ISSN 2616-5775 Vol. 6, Issue 7: 131-137, DOI: 10.25236/AJCIS.2023.060717

Research on remote sensing image object detection based on convolutional neural network

Meiyan Li[1,2,a], Joey S. Aviles[1,b,*]

[1] Graduate School, Angeles University Foundation, Angeles City 2009, Philippines
[2] School of Information Engineering, Baise University, Baise, 533000, China
[a] li.meiyan@auf.edu.ph, [b] aviles.joey@auf.edu.ph
**Corresponding author*

Abstract: *With the update of remote sensing equipment and the improvement of technology, remote sensing images not only have advantages in quantity, but also improve in imaging quality. Convolutional neural network has been widely used in object detection, which can extract targets from a large number of images and has a very powerful feature extraction capability. The detection technology for remote sensing images has a promoting role. This paper introduces the application of remote sensing image, analyzes several common convolutional neural network technologies and features, and analyzes the application of convolutional neural network in remote sensing images. Through comparative analysis, the efficiency of convolutional neural network technology is relatively high in remote sensing image target detection, which indeed solves the problems of low detection accuracy and high miss detection rate of remote sensing targets, remote sensing target characteristics and mainstream algorithms in current remote sensing images.*

Keywords: *Remote sensing image, Feature ability, convolutional neural network, Object detection*

1. Introduction

Due to the characteristics of remote sensing images such as large field of view, high background complexity, special viewing angle, target rotation, and small targets, it not only provides more regions of interest but also brings more complex background information, posing a huge challenge to target detection.The traditional remote sensing image object detection methods are usually based on image processing methods, which first implement geometric feature extraction, texture processing, and threshold segmentation, and then use templates for matching and background modeling to accurately and comprehensively detect and recognize the target.Traditional image processing methods such as scale invariant feature transformation and gradient direction histogram are designed based on artificial experience. Although these methods have good detection and recognition performance in specific environments, they have a strong dependence on prior knowledge, which can lead to poor adaptability and generalization ability of model algorithms.In recent years, deep learning has shown a rapid development trend in object detection in natural scene images. Due to its success, many scholars have attempted to use deep learning methods for multi-disciplinary research, especially in remote sensing image object detection. Yang Y & Newsam S (2021) proposed an FPN feature pyramid network with profound influence and outstanding operability[1]. It can provide feedback on the top-level feature images in the network architecture layer by layer based on practical situations and related needs, and can also use reasonable methods to fuse them into the feature maps of the previous layer, in order to promote the improvement of low-level features in specific semantic intensity, and can also be implemented purposefully at various scales Targeted target detection.It should be pointed out that this feature map (multi-scale) has better robustness when facing objects of various sizes (especially small objects).

Object detection plays an important supporting role in computer vision, an important branch of artificial intelligence, and is one of the important research trends in current scientific research. Remote sensing image target detection and recognition is one of the most fundamental and essential tasks in the field of optical remote sensing image processing. The characteristics of remote sensing images mainly include large field of vision, high background complexity, special angle of view, target rotation, small targets, etc[2]., which makes remote sensing images provide more regions of interest and more complex background information, which brings great challenges to the task of target detection. Due to the rapid development of remote sensing technology, there are more and more remote sensing detectors. At the same time, with the higher resolution of remote sensing images, the multispectral remote sensing images are becoming larger and larger, which leads to the increasing amount of remote sensing image data. Massive remote sensing image data not only provides more abundant resources for remote sensing target detection, but also puts forward new requirements for it, that is, how to effectively and accurately detect and recognize remote sensing image data, which is also a hot issue in the field of remote sensing image processing. However,due to the sensor's own reasons, such as physical and technical characteristics, observation angle and imaging mechanism, there are different noises in the collected remote sensing images, causing inevitable interference to the remote sensing image data. In addition, the acquisition of remote sensing images will inevitably interfere with the outside world, such as weather, cloud occlusion, illumination, object color, etc., which makes the target detection task vulnerable to various external factors in different environments[3].

