A NOVEL FRAMEWORK FOR REMOTE SENSING IMAGE SCENE CLASSIFICATION

Shulong Jiang[1,2], Hongrui Zhao[1,2*], Wenjia Wu[1,2], Qifan Tan[1,2]

[1]Department of Civil Engineering, Tsinghua University, Beijing 10084, China –
(jsl15@mails.tsinghua.edu.cn, zhr@tsinghua.edu.cn, wu-wj14, tqf17@mails.tsinghua.edu.cn)
[2]3S Center, Tsinghua University, Beijing 10084, China

Commission III, WG III/1

KEY WORDS: Scene Classification, Deep Learning, Convolutional Neural Network, Fully-connected Layer, XGBoost,

ABSTRACT:

High resolution remote sensing (HRRS) images scene classification aims to label an image with a specific semantic category. HRRS images contain more details of the ground objects and their spatial distribution patterns than low spatial resolution images. Scene classification can bridge the gap between low-level features and high-level semantics. It can be applied in urban planning, target detection and other fields. This paper proposes a novel framework for HRRS images scene classification. This framework combines the convolutional neural network (CNN) and XGBoost, which utilizes CNN as feature extractor and XGBoost as a classifier. Then, this framework is evaluated on two different HRRS images datasets: UC-Merced dataset and NWPU-RESISC45 dataset. Our framework achieved satisfying accuracies on two datasets, which is 95.57% and 83.35% respectively. From the experiments result, our framework has been proven to be effective for remote sensing images classification. Furthermore, we believe this framework will be more practical for further HRRS scene classification, since it costs less time on training stage.

1. INTRODUCTION

Remote sensing is a non-contact technology for surface observations that enables rapid and large-scale acquisition of surface information. The automatic extraction of ground object information from remote sensing images is a hot topic in the field of remote sensing image analysis (Zhang et al. 2016). With the continuous improvement of the spatial resolution of remote sensing images, the information extraction methods for remote sensing images are not satisfied with pixel-based and object-based methods (Fu G, Liu C, Zhou R, et al. 2017). People want to mine a higher level of semantic information from the image, and ground objects forms different semantic scene categories through different spatial distribute pattern (Bratasanu et al. 2011; Lienou et al. 2010). The scene not only contains the information of the ground objects, but also includes the spatial relationship between the ground objects and the environment. Scene categories of images are people's overall understanding of an image, and contain the contextual information of objects in the image. For example, Fig.1 (a), (b), (c) can be classified into the commercial area, residential and forest scene theme respectively. With the rapid growth of the amount of remote sensing image data, the semantic information of automated mining images is even more important. The need for automatic annotation methods for image scenes is also urgent.

In recent years, methods based on deep learning technology have made great breakthroughs in some computer vision task, for example, image classification and target recognition (Lecun Y et al. 2015). The convolutional neural network model has good feature extraction and classification ability in natural scene images. Deep learning technology has also gradually attracted the attention of remote sensing communities. And deep learning technology have shown good performance in HRRS scene classification (Zou Q et al. 2015). For example, models such as CaffeNet, VGGNet and GoogLeNet model have been successfully applied in remote sensing scene classification tasks. The structures of these models are mostly composed of multi convolutional layers and fully connection layers. Convolutional layers are used to extract image features, fully connection layer for classification. However, the number of parameters in the fully connected layer accounts for almost 80% of the total number of model parameters, which greatly increases the training and use cost of the model, and poses a high risk of over-fitting for the lack of enough labeled data.

(a) commercial (b) residential (c)forest

Figure 1 Three scenes with different sematic class

In order to solve this problem, this paper presents a framework for HRRS images scene classification, using XGBoost classifier instead of fully connected layer classification. XGBoost is an ensemble learning algorithm. It is a very popular algorithm in academia and industry and has also achieved good results in a lot of data mining competitions such as Kaggle (the largest and most diverse data community in the world). The XGBoost

* Corresponding author should be addressed to Hongrui Zhao, Email: zhr@tsinghua.edu.cn.

system provides open source toolkits, and users can expand on their own needs. (Chen et al. 2015). In general, the framework consists of two parts: (1) fine-tuning the pre-trained CNN model, transfer the CNN pre-trained on the large-scale dataset ImageNet(Fei-Fei L et al. 2009), and using the CNN for images feature extraction; (2) using XGBoost classifies the features to get the scene category of the images.

