Sonar image recognition based on fine-tuned convolutional neural network

Zhaotong Zhu*, Youfeng Hu

Xi'an Precision Machinery Research Institute, CSIC, 710075, Xi'an, China

Abstract. To solve the problem of sonar image recognition, a sonar image recognition method based on fine-tuned Convolutional Neural Network (CNN) is proposed in this paper. With the development of deep learning, CNN shows impressive performance in image recognition. However, massive data is needed to train a CNN from beginning. Through fine-tuning pre-trained CNN can help us training CNN from relatively high starting points, based on those pre-trained CNNs, only few data is needed to retrain a CNN which focus on sonar image recognition. A scaled model experiment shows that based on the architecture of AlexNet, compared with the traditional learning method, the transfer learning method can achieve higher recognition accurate rate of 95.81% and less training time. Moreover, this paper also compared 6 pre-trained networks, among those networks, VGG16 can achieve the highest recognition rate of 99.48%.

1 Introduction

Sonar imaging technology, including real aperture side-scan imaging, real aperture multi-beam imaging, synthetic aperture imaging, reversed synthetic aperture imaging etc., can achieve high resolution two-dimensional sonar images in the range on hundreds meters. With the gradual maturity of sonar imaging technology, how to process a large number of sonar imaging data quickly and efficiently has become an urgent problem to be solved. To solve this problem, Underwater Automatic Target Recognition (UATR) has become a hot topic in the field of underwater target recognition.

The conventional sonar imaging recognition method includes two categories [1]. The first category is template matching and the other is feature matching. The template matching method needs to maintain a big template library, which contains most of possible imaging conditions. This kind of large-scale template library leads to inconvenient in deployment of actual system. For the feature matching method, considering the randomness of features selection and the deviation in feature extraction, many proposed feature matching method can only get ideally result in specific data and the performance tends to fall when experiment condition changed [2].

With the development of artificial intelligence in big-data era, deep learning has achieved great success in pattern recognition fields, such as image classification, target detection, natural language processing etc. Commonly used deep learning algorithms include deep belief network (DBN), convolutional neural network (CNN), recurrent neural network (RNN) etc. Among

them, CNN is the most widely used algorithm in computer vision. Similar to SAS image recognition, the conventional optical image classification problem also can be divided in two steps of feature extraction and feature classification. Usually, the feature extraction algorithm is hand-designed based on statistics or physical property. However, in the latest computer vision research, this model is completely replaced by CNN which rich hierarchical feature is learned automatically [3, 4]. CNN has made a tremendous progress in optical image classification. In 2012, ref. [3] using deep CNN in Imagenet Large Scale Visual Recognition Challenge (ILSVRC) achieved a top-5 error rate down to 15.3%, much better than at the best performance ever. In 2014, ref. [5] proposed a 22-layer-deep GoogleNet and the ILSVRC top-5 error rate is further reduced to 6.67%. In ILSVRC, the performance is improved every year by using various forms of CNN, the latest result is made by [6] in 2016, which the top-5 error rate is reached 2.25%.

Although there are some differences between sonar image and optical image, we can still learn from the methods which have good performance in the field of optical image processing. With the great success of CNN in the field of optical image recognition, this kind of method is also introduced into the field of sonar image processing. Ref. [7] uses CNN to recognize mine targets in sonar images.

Transfer learning is a method that the knowledge learned from the source domain can be transferred to the target domain. Transfer learning is been used in radar target recognition [8] and some other fields like medical image recognition [9] to solve the problem of lack of training data. Usually, the source domain knowledge is

* Corresponding author: Johntonzhu@163.com

learned from a large dataset and the target domain knowledge is based on a relative smaller dataset focus on specific tasks. By taking the pre-trained network in computer vision as the basis, the architecture and most of the parameters in the network is inherited from the original network, only the last few layers should be modified to realize specific domain recognition. This method, also known as fine-tuning, cannot only reduce the computational complexity in network training procedure, but also reduces the size of the data required for network training.

Aiming at the potential application of deep neural network in sonar image recognition, this paper firstly describes the development of deep neural network in the field of optical image recognition, and summarizes the application of deep neural network in sonar image recognition. Section 2 introduces the basic algorithm of CNN and the method of training specific neural networks by using transfer learning; section 3 introduces the experimental design and imaging results and the results of sonar image recognition by using the proposed method; section 4 compares the performance differences between traditional training methods and training methods based on transfer learning, then the performance differences of the networks based on six pre-trained networks are compared; the preliminary conclusions are given in section 5.

2 Deep convolutional neural network and transfer learning

2.1 Introduction to deep convolutional neural network

CNN is a kind of deep neural network with special architecture. Usually, the first few layers are constructed by convolution layer and pooling layer alternately, and then follow with several fully connected layers. The basic principle is that the convolution layer learns different features, and the pooling layer converges the spatial shape into the high-dimensional feature space. The multi-layer alternating convolution and pooling can learn the hierarchical feature representation. Finally, a classifier is learned by the fully connected layers in the multi-dimensional feature space.

