A new data augmentation method of remote sensing dataset based on Class Activation Map

Wei Zhang¹, Yungang Cao*¹
¹ The Faculty of Geosciences and Environmental Engineering, Southwest Jiao-tong University, Chengdu 611756, China
* Corresponding author’s e-mail: yungang@swjtu.cn

Abstract. Remote sensing image scene classification is a significant direction in the field of remote sensing research. The method based on deep learning has become the most popular method in recent years because it can realize the automatic feature extraction and classification of remote sensing images. The deep learning requires a large number of samples for training and consumes large computing resources, and the data augmentation can alleviate the problem of insufficient samples. The image manipulation is one of the most commonly methods, but it may cause the loss of key information in the image. In this paper, we proposed an improved supervised data augmentation method based on Class Activation Map (CAM) and image manipulation, and then used this method to augment the high-resolution remote sensing images of NWPU dataset. We utilized three CNNs networks to count the classification accuracy of the remote sensing images. The experimental results show that the proposed method increases the accuracy of scene classification by more than 0.4%. The CAM-based methods provide a new technical support for the scene classification of remote sensing images based on deep learning.

1. INTRODUCTION

After the launch of the first artificial satellite in 1957, the resolution of remote sensing images, especially the spatial resolution, has been greatly improved with the continuous progress of earth space science and technology in the past sixty years. At present, the spatial resolution of IKONOS, WorldView, GeoEye-1, GF-2, ZY-3, and other satellite images has reached meter or even sub-meter level. Compared with natural images, high-resolution remote sensing images have richer spectrum information, more complex shape and texture features, as well as scene semantic information [1],[2]. Therefore, it is necessary to extract meaningful information in high-resolution images through a series of interpretation processing. Scene classification is an important way to understand remote sensing image information, that is, to understand the “scene” expressed by image and label it as a specific semantic category. It is of great significance for the interpretation of images and the understanding of the real world, and it is also one of the hotspots in the field of remote sensing [3].

According to different feature extraction ways, scene classification methods can be divided into three categories: low-level visual feature-based method [4],[5], middle-level visual expression-based method [6],[7], and high-level visual information-based method [8],[9]. The traditional scene classification methods mainly extract the low-level and middle-level features, which represent the specific color, shape and texture information. From the perspective of the research status of remote sensing image scene classification, feature extraction and classification of remote sensing images tend to be automated and the deep learning method has become the most popular classification method in recent years. The deep learning is mainly to extract the high-level semantic information from the
image, representing the advanced abstract information [10],[11]. However, it is very difficult to improve the accuracy of remote sensing scene classification only by traditional methods or deep learning methods. Therefore, how to effectively improve the accuracy is still an important subject with great significance.

The deep learning requires a large number of samples for training and consumes more computing resources. However, it takes a long time to produce a large number of high-quality training samples, so how to improve the recognition accuracy of features with small-size samples is a difficult point at the present stage. Recently, the data augmentation is an effective method, which can greatly reduce the time required for manual labeling and improve the accuracy of object detection. The technology based on image processing is one of the commonly used methods, but this method may cause the loss of key information. In this paper, based on Class Activation Map (CAM) [12],[13] and traditional data augmentation methods, such as random cropping, random translation, and noise injection [14], a supervised improved data augmentation method is proposed, and the high-resolution remote sensing image dataset has been adopted as experimental data. The experiments indicated that the proposed method in this paper further improves the accuracy of object recognition compared with the traditional methods, and provides new technical support for the scene classification of remote sensing images based on deep learning.

2. METHODOLOGY

This paper proposes a supervised data augmentation method based on Class Activation Map. Firstly, this method applies CNNs to train remote sensing dataset to obtain an initial model. Then, the class activation map generated by this initial model can be utilized to obtain the probability of the target in the samples. Finally, the sample amplification has been carried out under the supervision of CAM to ensure that the key information will not be lost.

