FEATURE FUSION FOR CROSS-MODAL SCENE CLASSIFICATION OF REMOTE SENSING IMAGE

Wanxuan Geng[a], Weixun Zhou[a,*], Shuanggen Jin[a,b]

[a] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China - (20191211014, zhouwx)@nuist.edu.cn
[b] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai, China

KEY WORDS: Cross-modal, Remote sensing, Scene classification, Feature fusion, Siamese Network.

ABSTRACT: Scene classification plays an important role in remote sensing field. Traditional approaches use high-resolution remote sensing images as data source to extract powerful features. Although these kind of methods are common, the model performance is severely affected by the image quality of the dataset, and the single modal (source) of images tend to cause the mission of some scene semantic information, which eventually degrade the classification accuracy. Nowadays, multi-modal remote sensing data become easy to obtain since the development of remote sensing technology. How to carry out scene classification of cross-modal data has become an interesting topic in the field. To solve the above problems, this paper proposes using feature fusion for cross-modal scene classification of remote sensing image, i.e., aerial and ground street view images, expecting to use the advantages of aerial images and ground street view data to complement each other. Our cross- modal model is based on Siamese Network. Specifically, we first train the cross-modal model by pairing different sources of data with aerial image and ground data. Then, the trained model is used to extract the deep features of the aerial and ground image pair, and the features of the two perspectives are fused to train a SVM classifier for scene classification. Our approach has been demonstrated using two public benchmark datasets, AiRound and CV-BrCT. The preliminary results show that the proposed method achieves state-of-the-art performance compared with the traditional methods, indicating that the information from ground data can contribute to aerial image classification.

1. INTRODUCTION

Scene classification is a hot topic in remote sensing field, which aims to assign a semantic category to the image according to its content, and is also the most intuitive understanding of remote sensing image. Unlike the traditional land use classification, scene classification does not find the corresponding figure category of each pixel or object. Scene classification only focuses on the semantic features of the whole image, and the overall cognition of the image scene. Scene classification pays attention to the global macro information, and generally tends to classify a region as a whole according to the scene semantic information. Therefore, global cognition and semantic information are the two most important parts of scene classification. At present, high-resolution remote sensing image scene classification is widely used, such as urban functional zoning planning (Huang, 2018), vehicle (Schilling, 2018) and ship object detection (Wang, 2019.), etc.

The traditional scene classification of high-resolution remote sensing image is based on a single simple network, from a single perspective, that is, using satellite remote sensing image training model for classification and prediction (Liu, 2018) (Xu, 2020). Although this kind of method is more common, the model training is affected by the image quality of the dataset, and the single perspective will cause the mission of some scene semantic information, which eventually affect the classification accuracy (Cheng, 2017). With the development of remote sensing technology, multi-source and multi view remote sensing data become easy to obtain (Xiong, 2020). The traditional method of "one data source, one model" is slightly outdated. How to do scene classification of cross source data sets has become a major research hot topic.

To solve the above problems, this paper proposes a method based on cross modal model fusion features, which combines the air and ground perspectives, uses the advantages of aerial images and ground street view data to complement each other. We extract the features of similar scenes from different perspectives for fusion, and finally achieve the purpose of improving the accuracy of scene classification.

2. METHOD

2.1 Siamese Network

The cross-modal model is based on Siamese network Siamese network, which consists of two neural networks to form the whole Siamese structure. This kind of "Siamese" is realized by sharing weights by two networks (Liu, 2019). Therefore, Siamese network receives two inputs and transmits it to two neural networks sharing weights to form their own architecture. Finally, the feature representation of each network output is calculated by the same loss function The measurement between them can represent the correlation between the two inputs, thus evaluating the similarity between them. Fig. 1 illustrates the framework of Siamese networks, in which a CNN is the basic unit of the model. It is composed of several layers including convolutional layers, pooling layers, and the fully connected layers, and each plays a vital role in the whole architecture. The convolution layer extracts feature by convolution operation on the input image using convolution kernel, and obtains the feature map as the input of the next layer. The pooling layer compresses the feature maps obtained by convolution layer, and reduces the dimension while retaining important features and avoiding overfitting. The full connection layer is to expand the features obtained from the volume layer or pool layer into one-dimensional feature connection classifier for classification.

