Menfish Classification Based on Inception_V3 Convolutional Neural Network

Jiangchang Huang*, Weijie Gong and Hong Chen

Institute of Intelligent Technology, Shenzhen University, Guangdong 518060, PR China

*Corresponding author e-mail: huangjianchang517@gmail.com.

Abstract. Due to the impact of global environmental pollution deterioration, as well as uncontrolled fishing activities. The global fish stocks are becoming more and more serious, so it is necessary to pay attention to this issue. In this paper, the characteristics of traditional fish identification methods are vast, subjective factors are strong, and the current accuracy is not good in today's deep learning. A migration learning based on Inception_v3 network with excellent feature extraction ability is proposed. Fish recognition model This model has better recognition ability than traditional models and some simple convolutional neural networks, and further improves the accuracy of Inception_v3 for fish recognition.

1. Introduction

Recently, the video of "The Shout of the Ocean" released by the French NGO Sea Guardian Association in conjunction with TBWA has made human beings face up to the fact that we have always been proud of the marine resources with rich treasures. The total area of the ocean reaches 360 million square kilometers, accounting for about 71% of the Earth's surface area; the total volume of more than 1.35 billion cubic kilometers of water, accounting for about 97% of the total water on the earth. Therefore, the dark underwater world that people can't see should contain a wealth of resources. But is it really true? The consumption of marine fish alone has greatly satisfied the appetite of human beings. At the same time, the deeper level of marine pollution and fish consumption have reached excessive destruction. According to research by the Marine Guardian Association, it is possible for humans to have no marine fish at all in 2048. But this is not just an impact on our food and clothing problems. Without marine fish, the ocean, the main climate regulator on Earth, will stop working. Humans will not be able to survive the death of the ocean.

Since the beginning of the 20th century, people have had the habit of using cameras to monitor the fishing activities of fishing boats. Therefore, it is necessary to use machine vision to automatically classify image data. Scholars have made outstanding contributions in the field of fish identification. DU Weidong [1] et al. proposed a SVM-based decision fusion fish identification method with an accuracy rate of over 90%. LARSEN et al. used linear discriminant analysis to classify fish species shape and texture features with an accuracy of 76%. Wan Peng [2] et al. used a 3-layer back-propagation neural network to identify carp and carp, with an accuracy rate of 92.50%. However, these target recognition methods are all artificially extracted by researchers, and there are insufficient calculations, long time, poor effect, and poor portability. With the development of the field of deep
learning, the excellent performance of convolutional neural networks has made the contribution to scientific research work more and more prominent. Deep learning uses a deep neural network to simulate the learning process of the human brain, and achieves more accurate feature abstraction through layer-by-layer autonomous learning of input features by multiple layers of neurons, thereby realizing autonomous extraction features. The convolutional network works by means of weight sharing, which greatly reduces the number of parameters, so that the convergence rate is significantly faster than the fully connected back propagation network when the same recognition rate is achieved during training.

Therefore, based on the above analysis, this paper builds a more generalized model through the existing inception_v3 [3] convolutional neural network model, improves the recognition ability of fish for complex underwater scenes, and realizes many types of recognition and high accuracy. Target recognition model.

2. convolutional neural network

The two characteristics of convolutional neural networks[4] are: first, having a convolutional layer to extract features; second, the convolutional layer in the network works by weight sharing, greatly reducing the amount of parameters to be trained, making The speed is superior to the fully connected back propagation neural network at accuracy.

The convolutional neural network is composed of an input layer, a convolutional layer, a maximum pooling layer, a fully connected layer, and an output layer. The input layer is used to receive the image data to be input to the network, thanks to the special structure of the local weight sharing of the convolutional neural network, which reduces the complexity of the model, so that the data entering the input layer does not need to be pre-processed like other algorithms. Processing and feature extraction. Each convolutional layer in the model has its own convolutional kernel, which convolves the image from the input layer. During the convolutional operation, the convolutional kernel moves the specified step from left to right and top to bottom each time to perform convolutional calculation until all pixels of the image are calculated. This completes the feature extraction and compression of the image. The underlying convolutional layer extracts some of the simplest features such as straight lines and diagonal lines; the convolutional layer in the middle layer is capable of extracting deeper features such as radian and corners; the top-level volume is trained as the layer propagation of the neural network. The laminate already has the ability to extract more complex features such as colors, textures, and the like. There will also be a pooling layer between each convolutional layer. The role of the pooling layer is to extract the features of the image that has just undergone the convolutional extraction feature, which not only reduces the processing capacity of the next layer of data; The feature is also abstracted again, allowing the model to have greater tolerance for data with minor changes (small translation, scaling, and rotation), which improves the model's generalization performance and prevents overfitting. After the model learns through layers, the local features extracted in the convolutional layer are reassembled into a complete graph through the weight matrix in the fully connected layer. The final output layer is mapped to the probability space by the soft max classifier to derive the final classification result.