Therefore, it is necessary to choose the algorithm with high technical efficiency of convolutional neural network to solve the problems of low detection accuracy and high miss detection rate of remote sensing targets, remote sensing target characteristics and mainstream algorithms in current remote sensing images. At the same time, the shallow features of high-resolution optical remote sensing images should be used to improve the detection accuracy and performance of remote sensing targets, reduce the missed detection rate, and realize the high-precision detection of small-target remote sensing targets.

2. Related research

2.1 Object detection technology based on traditional edge detection algorithms

Edge is one of the most important features of images and is widely used in remote sensing image detection. Edge detection algorithms usually use the contrast between the target and background in remote sensing images for object detection.Currently, there are three commonly used edge detection operators: Roberts operator, Sobel operator, and Log operator. After analyzing various operators, relevant scholars have proposed the Canny algorithm. Compared to edge detection algorithms using other operators, this algorithm achieves the best edge detection performance and is most advantageous for human eye recognition and judgment[2].

2.2 Target detection technology based on traditional segmentation algorithms

Traditional object detection techniques based on closed value segmentation usually utilize the differences in grayscale characteristics between the target and background in optical remote sensing images for object detection.Specifically, the image is regarded as a combination of two types of regions with different gray levels, and an appropriate threshold is selected to determine whether each pixel in the image should belong to the target or background region, so as to obtain a binary image. There are many commonly used methods, mainly including Otsu threshold segmentation algorithm, watershed segmentation algorithm, and iterative global threshold segmentation algorithm.

2.3 Object detection technology based on visual saliency

Visual saliency based object detection technology is an intelligent detection algorithm model that simulates the human eye's ability to quickly focus on areas of interest. It mainly includes two detection methods: one is a top-down saliency detection method; The second method is a bottom-up significance detection method.The top-down saliency detection method is a method of actively searching for target saliency maps based on existing tasks; The bottom-up saliency detection method is a design method that imitates the human eye's instinctive response to things and then obtains saliency maps.For the first detection method, it takes longer and has a slower running speed, while the second detection method has a faster running speed. Currently, there have been many significant object detection methods (from bottom to top), typical of which are ITTI algorithm[4], GBVS algorithm[5], FT algorithm[6], etc.

However, it is particularly emphasized that it is difficult to obtain ideal object detection results directly using visual saliency algorithms in optical remote sensing images. Therefore, it is mainly suitable for object detection in simple background situations or for extracting candidate regions of objects.

2.4 Target detection technology based on traditional machine learning algorithms

For a long time before, traditional shallow structure machine learning algorithms played a significant role in the entire remote sensing image object detection, and the vast majority of object detection tasks were based on conventional machine learning algorithms.Machine learning algorithms first combine specific objectives and carry out targeted and comprehensive design around feature extraction algorithms; Then, the obtained feature vectors are conveyed to a specific feature classifier in a reasonable, standardized, and efficient manner; Finally, implement training.The disadvantage is that conventional machine learning algorithms are difficult to extract features, which not only requires a lot of time investment, but also requires improvement and optimization of the feature extraction algorithm to obtain a more reasonable algorithm. Another issue is that classifiers in traditional machine learning algorithms, such as SVM network structures, have poor generalization ability and lack the ability to express complex functions[7]. Therefore, at this stage, traditional machine learning algorithms are gradually being phased out.