The remainder of this paper is organized as follows. In Section 2, some related works are presented. Our framework for HRRS scene classification are described in Section 3. In Section 4, we introduce the datasets and experiment in detail. Our framework’s performance are shown and discussed in Section 5. Finally, Section 6 presents the conclusion of this paper.

2. RELATED WORK

In the early studies of remote sensing image scene classification, people usually used hand-designed low-level image features, such as color histograms, texture descriptors, GIST, scale invariant feature transform (SIFT), and histogram of oriented gradients (HOG) etc. People use these low-level features to describe images and distinguish image categories for a long time. But these classification methods based on the low-level features of the remote sensing images is very sensitive to noise. These traditional scene recognition methods rely on these image features hand-crafted by experts and require a lot of prior knowledge, and the generalization ability of these features is not strong.

Later, there were classification methods using the middle-level features of the images, and these methods were mostly based on the bags of the visual words (BoVW) model (Yang et al. 2010) which was evolved from text classification. The features of the images are clustered using k-means algorithm to obtain a codebook of visual features. Then the frequency of visual words in the images is counted for classification. The BoVW model can fuse multiple features to improve the classification accuracy. And features fusing strategy is also a hot research issue. In addition, the topic model is also a very effective method. The topic model adds a hidden variable, the topic, between the scene category and the visual words of the image. LDA (Blei et al. 2003) and pLSA (Bosch et al. 2006) are the two most commonly used topic models. The BoVW model and topic model have achieved good results in the field of computer vision and has become the mainstream approach in image classification tasks for a period of time.

At present, deep learning technology which was inspired by human visual mechanism has been widely used in the field of computer vision, especially the convolutional neural network model, which has become the preferred model for various visual tasks. Deep learning has achieved excellent results in handwritten digital recognition tasks and has opened a door to image classification tasks. Deep learning method can extract global features of images hierarchically and achieve better images classification results. The outstanding performance of deep learning in the different datasets also confirmed its powerful generalization ability in feature extraction.

In the field of remote sensing scene classification, CNN models have also gradually been used. Reference (Otavio et al. 2015) used pre-trained convolutional neural networks for the first time to process remote sensing scene classification tasks, and confirmed that CNN has stronger generalization ability from natural images to remote sensing images when other approaches based on low-level features. Hu fan et al. transferred the pre-trained CNNs(AlexNet, CaffeNet, VGGNet) to overcome the limitation of lacking the labeled remote sensing data, and evaluated the CNN features get from the fully-connected layers and convolutional layers respectively for scene classification(Hu et al. 2015). Marco Castelluccio et al., fine-tuned the CaffeNet and GoogLeNet on remote sensing datasets (Castelluccio et al.2015). And the results showed that CNNs achieved the highest accuracy so far. Then there are many literatures designed many models for HRRS scene classification task. Yanfei Zhong et al. proposed the large patch convolutional neural network (LPCNN) that achieved good results in small-scale remote sensing dataset (Zhong et al. 2016). Dimitrios Marmanis et al. proposed an approach of fusing the many hidden layers’ features to reduce the computational burden of the model (Marmanis et al. 2016). Qian Weng et al. combined the convolutional neural networks and extreme learning machine (ELM), and they test this method on the UC-Merced dataset and achieved satisfactory results (Weng al. 2017). Since then, transferring the pre-trained CNN model from nature datasets has become the dominant method for remote sensing image feature extraction.

3. PROPOSED FRAMEWORK

Scene classification task has two important stages: (1) feature extraction and selection; (2) classifier design. Correspondingly, the architecture of this framework consists of two parts. (Figure 2). (1) We transfer pre-trained VGG-16 model on large image dataset ImageNet in order to better extract the features of remote sensing images. The global features of HRRS images are extracted through this stage. (2) XGBoost is an ensemble method, which improve the accuracy of classification through iterative computation of weak (basic) classifiers. Features got by first stage are fed into the XGBoost classifier, and this stage outputs the scene category of the image.