2.1. Intersection over True

Intersection over Union (IoU) is a standard for measuring the accuracy of detecting the corresponding object in a specific dataset, and the higher the degree of correlation is, the higher the value is [15]. IoU calculates the ratio of intersection and union sets of “Predicted” and “Ground-truth”. In order to evaluate whether the retention of the amount of sample information, which is the value to quantitatively analyze key information loss, generated by the sample augmentation method is qualified, this paper proposed an information loss evaluation method - Intersection over True (IoT) based on the IoU. “Ground-truth” in IoT is defined as the position of the object prediction box in the original image obtained by the pre-training model; “Predicted” is the position of the new image boundary range in the original image after image processing, such as random translation and random cropping; “True” is the area of “Ground-truth”.

\[
\text{IoU} = \frac{\text{Intersection}}{\text{Union}} \quad \text{IoT} = \frac{\text{Intersection}}{\text{True}}
\]

Figure 1. The calculation method of Intersection over True.
2.2. Class Activation Map
Bolei Zhou et al. proposed the class activation map (CAM) in CVPR-2016, which can solve the defect problem of insensitive representation of class information in the field of image processing and recognition [13]. That is, CAM can intuitively explain why the neural network can determine the category of information in the image. The class activation mapping is a technique which can identify the importance of the image regions by projecting back the weights of the output layer on to the convolutional feature maps. The CAM calculation formulas for a certain target in the image are derived as follows:

\[
S_c = \sum_k w^c_k \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w^c_k f_k(x,y) 
\]

(1)

\[
M_c(x,y) = \sum_k w^c_k f_k(x,y) 
\]

(2)

where \(f_k(x,y)\) represents the activation value of the unit \(k\) of the last convolutional layer in the spatial grid \((x,y)\), and the result of each unit \(k\) passing GAP is \(\sum_{x,y} f_k(x,y)\). For the target category \(c\), the input value of the classifier SoftMax is \(S_c\), where \(w^c_k\) represents the weight of the unit \(k\) of the classifier of the object category \(c\). \(M_c(x,y)\) is the class activation map for class \(c\), where each spatial element is given by Eq. (2). And the CAM of category \(c\) can be obtained by resampling \(M_c(x,y)\) into the original image size. A class activation map for a particular category indicates the discriminative image regions used by the CNN to identify that category, as shown in Figure 2.

![Figure 2. The CAMs of two classes from ILSVRC. The maps highlight the discriminative image regions used for image classification, the head of the animal for briard and the plates in barbell.](image-url)

2.2.1. Random cropping or random translation method based on CAM
Image cropping is to randomly sample a part of the original image, and then adjust the size of this part to the original image size. Random translation is that images are moved several pixels along the horizontal, vertical or both directions. This method is highly practical because the appropriate shift model of images can traverse all levels of image features, so as to effectively improve the training effect of the model. However, both methods have some defects, that is the selected target area may not contain the real target area or a large number of key information is lost, which cause network to train error tags. As shown in Fig. 3-(I), the real label of the sample is “baseball diamond”. However, if random cropping method is adopted, a large number of key information about baseball diamond in the upper left corner is lost. In this case, the value of IoT is less than 0.5, which make the sample more inclined to “meadow”.

The flow chart of the random cropping and random translation method based on the CAM is shown in Figure 3-(I). Firstly, the CAM of sample images are obtained by pre-training the initial model. Then, according to the target probability information of the CAM, Map is converted into a binary map, and the approximate area of the target can be obtained through the binary map and regional connectivity. The processed samples are randomly cut or translated, and IoT is used to evaluate whether there is a loss of key information. If the IoT value is greater than 0.5, the sample is qualified, and vice versa.
2.2.2. CAM-based gaussian noise injection

Gaussian noise is one of the common methods in noise injection image augmentation. Gaussian noise caused by bad lighting and high temperature is a kind of noise with Gaussian distribution characteristics, which is obvious in RGB images. In this paper, an improved method is proposed by combining CAM with Gaussian noise. The flow chart is shown in Figure 3-(II). Firstly, CAM of sample image is obtained by pre-training initial model. The weight information added by Gaussian noise is calculated according to the probability information of the target of CAM. The area with higher probability of the target injects less noise data. The calculation formula of the weight information is shown in Eq. (3). The Gaussian noise injection augmentation data calculated by the CAM can retain more information about the target, and distinguish the foreground and background information. This method can make the deep learning network to have more targeted learning feature information.

