Remote Sensing Image Scene Classification along the High-speed Railway based on Convolutional Neural Network

Xunqiang Gong¹,²,*, Xiayi Ju¹, Kuan Qian¹, Tieding Lu¹ and Zhiping Chen¹

¹Faculty of Geomatics, East China University of Technology, Nanchang, China
²Research Center for Ecological Civilization Construction System of Jiangxi Province, East China University of Technology, Nanchang, China
* Corresponding Author: xqgong1988@163.com

Abstract - China's high-speed railway operation network is huge, and the environment along the line is complex. Therefore, there are many security risks. For that reason, real-time dynamic monitoring of hidden dangers along the high-speed railway is required. Traditional manual field monitoring methods are inefficient, while remote sensing technology can provide a means for dynamic monitoring because of its large-scale, real-time and cyclical aspects. In addition, feature extraction of classification model is simple and classification accuracy is not ideal by using traditional scene classification methods. In order to effectively avoid potential safety hazards along the high-speed railway, in other words, to classify the ground objects along the high-speed railway more accurately, the remote sensing images along the high-speed railway obtained from Google Earth are taken as the research object, and three convolutional neural network (CNN) models for scene classification are constructed and their classification results are compared in this paper. The experimental results show that the classification accuracy of the three models is above 89%, and the results of the ResNet model are optimal.

1. Introduction

With the rapid development of remote sensing technology, the data volume of remote sensing images is increasing sharply, and the resolution is continuously improved. In this case, how to fully extract and mine the valuable information from the numerous remote sensing images is urgently to be explored [1,2]. Scene classification is an important means of remote sensing image interpretation and information extraction, which refers to the identification of images that possess similar scene characteristics from multiple images through a large number of computer learning processes [3]. It has important application value in the fields such as land use/land cover, disaster monitoring, vegetation distinction, and tree species identification [4,5].

Remote sensing scene image features are diverse, including not only low-level visual features with poor generalization, but also many middle-level semantic features with rich information, good stability, and strong discriminative ability. In order to mine high-level semantic information from high-resolution remote sensing images and overcome the problem of semantic gap between the low-level features and the high-level semantic concepts, scene classification technology has caused a lot of research by scholars in the field of aerial and satellite image analysis. Remote sensing image scene classification first extracts the features of image blocks, and then uses classifiers to classify. The classification method can use traditional pixel-based or object-oriented classifiers, which mainly include maximum likelihood classification, random forest, neural network, support vector machine,
and sparse representation based classification [6-8]. In order to fuse the different types of features with different characteristics, a semantic allocation level (SAL) multifeature fusion strategy based on PTM (SAL-PTM) and a Dirichlet-derived multiple topic model (DMTM) for high spatial resolution imagery have been proposed [9,10].

Convolutional Neural Network (CNN) employs a multi-stage learning structure with global feature to adaptively learn image features. Compared with low-level visual features and middle-level semantic features, it can learn more abstract semantic features through a large amount of data, which can avoid the blindness of artificial feature design effectively [11,12]. In recent years, remote sensing image scene classification based on CNN has attracted extensive attention [13,14]. However, the application of this model in remote sensing image scene classification along the high-speed railway has not been systematically studied, in spite of the wide application of CNN in the remote sensing community. To this end, three commonly used CNN models (i.e. AlexNet model, VGG-16 model and ResNet model) are constructed to classify remote sensing image scenes along the high-speed railway in this paper. A comparison with the three commonly used CNN models is conducted in remote sensing image scene classification along the high-speed railway and one set of dataset and four reasonable evaluation indexes are used in this evaluation.

This paper is organized as follows: In Section 2, three commonly used CNN models are introduced. Experiments with dataset, indexes and procedures for comparative evaluation are conducted in Section 3, and the description of the results and analysis are given in Section 4. Finally, some conclusions are made in Section 5.