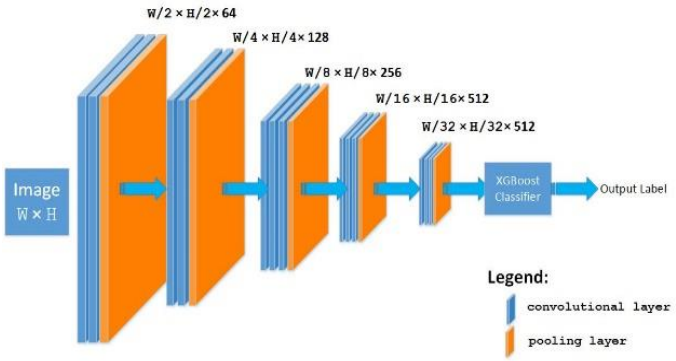

Figure 2. The architecture of our framework

3.1 Feature Extractor

The first stage utilizes the VGG-16 model (one of the most popular CNN models nowadays) to extract features from HRRS images. The architecture of feature extractor in VGG-16 has five parts which contains two or three convolutional layers and one pooling layer. The parameters of these parts are shown in table 1. The parameters of convolutional layers are expressed as “conv (kernel size weight * height) – (number of convolution kernels)”.

Part	Parameters
Part1	conv 3*3-64 conv 3*3-64 Max-pool 2*2
Part2	conv 3*3-128 conv 3*3-128 Max-pool 2*2
Part3	conv 3*3-256 conv 3*3-256 conv 3*3-256 Max-pool 2*2
Part4	conv 3*3-512 conv 3*3-512 conv 3*3-512 Max-pool 2*2
Part5	conv 3*3-512 conv 3*3-512 conv 3*3-512 Max-pool 2*2

Table 1 Parameters of 4 part in VGG-16 feature extractor

A convolutional layer get a feature map by computing the dot product between the receptive field and kernel. In general, an activation function is added behind each convolutional layer, such as the Sigmoid function, Rectified Linear Unit (ReLU), Tanh et al. This part uses the ReLU as the activation function. The formula for the ReLU function is :

$$f(x) = \max(0, x) \tag{1}$$

And ReLU function image shown in Figure 3.

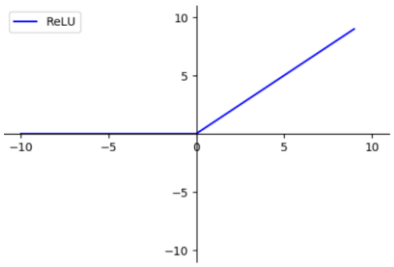

Figure 3 Image of the ReLU function.

The pooling layer is to downsample the image feature maps. There are two widely used pooling layers, the average pooling layer and the maximum pooling layer. The max-pooling layers used in this model, will return the max value from each sub-area, and the images are down-sampled by max-pooling layers, causing 1/2 reduction in each image's height and weight. Figure 4 is the image of the pooling layer.

Then we transfer pre-trained VGG-16 on large image dataset ImageNet in order to better extract the features of remote sensing images. The global features of HRRS images are extracted through this stage.

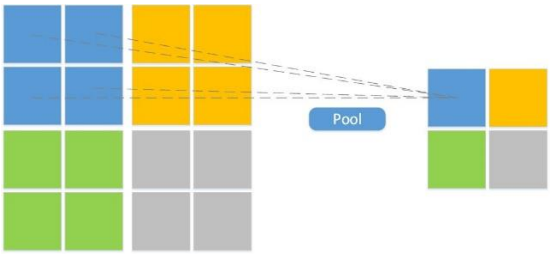

Figure 4. Illustration of pooling layer. The pooling layer with filters of size 2x2 and stride of 2 was shown above.

3.2 XGBoost Classifier

The XGBoost classifier replace the fully connected layer for classification. XGBoost is an ensemble method, which improve the accuracy of classification through iterative computation of weak (basic) classifiers. Features got by first stage are fed into the XGBoost classifier, and this stage outputs the scene category of the image.