2.5 Target detection technology based on deep convolutional neural networks

Deep convolutional neural networks have shown a rapid development trend, with extensive support from many large commercial companies. The biggest motivation for these companies to invest a large number of researchers in research is commercial interests, which stem from people's demand for intelligent living.It should be pointed out that due to the fact that most datasets are publicly available, such as pedestrian detection and facial data sets, they will also promote the development of related technologies. Due to the unique characteristics of remote sensing images, most units with aerospace and aviation equipment can obtain remote sensing images. Therefore, there are not too many datasets for remote sensing images, and there are fewer datasets for target detection in remote sensing images[8].Over time, many experts and scholars have applied deep convolutional neural networks to the field of remote sensing. Deng L & Yu D(2022) integrated R-CNN into remote sensing images for the first time to implement object detection[9]. LecunY et al (2019) applied Faster R-CNN to remote sensing image object detection tasks[10]. Lee C et al (2020)explored a new detection method based on the detection framework possessed by SSD, which can achieve variable input image scale[11]. Ranzato M et al (2021) proposed a remote sensing image object detection method based on YOLOv2, which significantly improves the accuracy of remote sensing object detection compared to traditional algorithms. However, the application effect for small object detection in remote sensing images is not very ideal. It should be noted that if there are differences in the specific types of targets, or if the distance during the image shooting process is uncertain, it will result in differences in the specific size of the targets.For remote sensing images, in order to better adapt to various characteristics of remote sensing images, such as differences in shooting angle, height, and resolution, multi-scale features need to be given high attention. Based on previous research results, Scherer D et al (2021)proposed a more advanced, practical, and efficient deep feature pyramid, which greatly enhances the network's ability to extract multi-scale target features[12].For small target detection in dense scenes, Wang T et al (2021)designed a multi-layer feature fusion structure to enhance the Semantic information of shallow features, and then further enhanced the feature representation capability of the network through the multi-scale Receptive field module. Gu J et al (2021)proposed a multi-scale object detection method using channel attention driven rotational invariance depth features for high-resolution remote sensing images, which have the characteristics of multi-scale, arbitrary direction, shape change, and dense arrangement[13]. This method achieved better detection performance than existing methods[14]. Xu B et al (2022)performed target detection tasks on cars in remote sensing images by fine-tuning the HDCNN network, while loffe S & Szegedy C(2021) performed forest fire detection on remote sensing images by fine-tuning the AlexNet network[15]. However, due to the relatively old AlexNet network and the fact that AlexNet only serves as a feature extractor, the idea of using shallow machine learning algorithms for target detection still failed to emerge. Srivastava N et al (2021) focuses on the intelligent detection and recognition of ship targets in optical remote sensing images, conducting research on key technologies such as feature extraction, feature fusion, and target detection and recognition of ship targets, achieving high-precision detection and recognition of sea surface ship targets.Y Lin M et al (2019)has conducted research on object detection in visible light remote sensing images based on attention mechanism[16]. Rawat W &Wang Z(2021)utilized a method similar to RCNN to combine convolutional neural networks with SVM for remote sensing vehicle target detection tasks. Potluri S et al (2021) conducted online analysis of difficult cases in the RPN segment, using the RealBoost algorithm to replace the Fast RCNN algorithm for vehicle detection in remote sensing images, and using ZF-Net as the basic network, making the entire detection process more complex.

ISSN 2616-5775 Vol. 6, Issue 7: 131-137, DOI: 10.25236/AJCIS.2023.060717

3. Structure of Convolutional Neural Networks

The structure of CNN is a combination of different structural layers, each part performing its own function to achieve feature extraction and classification. The input images will go through the internal structure of CNN in turn, and the internal feature information will be extracted through the internal structure. The features extracted from them will predict the loss function through reverse error propagation, and adjust the weight of the parameters in the network according to the prediction results, constantly reduce the value of the loss function to improve the learning effect, and finally complete the classification through the full connection layer. The following is a specific introduction to the composition structure.

3.1 Convolutional layer

Convolutional layer, as one of the important components of CNN, is mainly used to extract feature information through convolutional operations[17]. The reason why convolution operation has received widespread attention is mainly due to its extraction ability, which enhances certain required features through convolution operation while reducing unnecessary interference. Different positions in the input image have different features, and convolutional kernels use different scales to obtain different features. For deep convolutional neural networks, convolutional kernels can be used to extract features sequentially from simple to complex.

For example, for a two-dimensional convolutional array with an input size of 3x3, a 2x2 convolutional kernel is used. The shaded part in the figure is the process of calculating the first element. The corresponding elements of the shaded part are multiplied and then added to obtain the operation result of the first position of the output. Follow this step to slide and obtain the corresponding operation results for all elements.

