RESEARCH ON SE-INCEPTION IN HIGH-RESOLUTION REMOTE SENSING IMAGE CLASSIFICATION

Zhiling Cai [1], Qian Weng [2, *], Shaozhen Ye [3, *]

1 College of Mathematics and Computer Science，Fuzhou University, Fuzhou, China - cai_zhi_ling@163.com
2 National Engineering Research Centre of Geospatial Information Technology, Fuzhou University, Fuzhou, China - fzuwq@fzu.edu.cn
3 Research Institute of Intelligent Manufacturing Simulation, Fuzhou University, Fuzhou, China - yeshzh@fzu.edu.cn

KEY WORDS: Deep Learning, Transfer Learning, Convolutional Neural Networks Inception-V3, SENet, High Spatial Resolution Remote Sensing Images, Remote Sensing Image Classification

ABSTRACT: With the deepening research and cross-fusion in the modern remote sensing image area, the classification of high spatial resolution remote sensing images has captured the attention of the researchers in the field of remote sensing. However, due to the serious phenomenon of " same object, different spectrum" and "same spectrum, different object" of high-resolution remote sensing image, the traditional classification strategy is hard to handle this challenge. In this paper, a remote sensing image scene classification model based on SENet and Inception-V3 is proposed by utilizing the deep learning method and transfer learning strategy. The model first adds a dropout layer before the full connection layer of the original Inception-V3 model to avoid over-fitting. Then we embed the SENet module into the Inception-V3 model for optimizing the network performance. In this paper, global average pooling is used as squeeze operation, and then two fully connected layers are used to construct a bottleneck structure. The model proposed in this paper is more non-linear, can better fit the complex correlation between channels, and greatly reduces the amount of parameters and computation. In the training process, this paper adopts the transfer learning strategy, makes full use of existing models and knowledge, improves training efficiency, and finally obtains scene classification results. The experimental results based on AID high-score remote sensing scene images show that SE-Inception has faster convergence speed and more stable training effect than the original Inception-V3 training. Compared with other traditional methods and deep learning networks, the improved model proposed in this paper has greater accuracy improvement.

* Corresponding author
E-mail: fzuwq@fzu.edu.cn
* Corresponding author
E-mail: yeshzh@fzu.edu.cn

1. INTRODUCTION

With the development of remote sensing technology, remote sensing image data have been widely used in land resources management, urban management and national defense. Remote sensing image classification is one of the key basic technologies in modern remote sensing applications. This technology mainly uses computer to classify and recognize land feature information from images in an automatic or semi-automatic way. Classification maps can not only provide auxiliary discriminant information for target detection and recognition, but also can be used to produce and update land use type maps. One of the key basic technologies in the application of modern remote sensing technology [1,2].

Compared with low and medium spatial resolution remote sensing images, high spatial resolution remote sensing images (hereinafter referred to as "high resolution remote sensing images") can reflect more abundant details and semantic information of terrain objects, but at the same time it brings more complex problems of "homologous foreign objects" and "homologous foreign spectra", which are embodied in different semantic categories of scenes. There are the same object composition, but the scene of the same kind has different resolution and spatial distribution of objects. This kind of difference between and within classes is small, which makes the traditional pixel-based and object-based remote sensing image classification methods prone to excessive misclassification and omission, resulting in low classification accuracy. In recent years, scene-based image classification technology has become a research hotspot of high-resolution image classification [3]. According to image feature hierarchy, scene-based high-resolution remote sensing image classification methods can be divided into three categories:

1.Low-level visual feature-based methods: Low-level visual feature-based methods use various feature operators to extract features from low-level visual attributes (such as color, texture, spectral value) of high-resolution remote sensing images to describe images [4], such as color histogram (CH) [5], spatial envelope feature (GIST) [6] Local Binary Patterns (LBP) [7] and Scale Invariant Feature Transform (SIFT) [8]. This method has a good classification effect for high-resolution remote sensing images with uniform spatial distribution and structure pattern, but it has a poor effect for scenes with uneven spatial distribution.