* Weixun Zhou，zhouwx@nuist.edu.cn

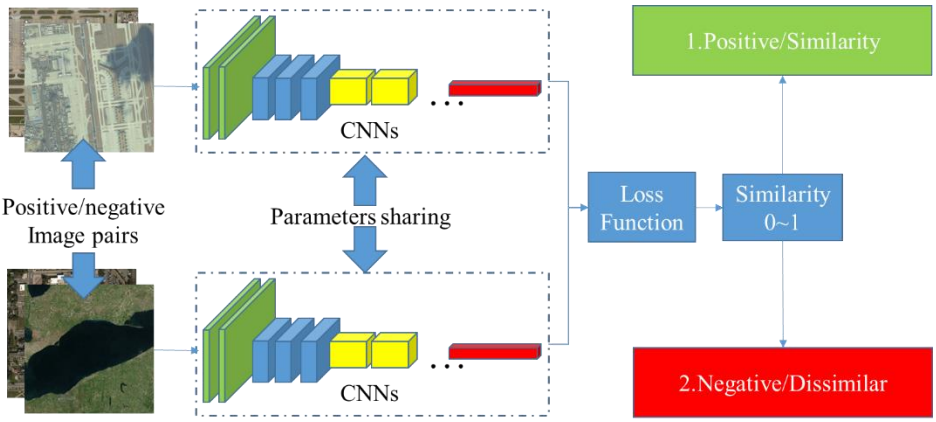

Figure 1. Example of the Siamese Network architecture.

2.2 Cross-modal model

After the brief introduction of Siamese network, we focus on the proposed method. Figure 2 shows the process of cross-modal feature fusion. In this method, we train the cross-modal model by pairing different sources of data based on Siamese network, that is, input aerial remote sensing image at one branch and input ground street view data at the other branch, and specify label to 0 or 1 (1 for the same scene, 0 for different scenes). Then, the model trained is used as the deep feature extractor to extract the deep features of the aerial / ground image pair named feature_a and feature_g. And the features of the two views feature_a and feature_g are fused in case of keeping dimension unchanged. The fused feature is named as feature_fusion. Do the same for the training set and the test set. Finally, a SVM classifier is trained with the fused features for scene classification.

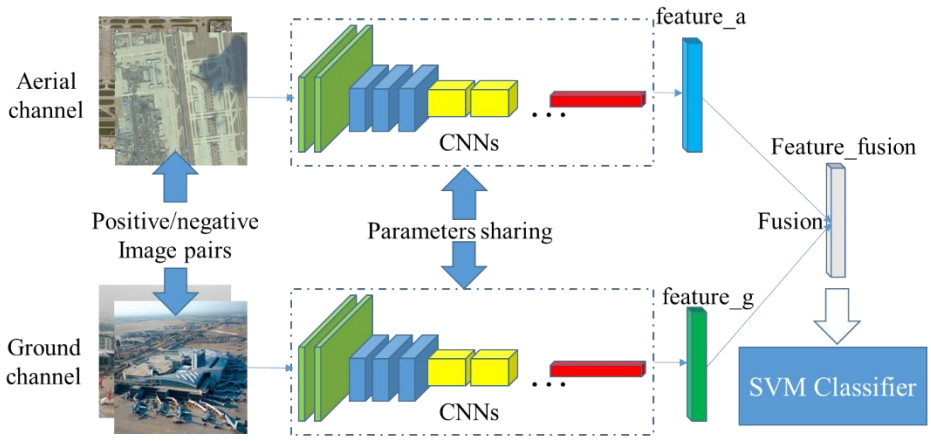

Figure 2. Architecture of the proposed cross-modal feature fusion.

3. EXPERIMENTS AND RESULTS

3.1 Dataset

Airound dataset (Machado, 2020) consists of 1165 pairs of images distributed in 11 categories, including airport, bridge, church, forest, lake, river, skyscraper, stadium, statue, tower and city park. Each sample is composed of a double group, which contains two images from different perspectives, i.e. ground Street perspective image and high-resolution RGB aerial image. All images are paired and manually checked to ensure their correctness. Figure 3 shows class distribution of AiRound and Figure 4 are some examples.