The tree model is generally used as a basic classifier in XGBoost System. The K trees ensemble mode is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F} \tag{2}$$

$$\mathcal{F} = \left\{ f(x) = w_{q(x)} \right\} \left(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T \right) \tag{3}$$

Where $\mathcal{F}$ is the space of functions containing all regression trees. Here $f(x)$ is the decision function, $q(x)$ represents the structure of each tree, and w represents the score in the leaf. For a remote sensing image feature vector, we will get the score of its corresponding leaf node on each regression tree. Finally add these scores to get the final prediction of the image. We defined the objective function:

$$\text{obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

$$\Omega(f) = \gamma \mathrm{T} + \frac{1}{2} \lambda w^2 \tag{5}$$

where l represents the loss function, which is used to measure the difference between the prediction result $\hat{y}_i$ and true result y_i. And Ω is a regular term that expresses the complexity of the model and avoids overfitting the model. Then we get the final classifier by optimizing this objective function.

4. EXPERIMENT

4.1 Datesets

In this paper we selected two datasets for experiment, UC-Merced dataset and NWPU-RESISC45 dataset. The UC-Merced dataset was provided by United States Geological

Survey (USGS), and UC-Merced dataset has become one of the most commonly used dataset in HRRS scene classification task (Yang et al. 2010). This dataset consists of 21 classes of scene with the spatial resolution 0.3 meter. There are 100 images of 256×256 pixels in each category. Figure 5 show some sample images for each class included in the UC-Merced dataset.

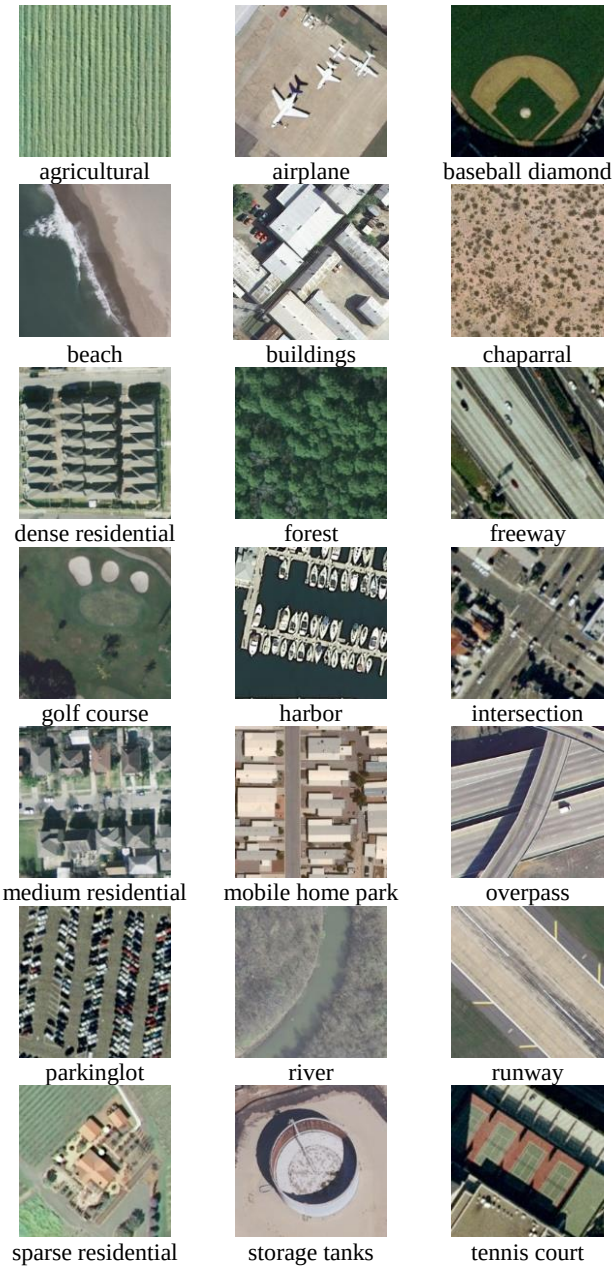

Figure 5: Some images from UC-Merced dataset

Another dataset is NWPU-RESISC45 dataset (Cheng et al. 2017). This dataset was provided by Northwestern Polytechnical University (NWPU). As far as we know, the NWPU-RESISC45 dataset is the most challenging dataset in HRRS image scene classification tasks because it has larger scale on scene categories and image number than other datasets. Furthermore, images of each scene category in NWPU-RESISC45 dataset has rich variations, such as illumination, resolution, shooting angle, background, etc., which also increases the difficulty of scene classification. NWPU-RESISC45 dataset images are divided into 45 scene classes, and spatial resolution of images varies from 30m to 0.2m. The scene categories in NWPU-RESISC45 dataset are: airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, dense residential, desert, forest, freeway, golf course, ground track field, harbor, industrial area, intersection, island, lake, meadow, medium residential, mobile home park, mountain, overpass, palace, parking lot, railway, railway station, rectangular farmland, river, roundabout, runway, sea ice, ship, snowberg, sparse residential, stadium, storage tank, tennis court, terrace, thermal power station and wetland.