2.Method based on middle-level visual representation: The method based on middle-level visual representation is to encode low-level local visual features of high-resolution remote sensing images to form global feature representation of scene images. The commonly used coding models are: Bag of words (BoVW) [9], Spatial Pyramid Matching. SPM [10], Locality-constrained Linear Coding (LLC) [11], Probabilistic Latent Semantic Analysis (pLSA) [12], Improved Fisher Kernel (IFK) [13], Vector of Local Aggregated Descriptors VLAD [14] etc. Compared with the method based on low-level visual features, the classification accuracy of the method based on middle-level visual representation has been greatly improved, but it is still limited by low-level visual features and coding methods, and can not achieve the optimal classification accuracy[15,16].

3.Method based on advanced semantic features: The development of deep learning provides a new idea for high-score image scene classification. In-depth learning can learn more abstract and discriminative features. Convolutional Neural Networks (CNN) has a large number of training label data and continuous multi-layer convolution process. It has been widely used in computer vision tasks [17,18]. AlexNet [19], CaffeNet [20], VGNet [21], Google LeNet [22], ResNet [23] and other deep network structures have emerged successively, which have achieved remarkable results in image classification tasks.

At present, most convolutional neural networks improve network accuracy by deepening network layers, but the following problems are over-fitting and network degradation [24] [25] [26]. Therefore, an improved Inception-V3 model is proposed and applied to scene classification of high-resolution remote sensing images. Firstly, aiming at the problem of small number of samples in high-resolution remote sensing scene set, this paper improves Inception-V3 network by adding Dropout layer before the last full-connection layer of the network to prevent over-fitting of the model. Then we embed the SENet module into the Inception-V3 model for optimizing the network performance. Then, using migration learning, the Inception-V3 model pre-trained on large natural image set (ImageNet) is migrated to the highest level. In order to reduce the difficulty of training and improve the classification effect of the model, remote sensing scene set is divided and fine-tuning training is carried out.

2. RELATED WORK

2.1 Inception-V3

Inception model was proposed by Christian Szegedy and others in 2015. Inception model decomposes the original large convolution kernel into small convolution kernels with the same operation by improving the structure of neural network. The spatial decomposition of asymmetric convolution is carried out, combined with the use of auxiliary filters. At the same time, the model further reduces the size of the feature graph. Effective preservation of image features while reducing computational complexity. In addition, Inception introduces the BN (Batch Normalization) method, which achieves efficient operation of the network by decomposing convolution and regularization, and speeds up the convergence speed of training. Inception-V3 model divides two-dimensional convolution layers into two one-dimensional convolution layers, which not only reduces the number of parameters, but also reduces the over-fitting phenomenon. Its network structure is shown in Figure 1.

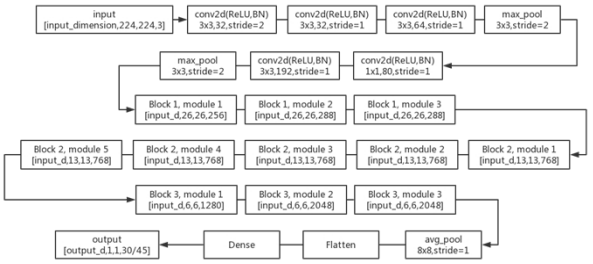


Figure 1. Inception-V3 network architecture

2.2 Transfer Learning

Traditional in-depth learning methods assume that training data and test data have the same input characteristics and the same distribution. However, in practical applications, it is very difficult to obtain training data with the same feature space and distribution as test data. Transfer learning solves the training problem of a small number of labeled samples through model and parameter transfer, and improves the efficiency of model learning.

Transfer learning can improve the efficiency of learning and training in one field (target domain) by transmitting information from related fields (source domain), and it can effectively solve the problem of information island. Transfer learning takes advantage of the powerful functions of deep neural network and image network data set, which can transfer the knowledge learned from the natural image processing model suitable for large data volume to the high-resolution remote sensing image field suitable for small data volume, and realize personalized

migration. The source domain is the labeled sample, and the target domain is the unlabeled sample. The probability distributions P and Q represent the edge distributions of the source domain and the target domain, respectively. A small number of labeled samples in the target domain are defined as annotated samples in the source domain and the target domain. By supervised learning in the source domain, deep transfer learning builds a transferable neural network that can learn cross-domain differences, and establishes a classifier to minimize the target risk. For convolutional neural networks, there are two strategies for using transfer learning: freezing and fine-tuning. These two strategies are shown in Figure 2. Firstly, the training parameters obtained from the pre-training model are used to initialize the target network, and then the target data are used to train. Freezing means freeze some layers, that is, keep the weight of some layers unchanged, and train the rest.

