Deep convolutional neural network structure design for remote sensing image scene classification based on transfer learning

Xiaoxia Zhang¹³, Yong Guo¹ and Xia Zhang²

¹College of geophysics, Chengdu University of Technology, Chengdu, China;
²Institute of intelligent manufacturing technology, Chongqing institute of Green and Intelligent Technology, Chinese Academy of Science, Chongqing, China

³E-mail: 2017010117@stu.cdut.edu.cn

Abstract. To obtain an ideal scene classification effect when applying a deep convolutional neural network (DCNN) to remote sensing images, a DCNN named ConNet-3F based on transfer learning was constructed. Firstly, the initial 13 layers of VGG16 were used as the main architecture, and then 6 layers were added to the end of it to extract deeper features. The pruned model of VGG16 acted as the feature extractor to extract features from remote sensing images. Subsequently, the parameters trained on the ImageNet were set as initialization. Finally, parameters of all layers were trained on the remote sensing data to obtain the favorable model. The results on the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset demonstrate that this method achieves higher classification accuracy compared with other mentioned methods. Especially, even if the ratio of training data was reduced on the NWPU-RESISC45 dataset, classification accuracies of the proposed method always stay above 90%.

1. Introduction
It was found deep learning (DL) has obtained great success in lots of areas, such as natural language processing, computer vision, speech recognition, especially in the field of nature image recognition. Different from the rapid development of natural image processing research with the method of DL, remote sensing image scene classification has encountered some bottlenecks mainly due to the insufficient of labeled training data, because low resolution and great loss of information caused by noise lead to time-consuming and requiring professional knowledge to annotate training data manually. Therefore, in the domain of remote sensing, the current researchers devote to transfer learning, the common practice is to transfer the pre-trained model and keep its parameters unchanged [1], or fine-tune the parameters of the last few layers [2]. It is an ideal selection of transfer learning for solving the problem of small samples and making knowledge learned from nature image datasets transfer to remote sensing data [3]. When the source domain is not identical to the target, it is necessary to adjust all the parameters in the training process. On the other hand, training a relatively deeper network is more likely to extract rich hierarchical features of remote sensing images, which in turn needs sufficient labeled data to obtain the favorable result.

Concerning the dilemma mentioned above, we addressed a strategy through transferring learning and designed a relatively deeper network that could learn more high-level feature representations in this work. Because previous studies have demonstrated that the transferring results have better
performance than other approaches on remote sensing datasets [4,5]. In the previous survey papers [6,7], the authors proposed transfer parameters and structures to train networks and evaluated them on the remote sensing data, the trained classification models based on transfer learning achieve more accuracy. Whereas these transfer learning methods keep the parameters of the first few layers unchanged or simply change the top layer but no other changes, they may be inappropriate to apply to remote sensing image classification.

The remote sensing image has the character of relatively low resolution and low signal-to-noise ratio, the information contained in it is far less than optical images. Therefore, it is impractical to train a deep convolutional neural network as many layers as optical images. For this reason, a network structure named ConNet-3F was proposed in this work, which adopted the pruned VGG16 as the main structure for transferring learning, and then a depth-wise separable convolutional layer [8], a max-pooling layer, a flatten layer and three fully connected layers were added to it. This is precisely the main direction of this paper.

The remainder of this paper is organized as follows. Section 2 gives an introduction of the experimental datasets used to train and test the model. Section 3 provides a review of the proposed ConNet-3F. Experimental results on the two datasets are presented in Section 4, and Section 5 provides a brief conclusion of this paper.

2. Experimental datasets

![Figure 1. Ten types of target examples: optical image(top) and SAR image(bottom).](image)

Table 1. Number of training and testing images of MSTAR.