2. Three Commonly Used Convolutional Neural Network Models

Convolutional Neural Network (CNN) is a multi-layer neural network, generally composed of an input part, a feature extraction part which composed of a convolutional layer and a pooling layer as well as a classifier formed by a fully connected layer. The convolutional layer and the pooling layer are the core modules for the implement of CNN feature extraction and selection functions. The convolutional layer obtains the feature map by sliding the convolution kernel on the input data so as to extract features. The pooling layer averages the values within the limited range of the pooling core and then to perform sampling to greatly reduce the feature dimension and avoid over-fitting effectively. The image feature reflection in CNN is a forward propagation process in which the output of the previous layer serves as the input of the current layer. The output of layer \( l \) can be expressed as follows [15]

\[
x' = f(Wx^{l-1} + b')
\]

where \( l \) is the number of layers, \( x \) is the feature map, \( W \) is the reflecting weight matrix of the current network layer, \( b \) is the additive bias term of the current network, and \( f \) is the activation function.

2.1. AlexNet Model

The main advantages of the AlexNet model compared to the previous neural network model are the use of the ReLU activation function and the DropOut method. On the one hand, ReLU can effectively improve the gradient disappears, and it only needs one threshold to obtain the network activation value, which can accelerate the convergence of stochastic gradient descent. On the other hand, by introducing the DropOut layer with multiple weight combinations after each fully connected layer can reduce the over-fitting problem of the model. The Dropout layer controls the activation state of neurons according to a certain probability through a threshold [16]. This structure significantly reduces the complex mutual adaptation relationship between neurons, thereby ensuring that the features extracted by the network model are independent to each other. The AlexNet model designed in this paper is shown in Fig. 1. It contains three 3×3 convolutional layers and three fully connected layers of 128 nodes, 128 nodes, and 7 nodes, respectively. ReLu activation function is used after the convolutional layer to solve the problem of gradient disappear. After each convolutional layer, the maximum pooling technique is used, and the 0.5 Dropout method is used in the fully connected layer after the pooling
layer to prevent over-fitting.

2.2. **VGG-16 Model**

Compared with the AlexNet model, the VGG-16 model has two obvious changes. One is to use multiple consecutive 3×3 convolution kernels to replace the original 5×5 and 7×7 convolution kernels of the AlexNet model, and the convolution layers use very small receptive fields. The other one is that the VGG-16 model has multiple convolutional layers, which far exceeds the AlexNet model in terms of model depth. This allows the CNN model to extract more accurate and deep features, thereby improving the classification accuracy of the network model. The VGG-16 model used in this paper is shown in Fig. 2. The input is a 100×100 RGB image, and 5 convolution pooling groups are used, including 2 consecutive two-layers and 2 consecutive three-layer convolutional layer groups, respectively. All the convolutional layers use a 3×3 convolution kernel. After each convolutional layer group, there is a pooling layer. The pooling layer uses a 2×2 pooling window with step size of 2. The 5 convolutional pooling groups are followed by 3 fully connected layers of 128 nodes, 128 nodes and 7 nodes, respectively. The last fully connected layer is the output layer, corresponding to the 7 categories of model output [17].

![Fig. 1 AlexNet model](image1)

![Fig. 2 VGG-16 model](image2)

2.3. **ResNet Model**

The ResNet model is an improvement of the VGG-16 model. By simply superimposing convolutional layers to increase the depth of the network model, the network degradation becomes more and more obvious. In order to solve this problem, the ResNet model designed a residual module [18]. As shown in Figure 3, assuming that \( x \) is the input of the network, the hidden layer of the deep network can be formally represented as \( H(x) \), and the hidden layer output of the residual block is represented as \( H(x) = H(x) + x \), that is, the network degradation problem can be solved by adding identity transformation. The residual module is realized by shortcut connection, so that the stacked layers are suitable for a residual reflecting and the input and output elements of multiple convolutional layers are cascaded through the shortcut connection for neuron intelligent superposition. This approach not only does not add additional parameters and calculations to the network model, but also accelerates the training speed of the network model. The ResNet model used in this paper is shown in Fig. 4. The input is a 100×100 RGB image, and 22 convolution groups are used, each of which is composed of 3 gradually increasing feature layers, adopts the structure of the residual module, and finally connects the pooling layer and the fully connected layer. In Fig. 4, \( H \times \) means that the block in the virtual
frame is stacked \( n \) times as a unit.

Fig. 3 Residual module  
Fig. 4 ResNet model

3. Experimental Design

Experimental design includes three parts: introduction of experimental dataset, evaluation indexes as well as experimental procedures. By constructing an experimental dataset of high-resolution remote sensing images of a certain section of the Beijing-Shanghai high-speed railway, and selecting reasonable evaluation indexes, the above three CNN models are used in remote sensing scene classification.