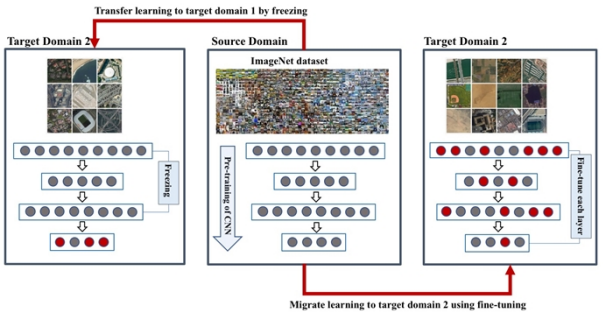


Figure 2. Freezing and Fine-Tuning in Transfer Learning

3. MODEL

3.1 Improved Inception-V3

Deep learning network is a multi-layer neural network with a large number of parameters, such as Inception-V3, which has a total of more than 20 million parameters. Overfitting is a serious problem for such networks. Dropout and Batch Normalization (BN) are two widely used methods to reduce over-fitting. As shown in Figure 3, dropout refers to the temporary removal of one or more neurons from the neural network by randomly discarding one or more neurons, making the network more sparse and compact, and easier to predict the output. BN is a method to improve the training speed of deep neural network by reducing the covariance offset within the neural network [27].

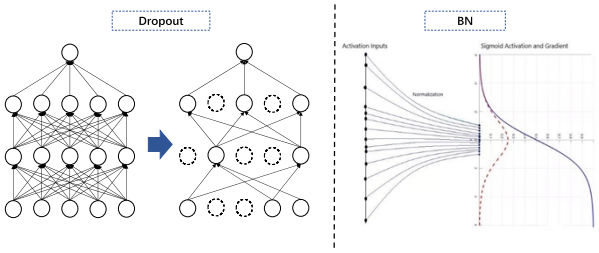


Figure 3. Dropout and BN schematics

The original Inception-V3 reduces the occurrence of over-fitting by using BN and improves the training speed. In order to solve the problem of less labeling samples and high labeling cost of remote sensing data, Dropout operation is added before the output layer Softmax classifier to improve the original Inception-V3 network by combining with the convolution layer BN operation. Adding Dropout structure before the last full connection layer can avoid the variance offset caused by the simple combination of BN and Dropout, and further reduce the over-fitting phenomenon of convolutional neural network. Its structure is shown in Figure 4.

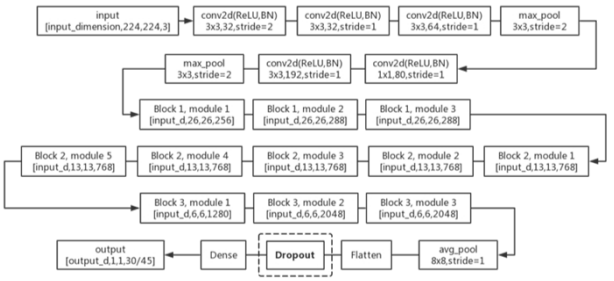


Figure 4. Improved Inception-V3 Network Architecture

3.2 SENet

In 2017, based on the relationship between feature channels, Ho Jie and others proposed Squeeze-and-Excitation Networks (SENet). Squeeze and Excation are two key operations, so SENet is named after them. SENet explicitly models the interdependence between feature channels. In addition, instead of introducing a new spatial dimension to fuse feature channels, SENet adopts a new strategy of "feature re-calibration". Specifically, it is to automatically acquire the importance of each feature channel by learning, and then according to this importance to enhance useful features and suppress features that are not very useful for the current task.