| Class(abbreviation) | Depression | Number | Depression | Number |
|---------------------|------------|--------|------------|--------|
| 2S1(2S)             | 17°        | 299    | 15°        | 274    |
| BMP2(BM)            | 17°        | 232    | 15°        | 196    |
| BRDM2(BR)           | 17°        | 298    | 15°        | 274    |
| BTR60(R6)           | 17°        | 256    | 15°        | 195    |
| BTR70(R7)           | 17°        | 233    | 15°        | 196    |
| D7(D7)              | 17°        | 299    | 15°        | 274    |
| T62(T6)             | 17°        | 299    | 15°        | 273    |
| T72(T7)             | 17°        | 232    | 15°        | 196    |
| ZIL131(ZI)          | 17°        | 299    | 15°        | 274    |
| ZSU234(ZS)          | 17°        | 299    | 15°        | 274    |
MSTAR dataset. In this work, the first dataset applied to test our network is the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [9]. Ten types of target images at 17° depression angle are for training and images at 15° depression angle for testing. Ten types of examples of SAR images and their corresponding optical images are shown in Figure 1, the number of training and testing images are depicted in Table 1. To facilitate drawing, abbreviation of the names for the ten classes are also depicted in Table 1. For the aim of checking whether the model meets desired accuracy, the training data is divided into two parts: 75% of the data for training and the other 25% for validation.

NWPU-RESISC45 dataset. The NWPU-RESISC45 dataset created by Northwestern Polytechnical University (NWPU) is applied to test the reliability of CovNet-3F when the ratio of training data decreased [10]. This dataset contains 31,500 images, covering 45 scene classes with 700 images in each class. Each image has a size of 256*256 pixels in red-green-blue (RGB) color space. 6 types with low discrimination among them are applied to the experiment, which are airport (AIR), bridge (BRI), freeway (FRE), overpass (OVE), railway (RAI), runway (RUN). Six types of examples of remote sensing images are shown in Figure 2. The training data is split into training and validation. Besides, 100 images for each category are set aside to test the proposed model.

![Figure 2. Six types of scene examples.](image)

Training a desired network needs a large number of training samples, however, in the remote sensing image scene classification task, the training samples are too insufficient to train a favorable model. Fortunately, a large number of samples are often produced by data augmentation. In this work, data augmentation methods that are suitable for remote sensing image classification are adopted, these augmentation methods include rotation, horizontal flip, shear range and scaling. Besides, each pixel of the image is divided by 255 before training, which provides a more efficient way of data processing and improves the convergence rate of the proposed model.

3. Methods

3.1. Structure of ConNet-3F
The proposed ConNet-3F network structure is straightforward. Core of the network is the feature extraction part, which is transferred from VGG16. The auxiliary part is added to the end of the core network and it contains a depth-wise separable convolutional layer, a max-pooling layer, a flatten layer and three fully connected layers. Because the depth-wise separable convolution can separate spatial and channel feature learning, it has the ability to use less data to learn more feature representations. Fully connected layers contribute to improve nonlinear expression ability of the model. Besides, the parameters of $W$ in the fully connected lays are punished by the method of L2 regularization, which reduces the complexity of the network. The main function of the feature extraction model is to obtain multi-scale feature maps from the bottom to top level convolutional layers. Therefore, the initial 13 layers of VGG16 as a whole is treated as the feature extraction model, which removes the top 3 layers, then six layers are added to the tail of it. It is a favorable architecture
for remote sensing image scene classification demonstrated by experiments. The structure of the proposed ConNet-3F is shown in Figure 3.

**Figure 3.** Structure of ConNet-3F used for remote sensing image classification (‘Conv’ represents the number of features maps, ‘@’ denotes the filter size).

### 3.2. Mainly idea of ConNet-3F

CNN is one of the most effective data processing methods, it is characterized by few computational parameters, mainly due to locally connected and weight-sharing mechanism. A complete CNN architecture is composed of convolutional layers, pooling layers and fully connected layers. Convolutional layer is the most applied among them.

Two important elements one kernels (or filters) and the other biases are in the convolutional layers. The kernel contains a small receptive field and is trained to extract features from an image. The relationship of convolutional layers between $l$ and $(l+1)$ in the forward propagation can be expressed as follows.

\[
 z^{(l+1)} = W^{(l+1)}a^{(l)} + b^{(l+1)} \\ a^{(l+1)} = f(z^{(l+1)})
\]

(1)

(2)

Where, $z^{(l+1)}$ is the intermediate result of the convolutional layer $(l+1)$, $a^{(l)}$ represents the activation value of layer $l$, $W^{(l+1)}$ and $b^{(l+1)}$ are weights and bias vector of layer $(l+1)$, $f(\cdot)$ is the activation function, activation function of Relu is applied in this work.