If the detection is for a color image, a three-dimensional convolutional kernel is used. If the color image is 6 * 6 * 3, 3 represents three color channels. Similarly, two-dimensional convolution uses a 3 * 3 filter. The three-dimensional convolution uses a three-dimensional filter with dimensions of 3 * 3 * 3, corresponding to the red, green, and blue channels. Slide the 3D convolutional kernel of 3 * 3 * 3 on the color image, and perform operations on the corresponding values of the convolutional kernel with the corresponding channels of red, green, and blue, in order to obtain the output results at the corresponding positions. If the input image is (representing the width, height, and number of channels of the image), the size of the convolutional kernel is F * F, the step size is S, the size of the pad filled is P, the number of convolutional kernels is K, and the resulting output image is . The calculation formulas are shown in (1) ,(2)and (3).

$$W_2 = \frac{W_1 - F + 2P}{S} + 1 \tag{1}$$

$$H_2 = \frac{H_1 - F + 2P}{S} + 1 \tag{2}$$

$$D_2 = K \tag{3}$$

3.2 Activate layer

The problems we encounter in real life are usually nonlinear, so we need to enhance the nonlinear ability of convolutional neural networks. The role of the activation layer is to increase the ability of nonlinear processing, use a series of activation function for nonlinear processing, and then simulate the function model we need to solve. The activation function has the advantage of fast convergence speed of Rectified-Linear Unit t (ReLU), and can avoid the disappearance of gradient, so it has become the activation function that is currently used more[18]. Other commonly used methods include Sigmaid, Tanh, Softmax, etc.

3.3 Pooling layer

After completing the convolution operation, we will obtain the extracted feature information, but these features are still relatively complex and massive. In order to improve subsequent processing

ISSN 2616-5775 Vol. 6, Issue 7: 131-137, DOI: 10.25236/AJCIS.2023.060717

efficiency, the pooling layer can be used to simplify the information of the extracted features and remove redundant information. If there is no pooling layer, directly processing the extracted information not only requires a huge amount of computation, but also time-consuming and laborious. The operation of the pooling layer has many benefits. Firstly, it can retain important information more specifically while ensuring the invariance of important features. The second is to reduce the dimensionality of the extracted data features, eliminate overly redundant information, and ultimately simplify the model and improve processing efficiency[19].

3.4 Fully connected layer

Fully Connected Layers (FC) generally appear at the end of the network, and its main function is to label features. Process the extracted feature information and classify all features based on the learned information. The main operation adopted is the multiplication of matrix vectors, which weights the obtained feature information and ultimately obtains the classification result[20].

4. Feature Extraction Network of Convolutional Neural Networks

4.1 LeNET

LeNet is a simple network with only seven layers. Composed of basic structural layers, with a simple hierarchy and fewer parameters used. This network can effectively utilize the structural information of images and is the foundation of other deep learning models[21].

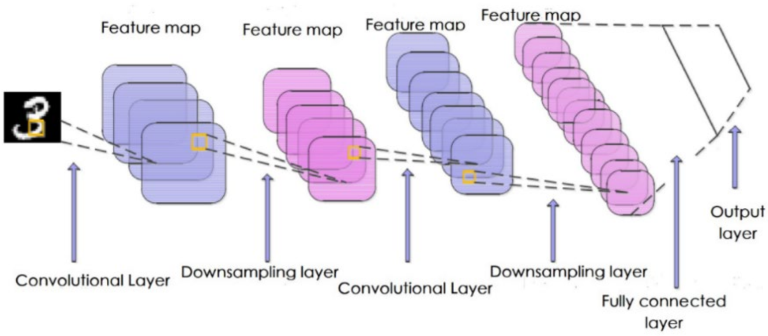

Figure 1 LeNet network

Figure 1 show the specific structure of LeNet. Input a black and white image into the network, and the convolutional layer in the network structure selects the features in the image. After passing through the pooling layer to retain the features of important areas, it finally reaches the fully connected layer. Softmax is used to determine what the target in the image is. This network has good performance in handwritten font recognition.

4.2 AlexNet

Although LeNet has achieved good results in small-scale issues such as handwritten digit recognition. But in practical situations, some datasets are even larger, and this network is not suitable, which is also the reason why the network has not been widely applied. The emergence of AlexNet has proven the feasibility of convolutional neural networks in large-scale data processing, and proposed the use of GPU to improve computing power, becoming an important driving force for the development of convolutional neural networks[22].