\[ G_c(x, y) = 1 - \sum_k w_k f_k(x, y) \quad (3) \]

where \( G_c(x, y) \) represents the weight information of the sample adding Gaussian noise, which is calculated by CAM. \( G_c(x, y) \) is resampled to the original image size and multiplied by the Gaussian noise matrix. Finally, the original image is added to obtain the enhanced image.

![Figure 3](image-url)

Figure 3. The experimental flowchart via CAM-based data augmentation methods. (I) the process of random cropping or random translation method based on CAM. (II) the process of CAM-based gaussian noise injection method.

3. EXPERIMENTS

3.1. Study Dataset

NWPU-RESISC45 is a remote sensing dataset constructed by Northwestern Polytechnical University, which can be used as a publicly available benchmark for remote sensing image scene classification [16]. This dataset contains 31,500 images with a pixel size of 256*256, and consists of 45 scene classes. These 45 scene classes are as follows: 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.

3.2. Experimental Environment and Parameters

The proposed model was implemented using the PyTorch framework, which was first introduced by
Facebook in 2017. In this paper, we utilize fixed-sized images (256×256 px) in the augmented remote sensing dataset to train three classic convolutional neural networks (ResNet-18, SqueezeNet, and DenseNet-121), and start training the model with a minibatch size of 16. The learning rate was initially set to 0.001 and adjusted by the exponential decay function:

\[ \alpha = k^{\text{epochnum}/\text{step}} \ast \alpha_0 \]  

(4)

where \( k \) is a coefficient with a value of 0.85, \( \text{epochnum} \) is the number of iterations, and the \( \text{step} \) size is 10. The network converged in 200 epochs. The experiment performed on an NVIDIA GTX 2080Ti (GPU), a 2.4 GHz Intel E5-2620 v3 (CPU), and a 16 GB RAM.

### 3.3. Comparisons and discussion

To evaluate how effectively the proposed method executes the augmentation task, we first augment the remote sensing dataset (NWPU-RESISC45) by using three groups of methods. The methods are random cropping/CAM-based random cropping (method 1), random translation/CAM-based random translation (method 2), and Gaussian noise injection/CAM-based Gaussian noise injection (method 3). And then we selected three classic convolutional neural networks - ResNet-18, SqueezeNet, and DenseNet-121 [17]-[19] – to verify the quality of these datasets. The classification accuracy results are shown in Table I: 1) The accuracy of using ResNet-18 for scene classification tasks on remote sensing datasets is good. And compared with the original methods, the three CAM-based augmentation in this paper have improved the accuracy of scene classification by more than 0.4%; 2) The overall classification accuracy of SqueezeNet in the results is not as good as other methods, and the result of method 2 is poor. SqueezeNet is a lightweight network with fast training speed and very few parameters, but its feature extraction ability is poor. Therefore, the insufficient extraction of key information of the object leads to the insignificant improvement effect of the CAM-based method. 3) Compared with ResNet, DenseNet by using CAM-based method has poor accuracy improvement, but has the highest overall classification accuracy.

| Method   | ResNet-18 Original | ResNet-18 Proposed | SqueezeNet Original | SqueezeNet Proposed | DenseNet-121 Original | DenseNet-121 Proposed |
|----------|--------------------|--------------------|--------------------|--------------------|-----------------------|-----------------------|
| Method 1 | 90.89              | 91.33              | 83.39              | 83.88              | 92.17                 | 92.53                 |
| Method 2 | 90.98              | 91.49              | 80.19              | 80.02              | 91.47                 | 91.68                 |
| Method 3 | 88.02              | 88.42              | 81.41              | 81.55              | 90.91                 | 91.17                 |