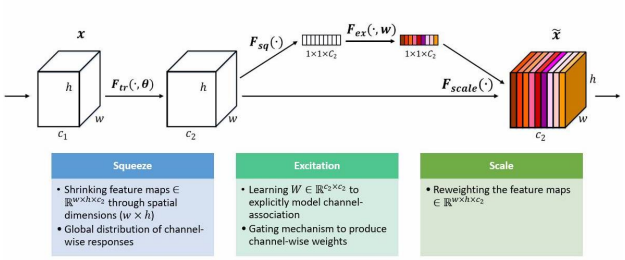


Figure 5. SENet Network Structure Diagram

The figure above is a schematic diagram of SENet. Given an input x, the characteristic channel number is c1. After a series of convolutions and other general transformations, a characteristic channel number is c2. Unlike traditional CNN, SENet uses three operations to re-calibrate the previous features.

First is Squeeze operation. SENet compresses the feature along the spatial dimension, transforming each two-dimensional feature channel into a real number, which has a global sense field to some extent, and the output dimension matches the number of input feature channels. It represents the global distribution of response on feature channels, and enables the layer close to the input to obtain the global sensing field, which is very useful in many tasks.

Secondly, Excation operation, which is similar to the gate mechanism in the circular neural network. The parameter W is used to generate weights for each feature channel, where the parameter W is learned to explicitly model the correlation between feature channels.

Finally, it is a Reweight operation. SENet regards the weight of Excitation output as the importance of each feature channel after feature selection, and then completes the re-calibration of the original feature on the channel dimension by multiplying the channel-by-channel weighted to the previous feature.

3.3 SE-Inception

SENet is simple to construct and easy to deploy without introducing new functions or layers. In addition, it also has good characteristics in model and computational complexity. Therefore, this paper introduces SE module into Inception-V3 model (referred to as "SE-Inception"), as shown in Figure 6. The dimension information next to the box represents the output of the layer.

This paper uses Global average pooling as Squeeze operation. Next, two Fully Connected layers form a Bottleneck structure to model the correlation between channels, and output the same number of weights as the input features. Firstly, the feature dimension is reduced to 1/16 of the input, then activated by ReLu, and then ascended back to the original dimension through a Fully Connected layer. The advantages of this method over using a Fully Connected layer directly are: (1) more nonlinearity, which can better fit the complex correlation between channels; and (2) greatly reducing the amount of parameters and computation. Then the normalized weights between 0 and 1 are obtained through a Sigmoid gate. Finally, the normalized weights are weighted to the characteristics of each channel through a Scale operation[28].

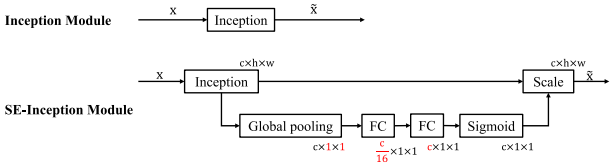


Figure 6. SE-Inception Network Architecture

SE-Inception has a 10% increase in model parameters relative to Inception-V3. Additional model parameters exist in the two Fully Connected designs of Bottleneck. Because the number of characteristic channels of the last stage in Inception structure is 2048, the model parameters increase greatly. If the SE settings on the three build blocks in the last stage are removed, 10% of the parameters can be obtained. Growth fell to 2%. At this time, the accuracy of the model is almost lost. In addition, in the existing GPU implementations, neither global pooling nor Fully Connected with a small amount of computation has been optimized, which results in a 10% increase in runtime SE-Inception over Inception-V3 on GPU. Nevertheless, the additional computational effort for theoretical growth is less than 1%, which matches the increase in CPU runtime. It can be seen that the embedded SE module in the existing network architecture leads to little increase in additional parameters and computational complexity.

With the improvement of image resolution, the spectral, texture and spatial characteristics of high-resolution remote sensing images and natural optical images are getting closer and closer. Because the first layers of deep learning network are used to acquire local features of images, this paper will freeze and migrate the first three layers of network parameters based on the pre-trained Inception-V3 on the large natural image set of ImageNet, and fine-tune the weight parameters of the latter layer. Since the last full connection layer of SE-Inception is after the newly added Dropout, it needs to be retrained. The migration process is shown in Figure 7.