2.1. Model building

After absorbing the theory of migration learning in deep learning, it is meaningful to select one of the existing neural networks with good image feature extraction ability to improve on the basis of this research goal. This study uses the inception framework in GoogLeNet. Although the mainstream VGG model also has good performance, the amount of calculation is very large. The inception parameter count is only about 25 million. Therefore, the inception_v3 framework with better performance and better computation is chosen. On the basis of inheriting all the weights of inception_v3, the full connection layer is added, and the model is fine-tuned to achieve the goal of this research.

The improved convolutional neural network is mainly composed of three Inception module groups, a fully connected layer, a soft max layer, six convolutional layers, and two pooling layers. In the
convolutional layer in the non-Inception module group in front of the network, 3x3 small convolutional kernels are used. Three convolutional layers with 3x3 convolutional kernels are equivalent in series to a 7x7 convolutional layer; however, the parameters are only 55% of the parameter amount of the latter and 2 times more ReLu [5] activation functions, making the network nonlinear enhanced. Thereby enhancing the learning ability of the network. Then, after the largest pooling layer with a size of 3x3 and a stride of 2, the image size is reduced and the output channel is increased to achieve size compression and feature abstraction of the image. Then you enter the core of the entire network. Three consecutive Inception module groups, each with its own multiple Inception Modules. These Inception modules have similar mechanisms, but there are a few differences in detail. Inception Module draws on the idea of Network In Network. Each Module is equivalent to a small network in the whole neural network. This small network is mainly formed by repeated concatenation of small convolutional layers, and also has the structure of branch network, and finally outputs.Aggregate on this dimension of the number of channels. Then the graph passes the 8x8 pooling layer to the fully connected layer and finally passes the softmax layer to give the image the probability of the classification to the classification target. The overall network process is shown in Figure 1.

![Figure 1. Overall network process.](image)

### 2.2. Inception Modules
The Inception Module in Inception_v3 has been optimized with two points. First, the idea of Factorization into small convolutionals was introduced to split a large n x n convolutional into two
smaller one-dimensional convolutionals, 1 x n and n x 1 convolutional. Not only reduces the parameters in the network, but also enhances the nonlinear of the model. Second, the module has three different structures: 35x35, 17x17, and 8x8. Take the first Inception Module in the first Inception module group as an example. There are 4 branches in this module, as shown in Figure 2. After the image is processed in each branch, it is merged on the output channel. The remaining two Modules in the current module group are similar in structure, except that the parameters are different. The second and third module groups are shown in Figures 3 and 4, respectively. The parameter settings of the three Inception modules are shown in Tables 1, 2, and 3.

Figure 2. Branches structure.

Figure 3. The second module group.
Figure 4. The third module group.

Table 1. The 1st module group.

| 1st Module | 2nd 3rd Module |
|------------|----------------|
| Branch_0   | 1x1x64         | 1x1x64         |
| Branch_1   | 1x1x48 5x5x64  | 1x1x48 5x5x64  |
| Branch_2   | 1x1x64 3x3x96 3x3x96 | 1x1x64 3x3x96 3x3x96 |
| Branch_3   | 3x3 Pooling | 1x1x32         | 3x3 Pooling | 1x1x32         |

Table 2. The 2nd module group.

| 1st Module | 2nd-5th Module |
|------------|----------------|
| Branch_0   | 3x3x384 | 1x1x192 |
| Branch_1   | 1x1x6 3x3x9 3x3x9 | 1x1x128 | 1x1x128 |
| Branch_2   | 3x3x9 3x3x9 3x3x9 | 7x1x128(160) | 1x7x128(160) | 7x1x128(160) | 1x7x128(160) |
| Branch_3   | 3x3 Pooling | 3x3 Pooling | 1x1x192 |

Table 3. The 3rd module group.

| 1st Module | 2nd 3rd Module |
|------------|----------------|
| Branch_0   | 1x1x192         | 3x3x320         | 1x1x320         |
| Branch_1   | 1x1x192 1x1x192 7x1x192 3x3x192 | 1x1x384 | 1x1x384 | 3x1x384 |
| Branch_2   | 3x3 Pooling | 1x1x448 3x3x384 3x1x384 | 1x3x384 |
| Branch_3   | 3x3 Pooling | 1x1x192 |

2.3. Pooling layer, full connection layer construction

In addition to the pooling layer in the Inception module group, there are two pooling layers in the model, which are located at the front and the tail of the network. The first pooling layer is the largest pooling layer with a pooled core size of 3x3 and a stride of 2. The ending pooling layer is a global
average pooling layer with a pooled core size of 8x8. The number of neurons in the fully connected layer is 1000, and the number of neurons in the soft max layer is 24. The Soft max layer classifies the results and returns the probability of each category to derive the classification results.