In order to further analyze and research the influence of the proposed methods on the different categories of the remote sensing dataset, we counted the classification accuracy change values of the 45 category samples in the test dataset (Fig. 4). In this paper, the CAM is defined as an effective improvement for a certain category if the test accuracy of the two methods based on CAM is improved compared with the original method. Otherwise, the CAM is an invalid method if the test accuracy of the two CAM-based methods is not significantly improved or decreased. We found that the CAM-based methods are effective for basketball court, commercial area, desert, freeway, golf course, island, mobile home park, mountain, parking lot, railway station, terraces, thermal power stations, and wetlands. The foreground and background information of these objects are quite different, so the key information in the images can be quickly highlighted by CAM. But the classification accuracy of the scenes, such as overpass, harbor, industrial area, rectangular farmland, sea ice, snowberg, storage tank, circular farmland, forests and chaparral, has not been improved. The invalid category of the experiment is shown in the Figure 5. The features of these scenes, especially the rectangular farmland, are not distributed in a certain area of the image, but in the entire image. The geometric transformation
or the noise injection methods are not loss the key information of the image, so the CAM-based augmentation method is not suitable for this kind of samples with uniform feature distribution.

**Figure 4.** The classification result statistics of different scene categories

**Figure 5.** The invalid category of the experiment. These scenes have poor classification accuracy improvement by using the CAM-based methods compared with the original methods.

### 4. CONCLUSION

In this paper, we combined CAM and data augmentation method to construct a supervised data augmentation method based on Class Activation Map. The CAM-based methods has a better ability to preserve the key information which can help CNNs extract more image features. Compared with original method, the proposed augmentation method has better compatibility because it improves the overall accuracy of scene classification. The use of the CAM can provide certain technical support for traditional data augmentation methods such as image manipulations to help CNNs solve the problem of insufficient samples and improve the classification accuracy. However, the CAM-based augmentation also still has the problem of poor performance in the amplification of samples with uniform feature distribution.

**References**

[1] Gamba, P. (2013). Human settlements: a global challenge for EO data processing and
interpretation. Proceedings of the IEEE, 101(3), 570-581.

[2] Gong, C., Han, J., & Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 105(10), 1865-1883.

[3] Xian, G., Xin, H., Zhang, L., et al. (2016). Support tensor machines for classification of hyperspectral remote sensing imagery. IEEE Transactions on Geoscience & Remote Sensing, 54(6), 3248-3264.

[4] Lowe, D. G. (2004). Distinctive image feature from scale-invariant key points. International Journal of Computer Vision.

[5] Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE (Vol.24, pp.971-987).

[6] Hu F., Xia G. S., Hu J., et al. (2015). Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing, 7(11): 14680-14707.

[7] Sheng G., Yang W., Xu T., et al. (2012). High-resolution satellite scene classification using a sparse coding based multiple feature combination. International journal of remote sensing, 33(8): 2395-2412.

[8] Hinton G. E., Osindero S., Teh Y. W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7): 1527-1554.

[9] Vincent P., Larochelle H., Lajoie I., et al. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(Dec): 3371-3408.

[10] Zhu X. X., Tuia D., Mou L., et al. (2017) Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine, 5(4): 8-36.

[11] Kussul N., Lavreniuk M., Skakun S., et al. (2017). Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5): 778-782.

[12] Selvaraju R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, IEEE International Conference on Computer Vision, pp.618-626.

[13] Zhou, B., Khosla, A., Lapedriza, A., et al. (2016). Learning Deep Features for Discriminative Localization. CVPR. IEEE Computer Society.

[14] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1).

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39(6), 1137-1149.

[16] Gong, C., Han, J., & Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 105(10), 1865-1883.

[17] He, K., Zhang, X., Ren, S., et al. (2016) Deep Residual Learning for Image Recognition. computer vision and pattern recognition.

[18] Iandola, F. N., Han, S., Moskewicz, M. W., et al. (2016). Squeezenet: alexnet-level accuracy with 50x fewer parameters and <0.5mb model size.

[19] Huang, G., Liu, Z., Laurens, V., & Weinberger, K. Q. (2016). Densely connected convolutional networks. IEEE Computer Society.