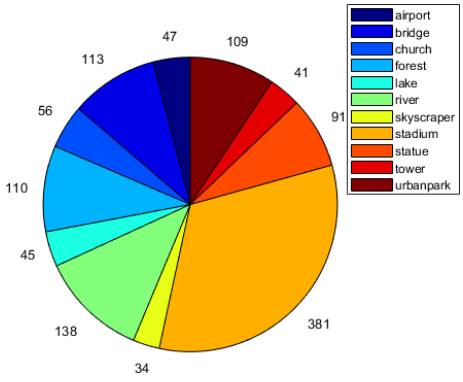

Figure 3. Class distribution of AiRound

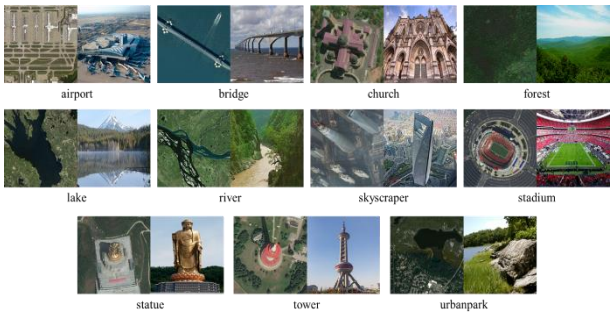

Figure 4. Examples of instances taken from AiRound (The left side of the image pair is aerial image, and the right side is ground street view)

The CV-BrCT dataset (Machado, 2020), which stands for Cross-View Brazilian Construction Type, comprises of approximate 24k pairs of images split into 9 urban classes. The pairs are composed of images from two different views: an aerial view, and a frontal view of a location. Figure 5 shows class distribution of AiRound and Figure 6 are some examples.

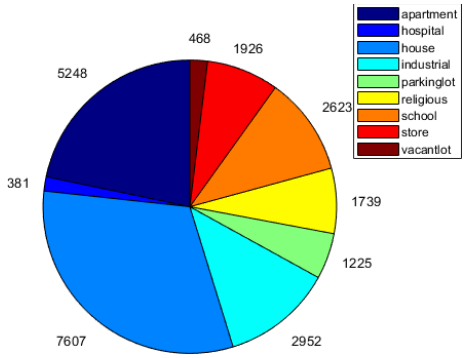

Figure 5. Class distribution of CV-BrCT

Figure 6. Examples of instances taken from CV-BrCT (The left side of the image pair is aerial image, and the right side is ground street view)

3.2 Experimental Details

In order to demonstrate the performance of our proposed method, we conducted four comparative experiments. All the training sets and test sets used in the experiment are the same for a fair comparison. Here is a detailed description of the four groups of experiments and Table 1 shows some parameters of cross modal model training, including the ratio of training set and test set, minimum batch, learning rate, etc.

I. **Single CNN**. A single CNN is used to test data from two different perspectives. AlexNet is used in this experiment.
II. **Siamese Network**. Siamese Network is used to calculate the similarity of two perspectives images to classify. The Siamese network used in this experiment is transformation of AlexNet.
III. **Dataset fusion**. The images of two perspectives in the same dataset are mixed together, that is, each category no longer distinguishes different perspectives. Then, we use the Alex Net to classify.
IV. **Cross-modal features fusion**. That is our proposed method.

Dataset	Ratio(train/test)	Batch size	Learning rate	Iteration
Airound	4:1	100	0.00006	3000
CV-BrCT	4:1	80	0.00008	10000

Table 1. Parameters of cross-modal model training setting

Because of the fusion, experiment III and IV only have one overall classification accuracy. All of our experiments were experimented on a PC with a 3.7-GHz 7-core CPUs, 16-GB memory and a NVIDIA GTX 1660s GPU.

3.3 Results and analysis

Figure 1 shows the results on Airound. It can be seen from the figure that the difference between the two view images in Airound dataset is small, and both can achieve nearly 80% accuracy with a single CNN. The accuracy of using Siamese network is slightly lower, and the data fusion is basically the same. The accuracy of our method is improved by about 3-4%. Figure 1 shows the results on CV-BrCT. This is a bit different from the Airound. In CV-BrCT dataset, the classification accuracy differences of two view data using a single CNN is high. The classification accuracy of aerial data is about 80%, while that of ground street view data is only 65%. Due to the great difference of classification between the two perspectives, Siamese network is about 5% lower in both perspectives, and the effect of dataset fusion is not good. But using our method, the accuracy can reach 80.64%, compared with aerial images. The improvement is not obvious, but for street view data, the improvement is very significant.