3.1. Experimental Dataset

High-resolution remote sensing images along a certain section of the Beijing-Shanghai high-speed railway from Suzhou to Bengbu (33.32°N, 117.31°E) are selected as experimental dataset. The image is a 3-band image after fusion processing, with a spatial resolution of 1 meter, and the shooting time is January 15, 2016. The experimental dataset contains seven types of land use/land cover scenes, i.e. road, building, factory, bare ground, agricultural, water, and vegetation [19]. There are 500 images of 100×100×3 in each category, and the scene example of each category is shown in Fig. 5.

3.2. Evaluation Indexes

Currently, the main evaluation indexes used are overall accuracy, precision, recall and F1-score [20]. The evaluation parameters of the scene classification results along the high-speed railway are: \( T_p \), the number of correctly recognized images; \( F_r \), the number of incorrectly recognized images; \( r_s \), the number of unrecognized images.

1) Overall Accuracy (OA) refers to the sum of correctly classified images divided by the total number of images. The formula for overall accuracy is

\[
OA = \frac{T_p}{T_p + F_r + F_s}
\]

2) Precision means the probability that all samples that are predicted to be positive are actually positive samples, its meaning is for its prediction results, and the formula is

\[
P = \frac{T_p}{T_p + F_r}
\]

3) Recall, its meaning is for its original sample, which means the probability of being predicted as a positive sample in the actual positive sample, its formula is

\[
R = \frac{T_p}{T_p + F_s}
\]

4) F1-score is an index that comprehensively considers precision and recall. It is the weighted average of precision and recall. The formula is

\[
F1 = \frac{2 \times P \times R}{P + R}
\]
3.3. Experimental procedures

This experiment is a third-party open source deep learning framework keras based on the python interface. In the experiment, the initialization parameter settings of the three CNN models are shown in Table 1.

| CNN        | Iteration | Batch size | Learning rate | Weight decay | Momentum |
|------------|-----------|------------|---------------|--------------|----------|
| AlexNet    | 100       | 32         | 0.001         | 0.0005       | 0.9      |
| VGGNet-16  | 100       | 32         | 0.001         | 0.0005       | 0.9      |
| ResNet     | 100       | 32         | 0.001         | 0.0005       | 0.9      |

The main steps of this experiment are as follows.

1) Use the GDAL library to crop the remote sensing images along the high-speed railway into 100×100×3 images, and manually mark them to generate sample datasets.

2) Normalization was made for the images, and then performing transformations including rotation, scaling, etc. on the normalized image data so as to realize the supplement and expansion of training samples.

3) Construct the AlexNet model, VGG-16 model and ResNet model respectively.

4) Taking 80% of the sample dataset as the training sample, using the stochastic method to initialize the network parameters, and setting the network learning rate, weight decay value, momentum parameter, etc., obtaining the entire model parameters through back-propagation and loss function iteration.

5) Use the remaining 20% sample dataset as a test set for testing to verify the classification performance of the model.

4. Experimental Results and Analyses

The above experimental dataset is adopt, and the AlexNet model, VGG-16 model, and ResNet model respectively are used to classify remote sensing scene images along the high-speed railway. Then the four indexes, i.e. overall accuracy, precision, recall and F1-score are used to evaluate the classification results of these three CNN models. 80% of the images in each type of category are selected as the training samples, and the remaining 20% are the test samples to be classified. Then we use AlexNet, VGG-16, and ResNet models respectively to classify remote sensing scene images, and a five-fold cross validation method is used to compare the classification results of these three models. The overall accuracy curves using the three CNN models of train sets and validation sets are shown in Fig. 6. It can be found that the overall accuracy curve gradually stabilizes after 10 iterations in the training sets of the three CNN models, indicating that the convergence speed is relatively fast, and it can be determined that the CNN model has effectively learned the dataset. When the number of iterations
reaches 100, the overall accuracy curve has converged significantly.