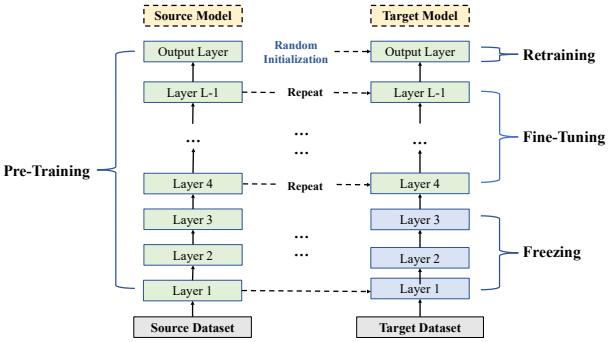


Figure 7. The Sketch of The Experiment

In the first stage, the structure of pre-trained Inception-V3 on ImageNet data set is improved, and then it is migrated to the high-resolution scene data set to train fine-tuning. The fine-tuned Inception-V3 is used to extract the feature vectors of high-resolution scene images, and a 2048-dimensional feature vector is obtained. In this stage, knowledge transfer is fully utilized and pre-training weights on ImageNet are used as initial values. Compared with random initial values, training time is greatly shortened. Then, the extracted features are input into the full-connected neural network. Since SE-Inception-V3 network can learn more abstract and easy-to-classify features of scene images, a single-layer full-connected neural network including Softmax is used, and the final classification results can be obtained by training a small amount of labeled image data.

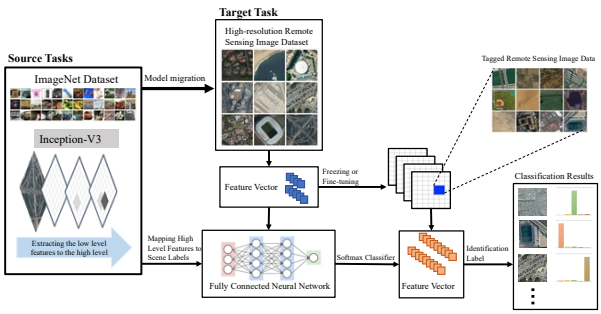


Figure 8. Flow chart of Scene Classification for High-Resolution Remote Sensing Images

4. EXPERIMENTS

4.1 Datasets

Aerial Images Datasets (AID) is a large high-resolution remote sensing scene data set released by Wuhan University in 2016. The data set contains 10,000 images in 30 categories. Images are mainly collected from Google Earth. The resolution of the scene image is 0.5-8 meters, and the image size is 600,600 pixels. The categories of the data set are shown in Figure 9. There are 220-420 images in each category.

Figure 7. Examples of 28 categories of Scene Images in AID

4.2 Experimental Setup

In this paper, we use Tensorflow 1.2 to implement Inception-V3 nd SE-Inception model, and test and run on Baidu Cloud Deep Learning Platform. The platform environment is 40GB of memory and Nvidia K40.

4.3 Evaluate Criterion

Five indicators were used to evaluate the performance: *train_accuracy*, *test_accuracy*, *train_loss*, *test_loss* and *confusion matrix*.

train_acc: The classification accuracy of the training set. The calculation formula is as follows:

$$train\_accuracy = \frac{\sum X_{train}^{predict=true}}{\sum X_{train}^{true}} \quad (1)$$

test_acc: The classification accuracy of the test set. The calculation formula is as follows:

$$test\_accuracy = \frac{\sum X_{test}^{predic=true}}{\sum X_{test}^{true}} \quad (2)$$

train_loss: Cross-entropy loss of training set. The calculation formula is as follows:

$$loss = -\sum\sum t_{train}^{i,j} log\left(p_{train}^{i,j}\right) \quad (3)$$

test_loss: The cross-entropy loss of the test set. The calculation formula is as follows:

$$val\_loss = -\sum\sum t_{test}^{i,j} \log\left(p_{test}^{i,j}\right) \quad (4)$$

In the experiment, the Inception parameter is updated by Adam algorithm[29], and the loss value is calculated by formula (5). The initial learning rate is set to 0.2. When *val_loss* is no longer reduced, the learning rate is updated by learning rate decay algorithm.

$$lr = lr \times factor \quad (5)$$

lr is the learning rate, factor is the attenuation factor of learning rate, the lower limit is set to 0. The batch size of the training set is set to 64, and the epochs number is set to 30. AID dataset construct training and test sets at 20% and 80% proportions for each category scenario.