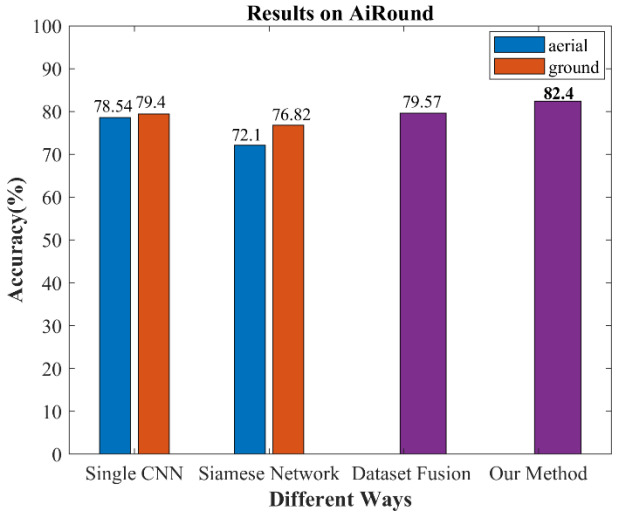

Figure 7. Results of different methods on Airound

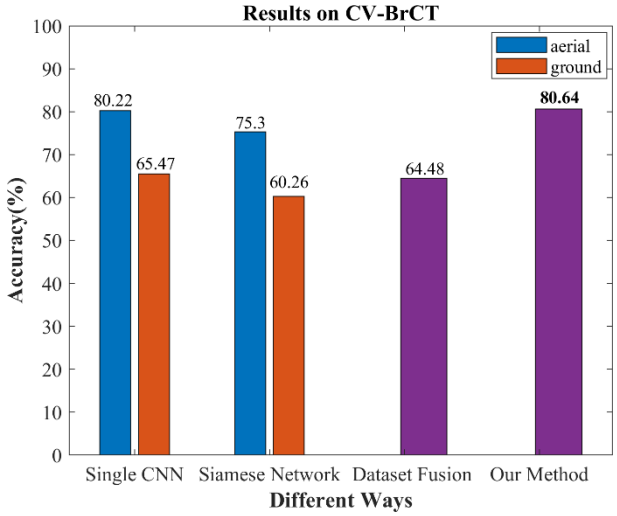

Figure 8. Results of different methods on CV-BrCT

4. CONCLUSION

In this paper, we propose a feature fusion method for cross-modal scene classification. Our method uses cross-modal training between aerial images and ground street view data, which learn features from different perspectives for fusion, and achieve cross source scene classification finally. In addition, the experiments also indicate that the information from ground data and aerial image can contribute to each other in scene classification. Comparison experiments on two datasets has demonstrated that there are performance improvements on both aerial image and ground view image.

ACKNOWLEDGEMENT

This work is supported by The National Natural Science Foundation of China (42001285)， The Natural Science Foundation of Jiangsu Province，China（BK20200813），The Natural Science Foundation of the Jiangsu Higher Education Institutions of China（20KJB420002）and The Startup Foundation for Introducing Talent of NUIST.

REFERENCES

Huang, B., Zhao, B., Song, Y.,2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sensing of Environment 214(2018):73-86.

Schilling, H., Bulatov, D., Niessner, R., Middelmann, W., Soergel, U.,2018. Detection of Vehicles in Multisensor Data via Multibranch Convolutional Neural Networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(11), pp. 4299-4316.

Wang, R., Xu, F., Pei, J., Wang, C., Huang, Y., Yang, J., Wu, J.,2019. An Improved Faster R-CNN Based on MSER Decision Criterion for SAR Image Ship Detection in Harbor. IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, pp. 1322-1325

Liu, Q., Hang, R., Song, H., Li, Z., 2018. Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification. IEEE Transactions on Geoscience and Remote Sensing, 56(1), pp. 117-126.

Xu, K., Huang, H., Li, Yuan., Shi, G., 2019. Multilayer Feature Fusion Network for Scene Classification in Remote Sensing. IEEE Geoscience and Remote Sensing Letters, 17(11), pp. 1894-1898.

Cheng, G., Han, J., Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE, 105(10), pp. 1865-1883.

Xiong, W., Xiong, Z., Zhang, Y., Cui, Y., Gu, X., 2020. A Deep Cross-Modality Hashing Network for SAR and Optical Remote Sensing Images Retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5284-5296.

Liu, X., Zhou, Y., Zhao, J., Yao, R., Liu, B., Zheng, Y., 2019. Siamese Convolutional Neural Networks for Remote Sensing Scene Classification. IEEE Geoscience and Remote Sensing Letters, 16(8), pp. 1200-1204.

Machado, G., Ferreira, E., Nogueira, K., Oliveira, H., Brito, M.,2021. AiRound and CV-BrCT: Novel Multiview Datasets for Scene Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 488-503.