(a) AlexNet  (b) VGG-16  (c) ResNet

Fig. 6 Overall accuracy curves of training sets and verification sets using three CNN models

The results of the overall accuracy of the three CNN models are shown in Table 2. According to the data in Table 2, the overall accuracy of the three CNN models is above 89%. This is because the CNN model as a classifier with a deep structure can mine the complex features hidden in remote sensing images and extract richer semantic features of ground objects. The overall accuracy obtained by Alex-Net model is 89.5%, but the overall accuracy of VGG-16 model is 92.1%. It is clear that the overall accuracy of the VGG-16 model is obviously better than that of the Alex-Net model. Among them, ResNet model has the highest overall accuracy, reaching 93.4%, which is higher than those for the two previous models. It can be seen that ResNet model achieves higher overall accuracy (i.e. 3.9%) than Alex-Net model and slightly higher overall accuracy (i.e. 1.3%) than VGG-16 model. Therefore, the ResNet model has better classification performance than the other models in remote sensing image scene classification along the high-speed railway.

The precision, recall and F1-score obtained by using the three CNN models (i.e. AlexNet model, VGG-16 model and ResNet model) are shown in Table 3, Table 4 and Table 5, respectively. It can be seen from Table 3 that the precision using the AlexNet model of bare ground is poor (only 0.73), but that of building is the best, reaching 0.92. The remaining are above 0.8. The recall of vegetation is the best (i.e. 1), and the other six are 0.77, 0.83, 0.93, 0.84, 0.73 and 0.71, respectively. The F1-score of different categories is between 0.77 and 0.90. From the data in Table 3, the average values of precision, recall and F1-score using the AlexNet model of the seven categories are 0.84, 0.83, and 0.83, respectively. The VGG-16 model has the lowest precision for water classification (only 0.78), while the classification results for factory and agricultural are better, with precision of 0.91 and 0.95 respectively from Table 4. It can be also seen that recall and F1-score of different categories are between 0.76 and 0.91. From the data in Table 5, the precision using the ResNet model for classification of road and building is low, which is 0.76 and 0.77, respectively, but that of vegetation is the best, with the precision of 0.97. And the recall and F1-score of different categories are between 0.79 and 0.90. It can be found that the average values of precision, recall and F1-score using the VGG-16 and ResNet model of the seven categories are above 0.85, which is slightly higher than those (i.e. 0.83) using the AlexNet model.

| Table 2 Overall accuracy of three CNN models |
| CNN      | OA     |
|-----------|--------|
| AlexNet   | 89.5%  |
| VGG-16    | 92.1%  |
| ResNet    | 93.4%  |

| Table 3 Accuracy, recall and F1-score of AlexNet model |
| AlexNet | P | R | F1 |
|---------|---|---|----|
| Road    | 0.84 | 0.77 | 0.81 |
| Building| 0.92 | 0.83 | 0.87 |
| Factory | 0.84 | 0.93 | 0.88 |
| Bare ground | 0.73 | 0.84 | 0.78 |
| Agricultural | 0.86 | 0.73 | 0.79 |
### Table 4 Accuracy, recall and F1-score of VGG-16 model

| Category   | P   | R   | F1  |
|------------|-----|-----|-----|
| Road       | 0.86| 0.80| 0.83|
| Building   | 0.84| 0.90| 0.87|
| Factory    | 0.91| 0.86| 0.88|
| Bare ground| 0.82| 0.89| 0.85|
| Agricultural| 0.95| 0.76| 0.84|
| Water      | 0.78| 0.80| 0.79|
| Vegetation | 0.85| 0.91| 0.91|
| Average value | 0.86| 0.85| 0.85|

### Table 5 Accuracy, recall and F1-score of ResNet model

| Category   | P   | R   | F1  |
|------------|-----|-----|-----|
| Road       | 0.76| 0.81| 0.79|
| Building   | 0.77| 0.81| 0.79|
| Factory    | 0.86| 0.90| 0.88|
| Bare ground| 0.90| 0.90| 0.90|
| Agricultural| 0.83| 0.86| 0.85|
| Water      | 0.89| 0.83| 0.86|
| Vegetation | 0.97| 0.84| 0.90|
| Average value | 0.86| 0.85| 0.85|

From the above results, it can be found that the average values of precision, recall and F1-score of the seven categories of the three CNN models are between 0.83 and 0.86. Among them, the ResNet model has certain advantages over the other two CNN models in terms of overall accuracy, precision, recall and F1-score. Its overall accuracy is 93.4%, and the average values of precision, recall and F1-score are 0.86, 0.85 and 0.85, respectively.