4.2 Experimental Protocol

We compare our framework with the VGG16 model, and we also added a traditional support vector machine (SVM) classification method to compare. In order to evaluate the accuracy and efficiency of our framework, we choose three indicators: overall accuracy (OA), time of model training, and kappa coefficient. The two datasets are divided into training set and test set according to the ratio of 80:20. And we randomly selected images from each category for training and testing. Training set is used for model training, and test set is used for evaluating the accuracy of the model.

The VGG16 model was pre-trained on ImageNet dataset, and we fine-tuned the model through both datasets to improve its generalization ability. We used the back propagation algorithm to train the convolutional layers and the fully-connected layers. Stochastic gradient descent algorithm was used by us to optimize parameters. The batch size for each iteration in the training is 64 and the learning rate is 0.001. We also used the dropout method to avoid overfitting. We use the TensorFlow (https://www.tensorflow.org/) framework to implement the deep learning model and use the sklearn library (https://www.scikit-learn.org) to implement the XGBoost and SVM classifier. And our program was run on a PC with 2 3.2GHz 8-core CPUs, 32GB memory and a NVIDIA TITAN X GPU for acceleration.

5. RESULTS AND DISCUSSION

Table 2 shows the performance comparison with the VGG16 and SVM model on UCM dataset and NWPU dataset. As we can see from the Table2, the CNN-XGBoost framework obtains the best classification accuracy than the VGG16 and SVM, and cost less time on training stage than VGG16 model. Our framework achieves satisfying accuracies on two datasets, which is 95.57% and 83.35%, respectively 6.05%, 2.1% higher than VGG16 model and 3.81%, 0.46% higher than SVM. In addition, our framework consumed 70 minutes less than the VGG16 model during the training time of the UC-Merced dataset. However, in the training of large-scale NWPU-RESISC45 dataset, our framework can save more time, saving about 2 hours. Compared to SVM method in training time, our framework consumes only a few minutes more than SVM, and there is almost no difference. The kappa coefficient is a ratio that represents the proportion of errors caused by classification and completely random classification. The kappa value of our framework is also higher than other two methods.

Figure 6 shows the per-class classification accuracies of our framework, SVM and VGG16 model. We can find that our framework performs better than VGG16 and SVM methods in most class. Classes building, sparse residential, medium residential and dense residential are the worst performance in three methods. From the sample images shown in Figure 5, we can see that the three categories buildings, medium residential,

and dense residential are particularly similar, and even human interpreters can hardly distinguish between them. The confusion matrix of the classification results gives detailed information. We use the confusion matrix (Figure 7-10) of the three classification methods to mine their classified error information. The VGG16 model identifies 22.5% of the buildings as dense residential class, and a large proportion of medium residential areas are also misidentified as dense residential category. The scene class with the lowest SVM classification result is dense residential, with 10% misclassification as the building scene class and 10% misrecognition as medium residential class. Although our method also has mistakes in these similar scene categories, our framework is still higher than the other two methods in the overall classification accuracy of the four categories: building, sparse residential, medium residential, and dense residential.

Dataset	Method	OA	Time of model training	Kappa
UC-Merced	VGG16	89.52%	1 hour 40 minutes	0.8899
	SVM	91.76%	24minutes	0.9135
	Our framework	95.57%	30 minutes	0.9535
NWPU-RESISC45	VGG16	81.25%	4 hour 2minutes	0.8083
	SVM	82.88%	1hour 52minutes	0.8250
	Our framework	83.35%	2 hour	0.8297