As shown in Fig.1, in CNN, all nodes in the convolution layers and pooling layers are arranged into a series of 2 dimensional matrices, which we also called "Feature Maps". In the convolution layer, the input of each hidden layer node contains only one node in the local neighbourhood of the previous layer. The node in the local neighbourhood of the previous layer is multiplied by a weight matrix, the result is applied with a nonlinear activation function (usually a rectified linear units, ReLU [10]). The operation results are used as the output values of the nodes in the convolution layer. Each hidden layer node can be regarded as a feature detector because it has a large response value when some feature it represents appears in its input. All nodes on the same feature map are limited to share the same weights, so each feature map detects the same features at different

locations in the image. Because of local connection and weight sharing, the number of independent parameters that need to be learned from the data in CNN is greatly reduced. In the next pooling layer, each pooling layer feature map correspond to a convolution layer feature map. Each node in the pooling layer takes the node of the local neighbourhood of the adjacent previous convolution layer as input, then the down sampling is carried out. The usual method is to retain the maximum value of all nodes in one local neighbourhood, while ignoring the remaining node values. A deep convolution network contains many combinations of convolution layer and pooling layer. Finally, the multi-classes classifier is realized via fully connected layers. Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

Fig. 1. Typical architecture of a convolutional neural network.

The great success of CNN in the industry benefits from: (a) the improvement of the algorithm; (b) the acquisition of massive data; (c) the popularization of high-performance computing resources such as graphics processing unit (GPU). The improvement of the algorithm is the key factor for the rapid development of deep neural network.

2.2 Sonar image recognition based on transfer learning

Usually, a CNN contains millions of parameters, and in the field of sonar image recognition, because of the difficulty of experiment, usually the total number of data available for training is only a few hundred to thousands.

It is doubtful to use such a small number of training data to train large numbers of parameters.

As shown in Fig.2, the idea of transfer learning using a pre-trained network is based on the idea that CNN can be considered as a universal image feature extractor and can be pre-trained on a data set (such as ImageNet) and then reused on a target data set (such as sonar image experimental data). The steps of sonar image recognition using transfer learning include: 1. select a pre-trained network, which is usually trained using a large number of labeled data sets; 2. transfer the parameters of some layers in the pre-trained network to the target network; 3. Remove the last fully connected layer from the original network, add the fully connected layer according to the target classification task to the last part of the transferred network to form a new network; 4. train the new network with labelled data in the target domain to complete the transfer of the entire network.

Fig. 2. Procedure of transfer learning.

3 Experiment designing and imaging results

3.1 Experiment designing

In order to verify the method proposed in this paper, a water tank experiment is designed. The experiment aims at realizing classification for four kinds of underwater targets. There are four strong scattering structures on all four kinds of targets. Consider the scale of real underwater targets is too large to test in water tank, for example, the scale of a real submarine is about 100m, and an experiment based on scaled-model principle [11] is designed. The maximum length of the actual

underwater target is reduced by 100 times to about 1m. Thus, in order to keep the relative relationship between the wavelength and the target size, the carrier frequency of the transducer is raised 100 times to 2 MHz from the 20 kHz commonly used in imaging sonar, and the band width is also raised to 75 kHz. Thus, the resolution of the experimental sonar imaging system can reach $1\text{cm} \times 1\text{cm}$ (range $\times$ azimuth, the same below), equivalent to $1\text{m} \times 1\text{m}$ resolution in the actual system. In order to ensure that an underwater target is completely covered by transducer beam, the imaging distance is set to 9m (the model scale is less than 1m) or 12m (model scale greater than 1m). At the same time, the rotation speed is set as 0.001rad/s, equivalent to 0.9m/s (imaging range of 9m) or 1.2m/s (imaging range 1.2m) at the target low-speed cruise mode, the direction is vertical to sonar line of sight vector, and the pulse repetition frequency is 5Hz. Based on the scaled model experiment, sonar images which are close to the real target can be obtained in the experimental tank.

The experimental platform is shown in Fig.3. The model is connected to a stepping motor through a metal bar, and a constant speed is maintained by controlling the motor. When each target is imaged, the rotation platform is rotated from head to rear relative to the transmit-receive transducer, with each target guaranteed to turn 180 degrees continuously. The experimental parameters are shown in Table 1, and a detailed description of the experiment is available in Ref. [12].

Fig. 3. Sketch of experiment design.

Table 1. Parameters of experiment.

Parameter	Value
Transmit signal bandwidth (kHz)	75
Carrier frequency (MHz)	2
Sample frequency (MHz)	5
Pulse repeat frequency (Hz)	5
Total rotation angle (degree)	180
Rotation velocity (rad/s)	0.001
Total sampling time (s)	3142
Range from transducer to target (m)	9 or 12

3.2. Imaging results

For every target, all the echoes from head to rear are pulse compressed through matched filtering, and then, sliding-window process is used. For single image, the sampling pulse number is set at 256. For every imaging result, the rotation speed is estimated by the position of highlights [13], and the estimation result is 0.001rad/s, consistent with the set rotation speed. Then, the Convolution Back-Projection (CPB) [14] algorithm is used to image the target based on the estimated rotation speed. The scaled image is achieved, and the theoretical resolution is 0.01m in both range and azimuth. Typical imaging results of 5 kinds of targets are shown in Fig.4.