4.4 Experimental Result

The original Inception's train_loss and test_loss indices in AID training set and test set are expressed by yellow dotted line and red dotted line respectively. The training results are shown in Figure 8 (a). It can be seen that at the beginning of training, the train_loss and test_loss values are higher because Inception is under-fitting, but with the increase of epochs, the train_loss values are higher. The values of train_loss and test_loss fluctuate occasionally during training, but after the 14th epochs, the values of train_loss and test_loss tend to be stable and the parameters converge basically. From the blue and green solid lines representing Inception's train_accuracy and test_accuracy indices in Figure 9(b), we can see that at the beginning of training, test_accuracy will fluctuate in the initial iteration because the gradient descent may fall on the local minimum of the non-optimal solution, but after the 14th epochs, train_accuracy and test_accura will fluctuate. Cy tends to be stable and achieves the highest classification accuracy.

方法	AID
LBP	26.26 ± 0.52
CH	34.29 ± 0.40
SIFT	13.24 ± 0.74
GIST	30.61 ± 0.63
BoVW+CH	47.77 ± 0.52
IFK+CH	64.83 ± 0.42
LLC+CH	49.36 ± 0.57
pLSA+CH	42.87 ± 0.54
SPM+CH	41.27 ± 0.49
VLAD+CH	44.78 ± 0.28
AlexNet	86.34 ± 0.43
VGG-16	86.87 ± 0.41
GooLeNet	83.84 ± 0.36
ResNet50	89.70 ± 1.05
Inception-V3	94.18 ± 0.40
SE-Inception	94.62 ± 0.23

The blue and green solid lines in Figure 8 represent the performance indicators of Inception on the AID training set and the test set, respectively. Compared with Figure 8 (a) and (b), the fluctuations of loss and accuracy in SE-Inception are significantly improved compared with the original Inception-V3 in both training and test sets. They tend to be stable after the 11th epochs, and converge earlier than the original Inception-V3. At the same time, the test_accuracy of SE-Inception is slightly higher than that of the original Inception-V3 according to the comparison of the blue solid line and the red dotted line in Figure 8 (b). It can be seen that SE-Inception can extract the high-level features of high-score scene images well, although the training parameters are reduced, thus improving the classification accuracy of high-score scene images. At the same time, due to the application of Dropout, SE-Inception can effectively reduce the over-fitting, improve the training speed and reduce the training difficulty on the premise of small training samples.

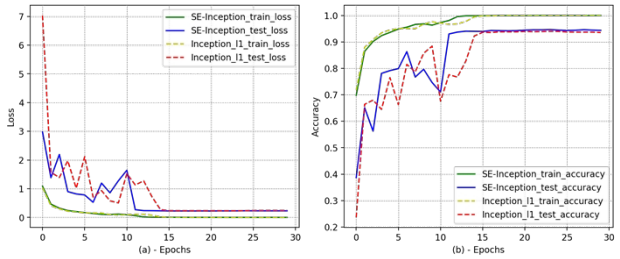


Figure 8. Inception and SE-Inception Training on AID Datasets

Table 1 comprehensively compares the average classification accuracy and standard deviation of feature classification methods at three levels: low-level visual features, middle-level visual representation and high-level semantic features after 30 random experiments. LBP, CH, GIST and SIFT are selected for low-level visual features. For LBP, CH and SIFT features, 16 *16 window is used to extract local features with 8 pixels sliding step size, and then the final feature vector is obtained by average pooling. Compared with middle-level and high-level visual feature methods, the classification accuracy of low-level visual feature methods is the worst. The classification accuracy of CH is better than that of the other three low-level features in AID. This is because most of the remote sensing scene images in the two datasets have color consistency. Using CH can extract all kinds of spectral information. GIST achieves better classification accuracy on AID data sets, because AID data sets contain a large number of artificial scene categories composed of various buildings. GIST can fully extract the spatial structure of these categories of images.

The middle-level visual representation uses six coding methods: visual word bag (BoVW), improved Fisher kernel (IFK), locally constrained linear coding (LLC), probabilistic latent semantics analysis (pLSA), spatial pyramid matching (SPM) and locally aggregated descriptor vector (VLAD), and combines the best CH features in the low-level features. Code. The dictionary sizes of BoVW, IFK, LLC, pLSA, SPM and VLAD are set to 4096, 128, 4096, 1024, 256 and 128 respectively, the number of topics of pLSA is set to 64, the pyramid level of SPM is set to 2, and the classifier uses SVM with linear kernel. The classification accuracy of IFK coding method is better than other coding methods. This is because BoVW, LLC and VLAD are coded based on feature dictionary, and IFK describes the scene image by calculating the probability density distribution of local features through the Gauss mixture model. This coding method can better describe the spatial distribution characteristics of remote sensing scene image.