The structure of Alexnet is shown in Figure 2. It has 8 layers, mainly composed of convolution layer and full connection layer. From the figure, we can see their corresponding layers. ReLU is used as the activation function, which improves the training speed.

ISSN 2616-5775 Vol. 6, Issue 7: 131-137, DOI: 10.25236/AJCIS.2023.060717

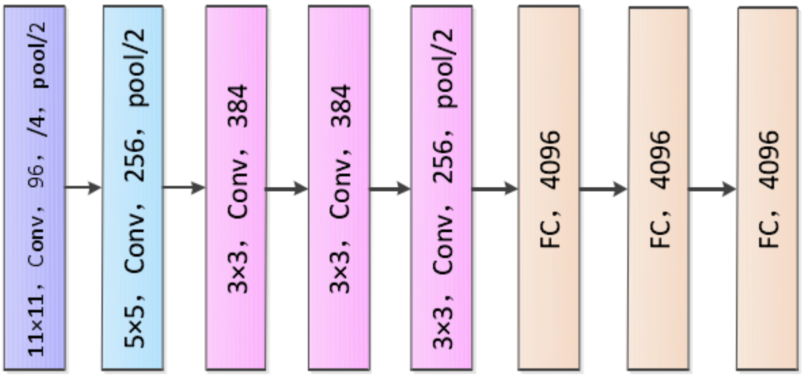

Figure 2 Alexnet network

4.3 ResNet

In 2015, He Kaiming and others proposed the ResNet network and achieved outstanding performance in world-class competitions. According to research findings, for shallow networks, the detection performance of the model can be improved by continuously stacking layers. By increasing the complexity of the model and thereby enhancing its expressive power, it exhibits better performance. But if the number of layers in the model increases to a certain extent, the performance of the model will not only not increase but also decrease. In response to this phenomenon, ResNet explores a better model structure by introducing residual blocks to improve the problem of rapid decline in network performance[23]. Due to the advantages of ResNet's own structure, it is very convenient to improve and adjust. By changing the number of channels and modules, the width and depth of the network structure can be improved, resulting in a network structure with different extraction capabilities without worrying about rapid performance degradation. As long as the network learns enough training data, it can achieve better performance by deepening the network. The specific structure is shown in Figure 3.

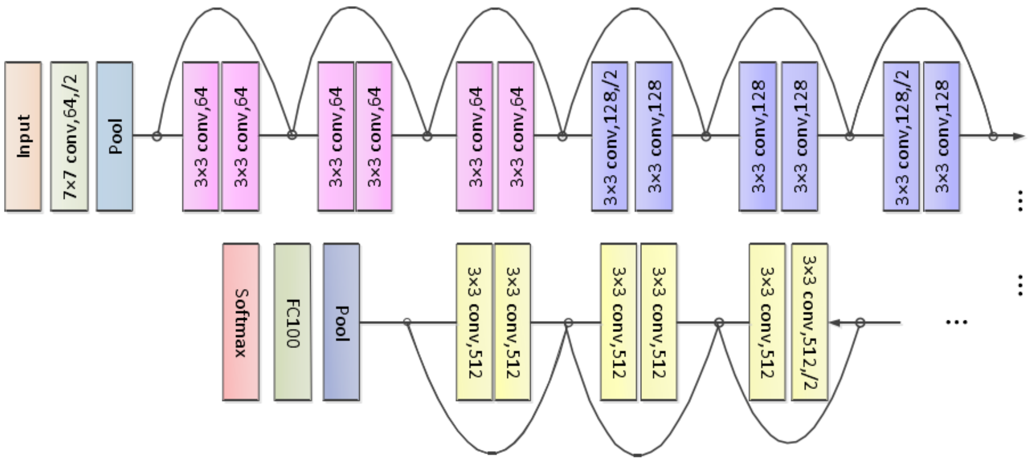

Figure 3 ResNet network

5. Conclusion

Currently, deep learning has been well applied in the field of computer vision, mainly through several core technologies, including Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), and Stacked Autoencoders (SAE). Compared to other technologies, CNN has a wider application in computer vision, mainly due to its obvious advantages, The performance is good in terms of accuracy and efficiency.At present, algorithms using convolutional networks for object detection have achieved hot development and enviable results. Compared with traditional object detection algorithms, deep learning algorithms have their own unique advantages, and their applicability is more universal. The obtained features can be transferred without the need to transform many algorithms.The cost of early development, later optimization, and maintenance is not high, and high-precision algorithms can be obtained, becoming a mainstream in the field of object detection. However, the current methods of object detection using