The neural network is trained with the cross entropy loss function, which can be represented as $J(W, b) \in \mathbb{R}$, then the residual is expressed as follows.

\[
 \delta^{(l)} = \frac{\partial J(W, b)}{\partial z^{(l)}} = \frac{\partial J(W, b)}{\partial z^{(l+1)}} \frac{\partial z^{(l+1)}}{\partial a^{(l)}} \frac{\partial a^{(l)}}{\partial z^{(l)}} = ((W^{(l+1)})^T \delta^{(l+1)}) \cdot f'(z^{(l)})
\]

(3)

Where, $\frac{\partial J(W, b)}{\partial z^{(l+1)}} \frac{\partial z^{(l+1)}}{\partial a^{(l)}} = (W^{(l+1)})^T \delta^{(l+1)}$, $\frac{\partial a^{(l)}}{\partial z^{(l)}} = f'(z^{(l)})$. According to the dimensional compatibility principle, $(W^{(l+1)})^T$ is the adjustment of $W^{(l+1)}$ adding the transpose. From equation (3), we can deduce the residual of each layer recursively. With the chain rule, $\nabla_{W(0)}J(W, b)$ and $\nabla_{b(0)}J(W, b)$ are defined in equation (4) and (5).

\[
 \nabla_{W(0)}J(W, b) = \frac{\partial J(W, b)}{\partial z^{(0)}} \frac{\partial z^{(0)}}{\partial W^{(0)}} = \delta^{(0)}(a^{(l-1)})^T
\]

(4)

\[
 \nabla_{b(0)}J(W, b) = \frac{\partial J(W, b)}{\partial z^{(0)}} \frac{\partial z^{(0)}}{\partial b^{(0)}} = \delta^{(0)}
\]

(5)
where, \( \frac{\partial f(W, b)}{\partial z(l)} = \delta(l) \), \( \frac{\partial z(l)}{\partial b(l)} = 1 \). The iterative formula of the gradient descent is expressed in equation (6), which can be derived from the Taylor’s expansion.

\[
x_{k+1} = x_k - \eta \cdot \nabla f(x_k), (\eta > 0)
\]  \( \text{(6)} \)

Then the iterative formulas of gradient descent for the parameters of \( W \) and \( b \) are given in equation (7) and (8). \( \eta \) represents the learning rate.

\[
w_{k+1} = w_k - \eta \cdot \nabla_w J(W, b)
\]  \( \text{(7)} \)

\[
b_{k+1} = b_k - \eta \cdot \nabla_b J(W, b)
\]  \( \text{(8)} \)

In the training process, the classification performance of the network is optimized based on the iteratively updating parameters of \( W \) and \( b \). Besides, stochastic gradient descent (SGD) with Nesterov momentum is used. The gradient descent method with momentum can accelerate the training process. Designing a reasonable initialization scheme and SGD with momentum will accelerate the convergence of the training process and achieve an ideal training effect. The updated formulas of momentum gradient descent parameters are as follows.

\[
V_{i+1} = \gamma V_i - \alpha V_i (W_i + \gamma V_i)
\]  \( \text{(9)} \)

\[
W_{i+1} = W_i + V_{i+1}
\]  \( \text{(10)} \)

where, \( V_i \) represents the gradient descent velocity of the \( i \)-th iteration, \( \gamma \) is the coefficient, it is set as 0.9. \( \alpha \) denotes the learning rate, \( W_i \) represents the result of the \( i \)-th iteration.

It is common to observe the validation accuracy improves steadily during the initial training phase, and it gets worse in the end. Early stopping contributes to mitigating overtraining, it also can be considered as a regularization method to avoid overfitting. Early stopping aims to solve the problem that the number of epochs needs to be set manually. The fundamental reason is that accuracy will reduce as the training continues. The general practice is to record the best coefficient values during training, when the patience is 10 epochs or more successive epochs not reaching the best accuracy, the accuracy can be regarded as no longer improved and should stop the training process.