Table 1 Comparison of classification accuracy

The method of obtaining high-level semantic features using deep learning network has the highest classification accuracy, which is about 30% higher than that of middle-level visual representation. Compared with four deep networks AlexNet, VGG16, GooLeNet and ResNet50, SE-Inception has better classification accuracy than AlexNet, VGG-16, GooLeNet, ResNet50 and original Inception-V3. This is because SE-Inception can establish the interdependence between feature channels. By learning, the importance of each feature channel is automatically acquired, and then according to this importance, useful features are enhanced and features that are not useful for current tasks are suppressed. In addition, the application of migration learning can make full use of the knowledge of large natural datasets, extract the image information of different proportion feature distribution through a large number of convolution and pooling, and use Inception Block to extract the complex feature information of remote sensing scenes by using multi-scale convolution kernels; at the same time, SE-Inception obtains the complex feature information of remote sensing scenes by the end of the paper. The random inactivation mechanism is added before the first layer, which further reduces the occurrence of over-fitting and improves the training efficiency.

The confusion matrix can directly reflect the classification effect of each category in the data set. Figure 9 shows the confusion matrix of SE-Inception-V3 on AID data sets. It can be seen that the classification accuracy of SE-Inception on AID is high, and the classification accuracy of 2-Baseball field, 3-Beach, 11-Forest, 13-Meadow, 15-Mountain, 25-Sparse Residential is 100%.

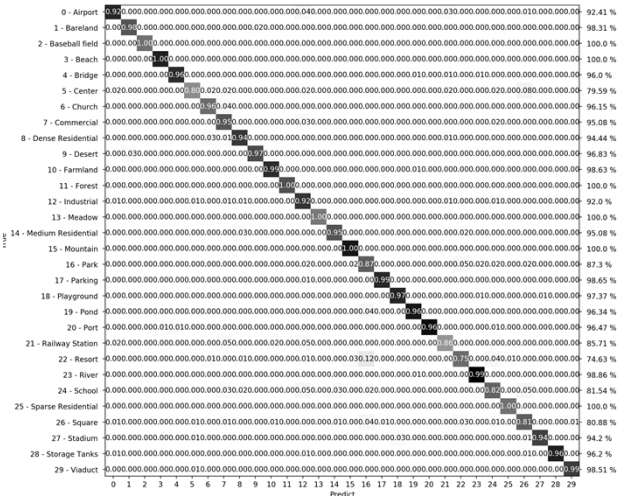
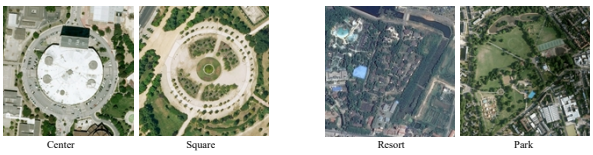


Figure 9. AID Dataset Confusion Graph of SE-Inception

Figure 9 shows that the classification accuracy of 5-Center and 22-Resort in AID data set is the lowest. As shown in Figure 10 (a) below, the remote sensing images of the roofs of Center and the central area have high similarity and similar characteristics, so it is impossible to distinguish Center from the Square better. Resorts are divided into Park by mistake. As shown in Figure10 (b) below, both resorts and parks have higher vegetation coverage, fewer buildings and similar planning methods. Therefore, parks and resorts can not be well distinguished.