Table2 Experimental results on the UC-Merced and NWPU-RESISC45 dataset

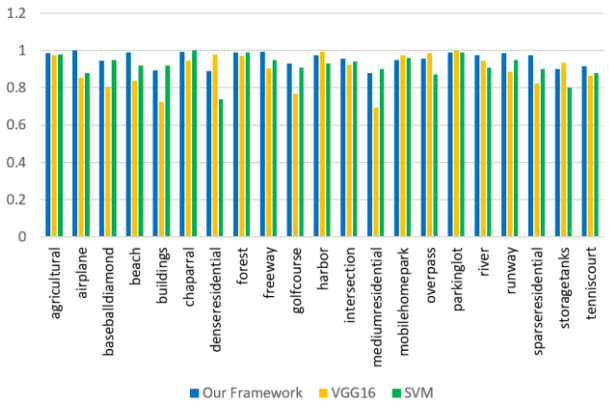

Figure 6 Comparison of the classification accuracy of each class of the three methods on the UC-Mercerd dataset. Blue represents our framework. Yellow represents the VGG16 model. Green represents the SVM.

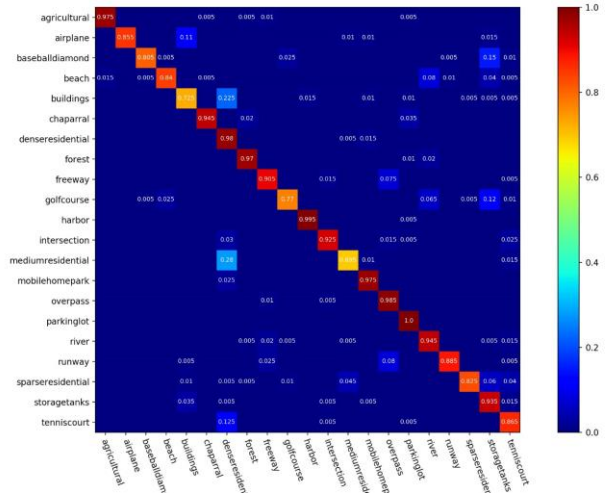

Figure 7 The confusion matrix of VGG16 model's classification result on UC-Merced dataset.

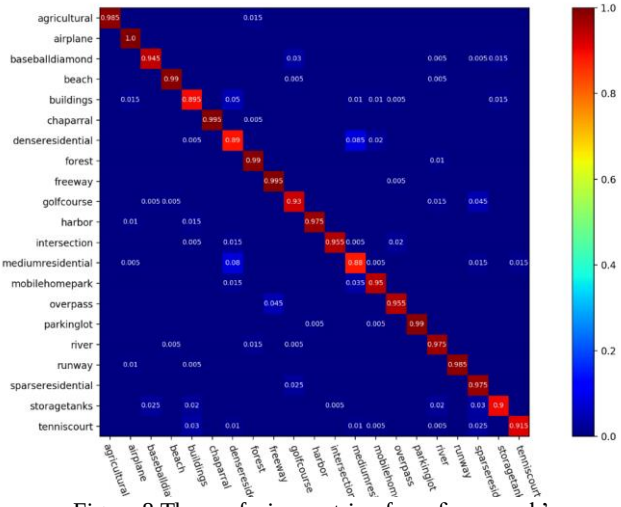

Figure 8 The confusion matrix of our framework’s performance on UC-Merced dataset.

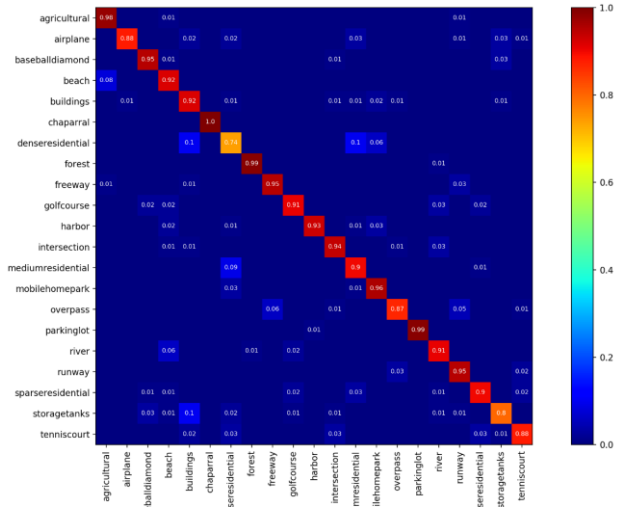

Figure 9 The confusion matrix of SVM classification result on UC-Merced dataset.