Natural images and remote sensing images share some low-level features, such as curves and edges, thus we can use the parameters trained from ImageNet as initial values. In this work, pre-training refers to using the pruned classical model of VGG16 and its parameters trained on the ImageNet as initialization. Then 6 layers were added to the end of the pruned model to complete feature extraction and scene classification. The combined model was used to train on the MSTAR and NWPU-RESISC45 dataset aforementioned. When using the modified model to train new data, it is necessary to train the parameters of all layers. For the problem of insufficient data, pre-training weights as initialization and fine-tuning not only accelerate the convergence speed of the target task, but also reduce overfitting to a certain extent.

Combining the training strategies aforementioned on the two standard databases, this paper aims to improve the performance of remote sensing image scene classification.

4. Experiments on the two datasets
The network is implemented by Keras with the back end of Tensorflow. The software is run on a desktop with Intel I5-9400F CPU. The Nvidia GeForce GTX 1660 GPU is supported by compute unified device architecture (CUDA) 10.0.

4.1. Experiment 1: classification results on MSTAR
In this section, the experiments are carried out to evaluate the proposed method explained above. The training and validation accuracies of the proposed ConNet-3F are shown in Figure 4.
Figure 4. Training and validation accuracies of the proposed ConNet-3F.

The horizontal and the vertical axes in Figure 4 represent the number of epochs and the accuracy of training and validation, respectively. The blue dashed line is training accuracy and the orange solid line denotes the validation accuracy. As the number of the epoch increases, the proposed network learns an ideal feature representation, leading the classification accuracy to improve and converge to the almost stable state. When applied to the test data, the network reaches an average classification accuracy of 99.58%. The confusion matrix of this experiment is shown in Figure 5.

Figure 5. Confusion matrix on the MSTAR dataset.

The proposed network and the training methods can achieve desirable performance as shown in Figure 5, especially classification accuracies of the ZIL131 and ZSU234 reach 100%. To verify the classification effect of the proposed network and training strategies, our method was compared with several paradigms based on the same dataset, such as VGG-S1 [11], CHU-Net [12], SVM [13], CNN [14], WKCNN [15] and CNN-SVM [16]. As shown in Table 2, our proposed method produces the highest average accuracy rates among the seven methods, including several classical methods. Figure 6 clarifies the overall accuracies of the ten classes.

Table 2. The performance comparison between ConNet-3F and the other six methods on the MSTAR dataset.

| Method     | VGG-S1 | CHU-Net | SVM  | CNN  | WKCNN | CNN-SVM | ConNet-3F |
|------------|--------|---------|------|------|-------|---------|-----------|
| Accuracy(%)| 98.9   | 99.09   | 90   | 92.3 | 98.39 | 99.5    | 99.58     |
In our work, the labeled images were chosen as source data to train the network for the remote sensing image classification task. As depicted in Table 2, the proposed network has some advantages over other methods including several classical methods. With an average accuracy of 99.58%, our network outperforms all the reference methods. Especially compared with the performance of VGG-S1, which modifies from the VGG-S and it has a similar structure with VGG16, ConNet-3F outperforms 0.68% than it. Figure 6 shows the performance of the proposed method and several paradigms (including VGG-S1, CHU-Net, WKCNN) on 10 types of targets classification problem, it can be seen that the overall accuracy of the proposed method is better than the other mentioned methods on the MSTAR dataset.

4.2. Experiment2: classification results on NWPU-RESISC45

In this part, 6 indistinguishable classes in the NWPU-RESISC45 dataset were classified by changing the ratio of training data. The ratios of training samples were set as 0.8, 0.7, 0.6 and 0.5 respectively. The experiments were taken under the same condition as Experiment 1. With respect to check whether the model met the favorable accuracy, the training data was also divided into two parts: training and validation. The classification accuracies on the test data are shown in Table 3. It is seen that the classification accuracy decreases with the ratio of training samples, other listed methods have the similar trend. More importantly, even the ratio of training data decreases from 0.8 to 0.5, the classification accuracy of the proposed method stays above 90%. Figure 7 shows confusion matrices of the proposed method under the training ratio of 0.8, 0.7, 0.6, and 0.5, respectively.

| Table 3. Average classification accuracy when changing the ratio of training data. |
|---------------------------------|---------|---------|---------|---------|
| Training Ratio                  | 0.8     | 0.7     | 0.6     | 0.5     |
| ConNet-3F                       | 95.17%  | 94.83%  | 94.3%   | 93.5%   |
| GoogLeNet[17]                   | 94.3%   | -       | -       | 92.7%   |
| unique metric learning[18]      | 59.33%  | -       | -       | 58.97%  |
| metric learning[18]             | 84.6%   | -       | -       | 82.4%   |