(a) Center and Square (b) Resort and Park

Figure.10 Legend of Center and Square, Resort and Park

5. CONCLUSIONS

Due to the lack of training samples for high-score remote sensing data sets, traditional deep learning network training is prone to over-fitting. To solve this problem, this paper proposes a network model of SE-Inception-V3, and adds Dropout layer before the last full connection layer of the original Inception-V3 to further reduce the occurrence of over-fitting. In network training, the migration learning technology is further applied. First, the pre-training parameters on ImageNet are transferred to the AID data set to fine-tune. Experiments on the AID high-resolution remote sensing scene set show that SE-Inception-V3 improves the classification accuracy, and at the same time, the training convergence speed is faster, low-level visual features, middle-level visual representation and. The comparison of classification accuracy of high-level semantic features shows that the classification performance of SE-Inception-V3 is better than other hierarchical feature classification methods. Compared with other commonly used deep learning networks, the classification accuracy of SE-Inception-V3 is also improved by 5%-10%. Combining the trained depth network model with a more efficient classifier to further improve the classification accuracy of high-resolution remote sensing scene images is a follow-up research issue.

ACKNOWLEDGEMENTS

The study is supported by the National Natural Science Foundation of China (No. 41801324) and the Fujian Natural

Science Foundation under Grant 2019J01244.The authors would like to thank Professor Gui-Song Xia for providing the AID dataset . The dataset is of high quality and the research community is fortunate to have such a fantastic resource.

REFERENCES

[1] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pp. 439–451, 2014.

[2] Blaschke T. Object based image analysis for remote sensing[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2010, 65(1):2-16.

[3] CHEN C,ZHANG B,SU H,et al. Land-use scene classification using multi-scale completed local binary patterns [J]. Signal,Image&Video Processing,2016,10(4):745-752.

[4] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang, "Aid: A benchmark dataset for performance evaluation of aerial scene classification," arXiv preprint arXiv:1608.05167, 2016.

[5] M. J. Swain and D. H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, no. 1, pp. 11–32, 1991.

[6] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.

[7] T. Ojala, Pietik, M. Inen, and Topi, "Multiresolution gray-scale and ro- tation invariant texture classification with local binary patterns," Pattern Analysis & Machine Intelligence IEEE Transactions on, vol. 24, no. 7, pp. 971–987, 2002.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[9] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in ACM Sigspatial International Symposium on Advances in Geographic Information Systems, Acm-Gis 2010, November 3-5, 2010, San Jose, Ca, Usa, Proceedings, 2010, pp. 270–279.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in Computer vision and pattern recognition, 2006 IEEE computer society conference on, vol. 2. IEEE, 2006, pp. 2169–2178.

[11] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality- constrained linear coding for image classification," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3360–3367.

[12] A. Bosch, A. Zisserman, and X. Mun˜oz, "Scene classification via plsa," in European conference on computer vision. Springer, 2006, pp. 517– 530.

[13] F. Perronnin, J. Sa´nchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in European conference on com- puter vision. Springer, 2010, pp. 143–156.

[14] H. Jegou, F. Perronnin, M. Douze, J. Sa´nchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," IEEE trans- actions on pattern analysis and machine intelligence, vol. 34, no. 9, pp. 1704–1716, 2012.

[15] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification." IEEE Geosci. Remote Sensing Lett., vol. 12, no. 11, pp. 2321–2325, 2015.

[16] Alshehhi R, Marpu, Prashanth Reddy, Woon, Wei Lee, et al. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2017, 130:139-149.

[17] Weng Q, Mao Z, Lin J, et al. Land-use classification via extreme learning classifier based on deep convolutional features[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(5): 704-708.

[18] Fan H, Xia Gui-Song, Hu Jingwen, et al. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery[J]. Remote Sensing, 2015, 7(11):14680-14707.

[19] Zhou W, Newsam S, Li C, et al. Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval[J]. Remote Sensing, 2017, 9(5):489.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural infor- mation processing systems, 2012, pp. 1097–1105.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Computer Science, 2014.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[24] Zhou W, Shao Z, Diao C, et al. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder[J]. Remote Sensing Letters, 2015, 6(10):775-783.

[25] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag- of-visual-words model for remote sensing image scene classification," Journal of Applied Remote Sensing, vol. 10, no. 3, p. 035004, 2016.

[26] Ma L, Li, Manchun, Ma, Xiaoxue, et al. A review of supervised object-based land-cover image classification[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2017, 130:277-293.

[27] Li X , Chen S , Hu X , et al. Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift[J]. 2018.

[28] [28] Cheng D , Meng G , Cheng G , et al. SeNet: Structured Edge Network for Sea–Land Segmentation[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(2):247-251.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Computer Science, 2014.