Figure 10-13 show us the classification performance of each class of the three classification methods on the NWPU-RESISC45 dataset. For cloud, desert, lake, forest, mountain, sea

ice and wetland scene class, they are easier to distinguish because their texture and color difference is more obvious. Our framework and other two methods have similar classification accuracy. But in category palace and church, their texture features are relatively close, and our classification performance is obviously better than the other two methods.

Summarizing the above discussion, we find that our framework has greatly improved the classification accuracy on dataset UC-Merced compared to other methods. For dataset NWPU-RESISC45, our overall accuracy is still the highest, but since the classification of dataset NWPU-RESISC45 is particularly difficult, our method does not provide much improvement in accuracy. However, it can save nearly half of the training time on dataset NWPU-RESISC45.

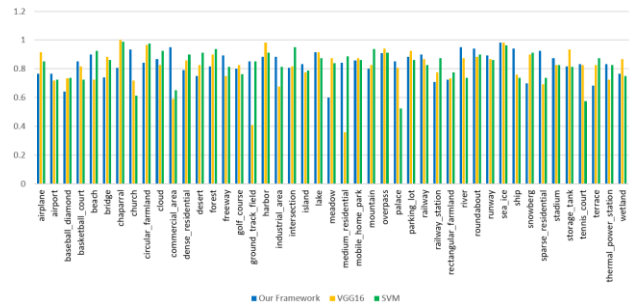

Figure 10 Comparison of the classification accuracy of each class of the two methods on the NWPU-RESISC45 dataset. Blue represents our framework. Yellow represents the VGG16 model. Green represents the SVM.

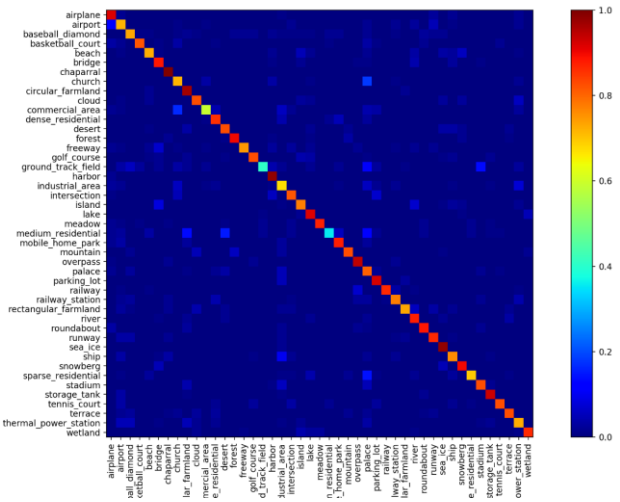

Figure 11 The confusion matrix of VGG16 model's classification result on NWPU-RESISC45 dataset.

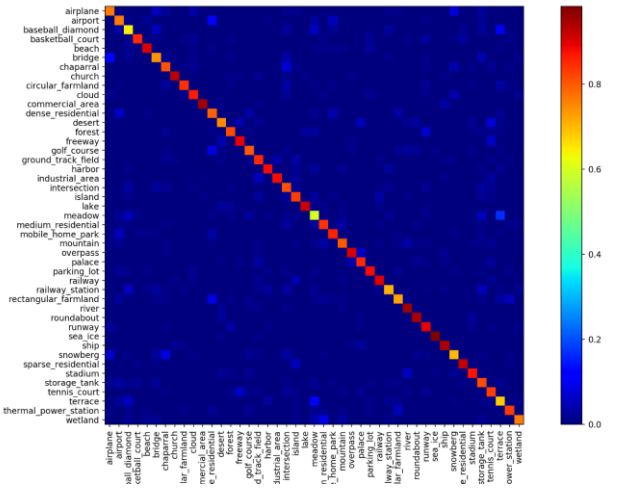

Figure 12 The confusion matrix of our framework’s performance on NWPU-RESISC45 dataset.