Figure 6. Accuracies of different methods on the MSTAR dataset.
As shown in Figure 7, it is observed that the average accuracy decreases along with the training ratio. In general, the classification accuracy of runway is the lowest among the 6 classes, which is indistinguishable from the airport image. Since choosing 6 classes of the most confusing images, it is inspiring to see that the classification accuracy always stays above 90%, which proves the training strategies and the proposed model are reliable when the ratio of training data decreased.

5. Conclusion

In order to overcome the difficulty of training a deep convolutional neural network resulting from scare labeled remote sensing images, a network named ConNet-3F was proposed, which used the initial 13 layers of VGG16 as the feature extraction model, then 6 layers were added to the end of the pruned VGG16 to enhance its nonlinear expression ability. Parameters trained on the ImageNet were set as initialization of the proposed model, the parameters of all layers were trained from initial training, several data augmentation methods were also used in the training process. Experiments on the MSTAR datasets achieve better performance on the overall results than that related methods, including some classical methods. Besides, our model is reliable when the ratio of training data decreased on the NWPU-RESISC45 dataset, the classification accuracy of the proposed method always stays above 90%. From the results of the two experiments, our method of DCNN contributes to balancing insufficient data and classification accuracy.

References

[1] Zhong C, Ting Z, Chao O 2018 End-to-End Airplane Detection Using Transfer Learning in Remote Sensing Images Remote Sensing 10(1) 139
[2] Chen S, et al. 2016 Target classification using the deep convolutional networks for SAR images IEEE Transactions on Geoscience and Remote Sensing 54(8) 4806-4817
[3] Yosinski J, et al 2014 How transferable are features in deep neural networks? 3320-3328
[4] Rostami M, Kolouri S, Eaton E, et al 2019 Deep Transfer Learning for Few-Shot SAR Image Classification Remote Sensing 11(11) 1374
[5] Fan H, Xia G S, et al 2015 Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery Remote Sensing 7(11) 14680-14707
[6] Shang R H, et al 2018 SAR Targets Classification Based on Deep Memory Convolution Neural Networks and Transfer Parameters IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(8) 2843-2846
[7] Wang Y Y, et al 2018 Ship Classification in High-Resolution SAR Images Using Deep Learning of Small Datasets Sensors 18(9) 2929
[8] Chollet F 2017 Xception: Deeplearning with depthwise separable convolutions Proceedings of the IEEE conference on computer vision and pattern recognition 1251-1258
[9] Ross T D, et al 1998 Standard STR ATR evaluation experiments using the MSTAR public release data set Algorithms for Synthetic Aperture Radar Imagery 566-573
[10] Cheng G, et al 2017 Remote Sensing Image Scene Classification: Benchmark and State of the Art Proceedings of the IEEE 105(10) 1865-1883
[11] Yosinski J, et al 2014 How transferable are features in deep neural networks? Advances in neural information processing systems 3320-3328
[12] Cui Z, et al. 2018 SAR unlabeled target recognition based on updating CNN with assistant decision IEEE Geoscience and Remote Sensing Letters 15(10) 1585-1589
[13] Srinivas U, et al 2014 SAR automatic target recognition using discriminative graphical models IEEE transactions on aerospace and electronic systems 50(1) 591-606
[14] Song H, et al. 2016 Sparse representation-based SAR image target classification on the 10-class MSTAR data set Applied Science 6(1) 26
[15] Tian Z, et al 2018 Classification via weighted kernel CNN: application to SAR target recognition International journal of remote sensing 39(23) 9249-9268
[16] Wagner S A 2016 SAR ATR by a combination of convolutional neural network and support vector machines IEEE transactions on Aerospace and Electronic Systems 52(6) 2861-2872
[17] Yang Y, Newsam S 2010 Bag-of-visual-words and spatial extensions for land-use classification 270-279
[18] Goel A, Banerjee B, Pizurica A 2018 Hierarchical Metric Learning for Optical Remote Sensing Scene Categorization IEEE Geoscience & Remote Sensing Letters 1-5