### 5. Conclusions

In this paper, three types of CNN (i.e. AlexNet, VGG-16, ResNet) models are constructed through the production of remote sensing image scene dataset along the high-speed railway to train and test the seven types of ground objects: road, building, factory, bare ground, agricultural, water and vegetation. A comparative evaluation is made to provide effective technical means for effectively distinguishing buildings and factories along the high-speed railway, so as to quickly detect hidden dangers along the high-speed railway. The experiment got the following conclusions:

1) The three CNN models of AlexNet, VGG-16 and ResNet are used to classify scene images along the high-speed railway with the overall accuracy of over 89%, and the highest is up to 93.4%. This shows that the CNN model has a good performance on the scene classification of remote sensing images along the high-speed railway, which can be considered to have certain practical value.

2) The ResNet model takes into account the length of the network and adds a residual module to compensate for the problem of gradient disappear of the increasing length. Therefore, its overall accuracy, precision, recall and F1-score are better than AlexNet and VGG-16 models. It can be considered that using the ResNet model to classify scene images along the high-speed railway has the best performance.
ACKNOWLEDGMENTS
This work was jointly supported by the National Natural Science Foundation of China (No. 41904031), the Natural Science Foundation of Jiangxi Province (No. 20202BABL213033), the Research Project of Teaching Reform in Colleges and Universities in Jiangxi Province (No. JXJG-19-6-13) and the Foundation of Research Center for Ecological Civilization Construction System of Jiangxi Province of East China University of Technology (No. 19GL02).

REFERENCES
[1] Li Z, Shen H, Cheng Q, et al. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. ISPRS Journal of Photogrammetry and Remote Sensing, 150: 197-212.
[2] Gong X, Shen L, Lu T. Refining training samples using median absolute deviation for supervised classification of remote sensing images. Journal of the Indian Society of Remote Sensing, 2019, 47(4): 647-659.
[3] Zhong Y, Cui M, Zhu Q, et al. Scene classification based on multi feature probabilistic latent semantic analysis for high spatial resolution remote sensing images. Journal of Applied Remote Sensing, 2015, 9(1): 95-64.
[4] Lu X, Zheng X, Yuan Y. Remote sensing scene classification by unsupervised representation learning. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(9): 5148-5157.
[5] Wang C, Lin W, Tang P. Multiple resolution block feature for remote-sensing scene classification. International Journal of Remote Sensing, 2019, 40(17-18): 1-21.
[6] Rodriguez-Galiano V F, Ghimire B, Rogan J, et al. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 2012, 67(1): 93-104.
[7] Pradhan B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Computers and Geosciences, 2013, 51(2): 350-365.
[8] Xu S, Mu X, Chai D, et al. Remote sensing image scene classification based on generative adversarial networks. Remote Sensing Letters, 2018, 9(7): 617-626.
[9] Zhong Y, Zhu Q, Zhang L. Scene classification based on the multi-feature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(11): 6207-6222.
[10] Zhao B, Zhong Y, Xia G S, et al. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(4): 2108-2123.
[11] Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285-1298.
[12] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.
[13] Maggiori E, Tarabalka Y, Charpiat G, et al. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 2016, 55(2): 645-657.
[14] Cheng G, Yang C, Yao X, et al. When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(5): 2811-2821.
[15] Tian Z, Zhan R, Hu J, et al. SAR ATR based on convolutional neural network. Journal of Radars, 2016, 5(3): 320–325.
[16] Jiang B, He J, Yang S, et al. Fusion of machine vision technology and AlexNet-CNNs deep learning network for the detection of postharvest apple pesticide residues. Artificial Intelligence in Agriculture, 2019, 1: 1-8.
[17] Kaur T, Gandhi T K. Automated brain image classification based on VGG-16 and transfer learning. 2019 International Conference on Information Technology (ICIT). IEEE, 2020.

[18] Lu Z, Jiang X, Kot A. Deep coupled ResNet for low-resolution face recognition. IEEE Signal Processing Letters, 2018, 25(4): 526-530.

[19] Shen L, Fang C, Wang J, et al. Automatic recognition of buildings along high-speed railway based on framework of scene interpretation. Remote Sensing Information, 2018, 33(3): 77-84.

[20] Giancardo L, Arevalo O, Tenreiro A, et al. MRI compatibility: Automatic brain shunt valve recognition using feature engineering and deep convolutional neural networks. Scientific reports, 2018, 8: 16052.