ISSN 2616-5775 Vol. 6, Issue 7: 131-137, DOI: 10.25236/AJCIS.2023.060717

convolutional neural networks still face many challenges in remote sensing images. Specifically, there are two aspects to this. Firstly, when encountering targets that are similar to the background in remote sensing image detection, the detection effect is poor.Secondly, remote sensing images have a large number of small and medium-sized targets, and the information available is limited. Therefore, there are higher requirements for the resolution and image quality of the targets in the image. If feature information cannot be accurately extracted, it can lead to serious cases of missed and false detections.

References

[1] Yang Y & Newsam S(2021).Bag-of-visual-words and spatial extensions for land-use classification. Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems.ACM, 2021:270-279.
[2] Lazebnik S & Schmid C(2021).Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. IEEE Computer Society Conference on Computer Visi Vaduva C et al (2019). Latent Dirichlet allocation for spatial analysis of satellite images. IEEE Transactions on Geoscience and Remote sensing, 51(5): 2770-2786.
[3] Joao Carreira, Fuxin Li & Cristian Sminchisescu.Object Recognition by Sequential Figure-Ground Ranking, In IEEE conference on computer vision and pattern recognition (Vol. 2, pp. 2169 –2178).
[4] Szegedy C et al(2021).Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition.1-9.
[5] He K et al(2021).Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.770-778.
[6] Penatti O A B et al(2021). Do deep features generalize from everyday objects to remote sensing and aerial scenesdomains?Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 44-51.
[7] Khan A et al(2022).A Survey of the Recent Architectures of Deep Convolutional Neural Networks. Computer Vision and Pattern Recognition, 2022.
[8] LecunY et al(2019).Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1(4):541-551.
[9] Deng, L. and Yu, D. (2013) .Deep Learning: Methods and Applications. Foundations and Trends? in Signal Processing, 7, 197-387.
[10] Lecun Y et al (2019). Deep learning. nature,521(7553):436-444.
[11] Scherer D et al(2021).Evaluation of pooling operations in convolutional architectures for object recognition. international conference on artificial neural networks,92-101.
[12] Gu J et al(2021). Recent advances in convolutional neural networks. Pattern Recognition, 354-377.
[13] Xu B et al(2022). Empirical Evaluation of Rectified Activations in Convolutional Network. Computer Science,2022.
[14] Loffe S & Szegedy C(2021).Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift. international conference on machine learning,448-456.
[15] Lin M et al (2019). Network In Network. arXiv: Neural and Evolutionary Computing, 2019.
[16] Dahl G E et al(2021).Improving deep neural networks for LVCSR using rectified linear units and dropout. international conference on acoustics, speech, and signal processing, 8609-8613.
[17] Dauphin Y N et al(2020).Equilibrated adaptive learning rates for non-convex optimization. neural information processing systems, 1504-1512.
[18] Redmon J et al (2019).You Only Look Once:Unified,Real-Time Object Detection. IEEE, 2016.
[19] Zoph B & Le Q V(2020). Neural Architecture Search with Reinforcement Learning. arXiv: Learning, 216-287.
[20] Zoph B et al (2019).Learning Transferable Architectures for Scalable Image Recognition. computer vision and pattern recognition,8697-8710.
[21] Zoph B & Neumann M(2019).Progressive Neural Architecture Search. european conference on computer vision, 19-35.
[22] JunhuaFang et al (2020).Distributed and parallel processing for real-time and dynamic spatio-temporal graph. World Wide Web: Internet and Web Information Systems,23(9).234-256.
[23] Hao Wang et al(2020). A diffusion algorithm based on P systems for continuous global optimization. Journal of Computational Science,44-65.