets/landuse.html.

3.2 Experimental Introduction

For input data, 75% images in each class serve as the training set and the remaining serve as the testing set. In experiments, the global feature descriptors are extracted from two pre-trained CNN models consisting of VGGNet-16 and ResNet-50, which are trained on the ImageNet dataset. The dimensions of global feature descriptors are 1×1000. The random forest is applied for training and classification. The confusion matrices performed the result of our proposed method; the training time is analyzed on the different methods to train.

3.3 Discuss Different CNN Models with PCA

The comparison between two types of CNN-based features classification is shown in Tables 1 and 2. The result of the classification on VGGNet-16-based and ResNet-50-based features without PCA transformation is displayed in Table 1 and the classification details of two types of CNN-based features with PCA transformation are performed in Table 2. The ratio of PCA transformation is set as 95%.The comparison in tables demonstrates that VGGNet-16 performs better classification effect. VGGNet-16-based features are better discrimination, especially in some similar categories, such as “beach,” “parking lot,” and “runway”, which is because the initial parameters in VGGNet-16 are much richer.

From two tables, we can see that using PCA transformation performs greatly for both types pre-trained CNN models, especially for the ResNet-50, which has improved about 8%. In general, PCA transformation help to improve the description ability of the global features.

3.4 Analysis of the Proposed Method

In this section, we research the confusion matrix on the 21-classes public remote dataset. Figure 2 describes the confusion matrix of proposed method, which includes feature fusion and PCA transformation. The confusion matrix is analyzed via using 25% of the training dataset. As confusion matrix is shown, the entry in the ith row and jth column means the rate of RSI belongs to the ith class and classifies to jth class. The classification results are proposed as percentages.

In Fig. 2, the average accuracies of classification for proposed method are 86.78%. It performs great ability (the accuracy of classification $\geq$90%) on the classes, such as “airplane,” “agricultural,” “baseballdiamond,” “chaparral,” “forest,” “golf course,” “harbor,” “overpass,” “tennis court,” “river.” The features extracted from these RSI include considerable information, which helps classify correctly.

Moreover, some classes obtain poor classification effect, such as “dense residential,” “medium residential.” This is the reason that high dimensional features of these classes are too similar to distinguish. In addition, several classes include plenty of building elements, which results an error decision.

Table 3 presents the comparison between state-of-art methods and our proposed method. These existing methods are detailed in [8,12–14]. As we can see

Table 1. Classification accuracies without PCA transformation. OA: overall average

Class	Airplane	Beach	Agricultural	Baseball-diamond	Buildings	Chaparral	Dense-residential	Forest	Freeway	Golfcourse	Harbor
ResNet-50	0.875	0.313	0.5	0.688	0.375	0.938	0.125	0.875	0.688	1.0	0.563
VggNet-16	1.0	0.688	0.75	0.438	0.625	1.0	0.5	1.0	0.813	0.875	0.813
Class	Intersection	Medium-residential	Mobile-homepark	Overpass	Parkinglot	River	Runway	Sparse-residential	Storage-tanks	Tennis-court	OA
ResNet-50	0.313	0.25	0.625	0.625	0.313	0.938	0.313	0.25	0.5	0.813	0.554
VggNet-16	0.5	0.563	0.5	0.688	0.635	0.75	0.375	0.625	0.625	0.813	0.693

Table 2. Classification accuracies with PCA transformation. OA:overall average

Class	Airplane	Beach	Agricultural	Baseball-diamond	Buildings	Chaparral	Dense-residential	Forest	Freeway	Golfcourse	Harbor
ResNet-50	1.0	0.438	0.438	0.25	0.563	0.875	0.188	1.0	0.625	0.875	0.875
VggNet-16	1.0	0.688	0.688	0.75	0.625	1.0	0.5	1.0	0.875	0.938	0.75
Class	Intersection	Medium-residential	Mobile-homepark	Overpass	Parkinglot	River	Runway	Sparse-residential	Storage-tanks	Tennis-court	OA
ResNet-50	0.438	0.75	0.688	0.688	0.325	0.688	0.625	0.313	0.563	0.875	0.639
VggNet-16	0.438	0.688	0.688	0.5	0.75	0.75	0.188	0.563	0.688	0.813	0.715

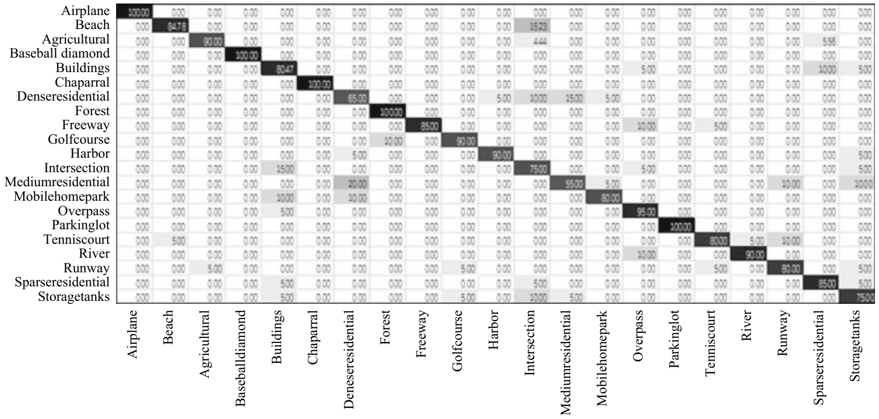

Fig. 2. Confusion matrix for 21-category performed result of the proposed method (86.78%)

that our proposed method enhances the classification effect 1.69% over the best-existed results in [8,12–14]. To sum up, our approach achieves a more favorable effect on this dataset because of the combination of feature fusion and PCA transformation.