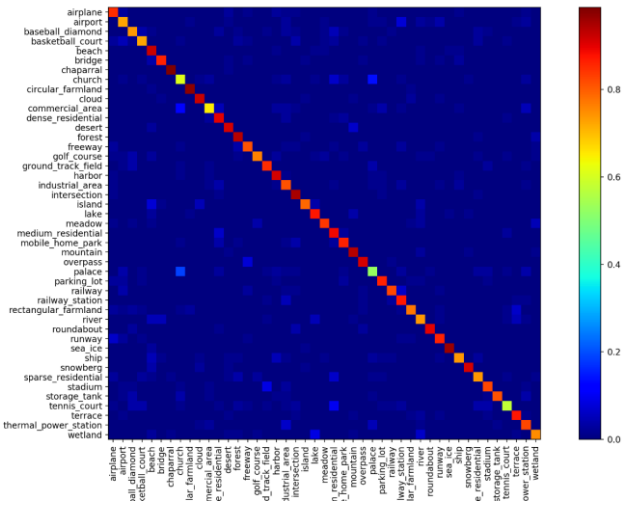

Figure 13 The confusion matrix of our SVM’s performance on NWPU-RESISC45 dataset.

6. CONCLUSION

We know that there are too many parameters in the fully connection layers, which limits the training and using of the deep learning model. Our framework is proposed to solve this problem. We use XGBoost system instead of the fully connection layer in this framework to complete the classification task. And our framework integrates feature extraction capabilities of convolutional neural network and advantages of the XGBoost classifier. We evaluated our framework through the UC-Merced dataset and NWPU dataset, and our framework took less training time but achieved higher accuracy. So, this framework has been proven to be effective for remote sensing images classification. Furthermore, we believe this framework will be more practical for further HRRS scene classification, since it costs less computing resources.

In future, we intend to use multi-source data to assist in remote sensing scene classification such as point of interest (POI), social media data, etc. We also need to explore new technologies to combine these data with location and remote sensing data. Furthermore, it is also very meaningful to explore the application mode of the HRRS image scene classification in the field of urban planning and image retrieval.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China “Eco-environment assessment of Yanhe watershed based on temporal-spatial entropy” [Grant number 41571414].

REFERENCES

Zhang L, Zhang L, Du B, 2016. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geoscience and Remote Sensing Magazine.* Vol.4, pp. 22-40.

G Fu， C Liu， R Zhou， T Sun， Q Zhang, 2017. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sensing.* Vol. 9, pp. 498-519.

Bratasanu D, Nedelcu I, Datcu M., 2011. Bridging the Semantic Gap for Satellite Image Annotation and Automatic Mapping Applications. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing.* Vol. 4, pp. 193-204.

Lienou M, Maitre H, Datcu M., 2010. Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation. *IEEE Geoscience & Remote Sensing Letters.* Vol. 7, pp. 28-32.

Lecun Y, Bengio Y, Hinton G., 2015. Deep learning. *Nature.* Vol.521, pp. 436-444.

Castelluccio M, Poggi G, Sansone C, et al., 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *Acta Ecologica Sinica.* Vol.28, pp. 627-635.

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

Hu F, Xia G, Hu J, et al., 2015, Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sensing.* Vol.7, pp.14680-14707.

Zou Q, Ni L, Zhang T, et al., 2015. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geoscience and Remote Sensing Letters.* Vol. 12, pp. 2321-2325.

Chen T, Guestrin C., 2016. XGBoost: A Scalable Tree Boosting System. *ACM*. pp. 785-794.

Fei-Fei L, Deng J, Li K., 2009. ImageNet: Constructing a large-scale image database. Journal of Vision. Vol. 9, pp. 1037-1037.

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

Zhong Y, Fei F, Zhang L., 2016. Large patch convolutional neural networks for the scene classification of high spatial resolution imagery. *Journal of Applied Remote Sensing.* Vol. 10, pp. 25006.

Marmanis D, Datcu M, Esch T, et al., 2016. Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geoscience and Remote Sensing Letters.* Vol. 13, pp. 105-109.

Cheng, G., Han, J., Lu, X., Remote sensing image scene classification: benchmark and state of the art. *Proc. IEEE.*

Weng Q, Mao Z, Lin J, et al., 2017. Land-Use Classification via Extreme Learning Classifier Based on Deep Convolutional Features. *IEEE Geoscience & Remote Sensing Letters.* 2017, PP(99): 1-5.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2015. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* Vol.3, pp. 993-1022.

Bosch A, Zisserman A, Muñoz X. 2006. Scene Classification Via pLSA. *Computer Vision – Eccv*, pp. 517-530.

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