Fig. 4. Imaging results of 5 kinds of scaled models.

4 Classification results by using transfer learning

4.1 Performance comparison between transfer learning and traditional training method

In order to compare the difference of training speed and final recognition accuracy between the traditional training method that start from the initial state and the training method based on transfer learning, a comparative experiment is designed based on AlexNet pre-trained network. Firstly, a CNN is designed with reference to AlexNet network structure, in which the output number of the last full connection layer of AlexNet is adjusted from 1000 to 5, and all parameters in the network are initialized with random values. Then the sonar image data is used to train the network.

At the same time, based on the transfer learning method proposed in Section 2.2, the weights of AlexNet pre-trained network are fixed, only the last full connection layer is replaced, and then the training is performed with sonar image data. The experimental results are shown in Fig. 5 which indicates that the transfer learning method can reach higher recognition accuracy with less training samples.

Fig. 5. Performance comparison of transfer learning and traditional training method.

4.2 Performance comparison of different pre-trained networks

In order to compare the performance of different pre-trained networks in recognizing sonar images obtained in this experiment, this paper selects several common pre-

trained networks for comparison. The selected pre-trained networks include AlexNet, GoogLeNet, vgg16, vgg19, RESNET 50, RESNET 101. Based on these pre-trained networks, transfer learning is carried out. The results are shown in Table 2.

Table 2. Performance of different pre-trained networks.

Name of pre-trained network	Recognition accuracy rate (%)	Training time (s)
AlexNet	95.81	235
GoogLeNet	94.76	611
VGG16	99.48	1334
VGG19	97.91	1314
Resnet-50	94.24	1394
Resnet-101	90.58	3407

5 Conclusions

This paper introduces an attempt to apply the new revolutionary technology CNN and transfer learning in the field of sonar image recognition. In this paper, a new method for sonar image recognition is proposed by using CNN and transfer learning technology. The experimental data obtained by scaled model with rotation platform imaging method are used to verify the effectiveness of the proposed method. Experiments show that the transfer learning method can achieve higher recognition rate and faster convergence speed with the less training data requirement than the traditional method which initializes all the values of the network randomly. By using transfer learning, the accuracy of final recognition is improved by 23.56% compared with the traditional method. In addition, the difference between training time and recognition accuracy is compared when different pre-trained networks are used for migration learning. The results show that the highest recognition accuracy can be obtained by using VGG16 as pre-trained network, and the recognition accuracy reaches 99.48% of the training time is 1334 seconds. The shortest training time is AlexNet which can reach 235 seconds, while the recognition accuracy of AlexNet is 95.81%. In addition, the recognition accuracy of all the selected pre-trained networks is over 90%. This shows that the method based on CNN pre-trained network and transfer learning can effectively realize target recognition in sonar image recognition.

The work was supported by post-doctor foundation of Shaanxi province (2017BSHQYXMZZ04) and post-doctoral foundation of Xi'an Precision Machinery Research Institute, CSIC.

References

1. Z. Zhu, S. Peng, J. Xu, X. Xu, Journal of Electronics and Information Technology. **37**, 1757 (2015)
2. Z. Zhu, S. Peng, J. Xu, X. Xu, Acta Acustica. **43**, 154 (2018)

3. A. Krizhevsky, I. Sutskever, G. E. Hinton, *International Conference on Neural Information Processing Systems* (APNNA, Doha, 2012).
4. K. He, X. Zhang, S. Ren, J. Sun, *2015 IEEE International Conference on Computer Vision (ICCV)* (IEEE, Santiago, 2015)
5. C. Szegedy, L. Wei, J. Yangqing, P. Sermanet, S. Reed, D. Anguelov, et al., *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Boston, 2015)
6. J. Hu, L. Shen, G. Sun. ArXiv e-prints 1709. (2017)
7. J. McKay, I. Gerg, V. Monga, R. G. Raj, *OCEANS 2017 - Anchorage* (IEEE, Anchorage, 2017)
8. F. Gao, T. Huang, J. Wang, J. Sun, A. Hussain, E. Yang, *Applied Sciences*, **7**, 447 (2017)
9. Z. Huang, Z. Pan, B. Lei, *Remote Sensing*, **9**, 907 (2017)
10. X. Glorot, A. Bordes, Y. Bengio, *International Conference on Artificial Intelligence and Statistics*, (JMLR, Fort Lauderdale, 2011)
11. J. Lu, Y. Zhang, Z. Xi, M. Zhang, *Technical Acoustics*, **34**, 121 (2015)
12. H. Yan, S. Peng, J. Xu, X. Xu, Z. Zhu, X. Zhang, Chinese patent, CN104297756A (2015)
13. C. Yeh, J. Xu, Y. Peng. *IET Radar Sonar and Navigation*, **4**, 528 (2010)
14. J. Xu, X. Jiang, Y. Zhu, B. Yuan, *Acta Acustica*, **26**, 400 (2001)