Table 3. Overall classification accuracies on dataset

	Approaches	Overall accuracies (%)
State-of-art	BOVW [12]	76.81
	Texture [12]	76.91
	spck++ [14]	77.38
	Approach of [13]	75.33
	Strategy 1 of [8]	85.09
	Our method	**86.78**

4 Discussion of Time

In this part, the time spent on the training process with different classifiers is compared. As Table 4 shown, applying Caffe framework to train the CNN architectures, VGGNet-16 and ResNet-50, based on GPU for acceleration need 4376 and 2437 s, individually. However, our proposed method with random forest for classifying based on the 21-category dataset just needs 24.81 seconds, which shortens over one hundred times. In addition, the classification accuracies of our method are 86.78%, which is higher than VGGNet-16 and ResNet-50. Moreover, after analyzing, PCA transformation helps to reduce the time for training, too.

Table 4. Time for training a model on dataset

Classifier	Pretain model	Accuracy (%)	Training time (s)
Softmax	VGGNet-16	78.3	4376
	ResNet-50	82.4	2437s
Random forest (with PCA)	VGGNet-16	70.8	15.73
	ResNet-50	63.7	10.45
	Our method	86.78	24.81
Random forest (no PCA)	VGGNet-16	72.3	54.07
	ResNet-50	56.9	55.45
	Fusion-feature	85.24	52.00

In Fig. 3a, two curves display two situations about the proposed method with several numbers of trees in a random forest. With the number increasing, the spending time increases, too. The gray line demonstrates the features without PCA transformation, training spending is slower than the other situation. In addition, as Fig. 3b shown, while the dimension (For proposed method, corresponding to the PCA ratio 90–99%, the dimension number of PCA transformation is separately 104, 116, 130, 147, 168, 194, 227, 274, 346, 487) number of PCA transformation increasing, the training process gets longer, too. On the other hand, when the PCA transformation ratio obtains 95% (The dimension numbers is 194), the classification accuracy attains the highest, which is 86.78%. Overall, whatever for time spending or classification performance, our proposed method performs a better effect on the total datasets.

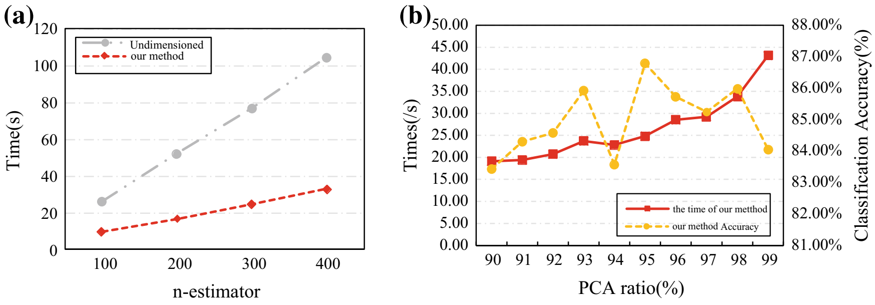

Fig. 3. **a** Shows time changing with different n_estimators; **b** shows the Pca transformation ratios influence the changing of training time and classification accuracy

5 Conclusion

In this paper, we investigate the global feature descriptors extracted from the last FC layer of CNN pre-trained models. Comparing to a single-model feature, the fusion features are more representative. Both pre-trained CNN-based features, VGGNet-16 and Resnet-50, are proposed to form the last global feature descriptors. The proposed method focuses on the most contribution features in both different CNN models. The result of experiment illustrates that feature fusion with a suitable dimension number of PCA transformation can enhance the performance of classification. Comprehensive evaluations of the public RSI scene classification perform the model training efficiency.

References

1. Bian X et al (2017) Fusing local and global features for high-resolution scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens 10(6):2889–2901
2. Rublee E et al (2011) ORB: an efficient alternative to SIFT or SURF. In: 2011 IEEE international conference on computer vision (ICCV). IEEE
3. Salti S, Tombari F, Di Stefano L (2014) SHOT: unique signatures of histograms for surface and texture description. Comput Vis Image Underst 125:251–264
4. Tombari, F, Salti S, Di Stefano L (2010) Unique signatures of histograms for local surface description. In: European conference on computer vision. Springer, Berlin, pp 356–369
5. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
6. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
8. Liu N et al (2018) Exploiting convolutional neural networks with deeply local description for remote sensing image classification. IEEE Access 6:11215–11228
9. Nogueira K, Penatti OAB, dos Santos JA (2017) Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn. 61:539–556
10. Kanter JM, Veeramachaneni K (2015) Deep feature synthesis: Towards automating data science endeavors. In: 2015 IEEE international conference on data science and advanced analytics (DSAA), 36678. IEEE, pp 1–10
11. Ke,Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004, vol 2. IEEE
12. Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. ACM, pp 270–279

13. Liu Q, Hang R, Song H, Li Z (2018) Learning multiscale deep features for high-resolution satellite image scene classification. IEEE Trans Geosci Remote Sens 56(1):117–126
14. Yang Y, Newsam S (2011) Spatial pyramid co-occurrence for image classification. In: 2011 IEEE international conference on computer vision (ICCV). IEEE, pp 1465–1472