2.4. Cross entropy loss function
The loss function used in this network is the cross entropy [6] loss function. Although the goal of this study is to identify and classify fish in 24, it is actually a single classification problem, that is, there is only one category in one sample (one picture). Therefore, the standard loss function of the single classification problem, cross entropy, is chosen. Its expression is:

\[
loss = -\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij} \log(\hat{y}_{ij})
\]  

Where \( m \) represents the number of samples; \( n \) represents the number of categories; \( y_{ij} \) the true label value, using the one-hot encoding; \( \hat{y}_{ij} \) is the predicted value derived from the neural network.

3. Method of reducing over fitting
For large-scale neural networks, due to its numerous parameters and complex structure, it is the main reason for the model over-fitting. Therefore, measures must be taken to reduce the over-fitting of the model, so that the model has good generalization ability. In this study, the dropout method and the structure of the Inception Module were taken to solve the problem.

The Dropout [7] method means that during the transmission of the neural network, the value of the neurons in the network is 0 with a certain probability that it does not participate in forward propagation and parameter update. Because such random discarding neurons reduce the interaction between neurons, they do not rely on the interaction of implicit nodes with fixed relationships. At the same time, dropout different hidden neurons is equivalent to training different networks, randomly discarding a part of neurons to make the network structure change, the whole process is equivalent to averaging many different neural networks, reducing the overall Fitting. Enhance the generalization ability of the model and reduce the training time of the model.

The structure can not only stack repeatedly to form a large network, but also can increase the branch network to combine information between the output channels, thereby greatly improving the utilization efficiency of parameters. A large number of 1x1 convolutional layers also play a crucial role, adding a layer of feature transformation and improving nonlinearity with a small amount of computation; while some larger convolutional kernels are also split into two One-dimensional convolutional, which also enables a large reduction in the amount of parameters. It is this sparse structure that expands the depth and width of the network, so that the accuracy can be improved while reducing the amount of calculation to reduce the over-fitting effect.

4. Experimental data, test method design
The data set for this experiment was obtained from crawler technology on the web and has a total of 24 marine life. After screening the captured images, there are 7999 images for each bio-picture. The number of each fish is balanced. The amount is more than 10 times the difference between those with less. There are 554 effective pictures of anemone fish, but the number of effective pictures of puffer fish is only 55. In the case of uneven distribution of such samples, it is not feasible to directly train these data; it will lead to over-fitting due to poor generalization performance of the model on the test set. There are two main types of methods for resolving sample imbalances. The first is from the data level, and the second is from the classifier level. In this paper, the first method is adopted, and the data is enhanced from the data layer to achieve balance between various samples. The data enhancement method adopted in this experiment uses the following four categories: flip transform, color jitter, noise
perturbation, and random cropping. Random factors are added to each of the above transformations. Since the probability that each random number generated is equal to zero, each time an enhancement is made, a different result is obtained. In this experiment, each transformation is performed 5 times, then for a picture, a total of 20 enhanced pictures can be obtained. For fish pictures with less effective data, it will increase by 20 times. Solved the problem of uneven quantity between fish. The picture after data enhancement is shown in Figure 5. A picture of 24 species of fish is shown in Figure 6.

5. Conclusion
The experiment was done under the Ubuntu 16.04 system environment, using the Tensorflow [8] deep learning open source framework. Finally, the average recognition accuracy was about 97.99%, and the recognition rate of the above fish reached 96.43%. The model has better generalization performance. Based on the deep learning, this paper builds a network model suitable for fish identification based on the existing Inception v3 convolutional neural network. In the case of a small number of data sets, migration learning achieves higher accuracy. And solve the problem of sample imbalance by using data enhancement methods. The final model has a high recognition accuracy and achieved the expected results.

![Figure 5. Data argumentation.](image1)

![Figure 6. The fish species.](image2)
Acknowledgments
I am very appreciated the Prof Chenghong and Prof Weijie Gong’ help and guidance.

References
[1] Weidong Du, et al. "SVM-based decision fusion fish identification method." Journal of Harbin Engineering University 05 (2015): 623-627.
[2] Peng Wan, et al. "Design of online identification device for freshwater fish species based on machine vision technology." FOOD&MACHINERY28.6 (2012): 164-167.
[3] Szegedy, Christian, et al. "Rethinking the Inception Architecture for Computer Vision." (2015).
[4] Cai, Y., et al. "Multi-modal vertebrae recognition using Transformed Deep convolutional Network." Computerized Medical Imaging & Graphics51 (2016): 11-19.
[5] Li, Yuanzhi, and Y. Yuan. "Convergence Analysis of Two-layer Neural Networks with ReLU Activation." (2017).
[6] Boer, Pieter Tjerk De, et al. "A Tutorial on the Cross-Entropy Method." Annals of Operations Research 134.1 (2005): 19-67.
[7] Krizhevsky, Alex, I. Sutskever, and G. Hinton. "ImageNet Classification with Deep convolutional Neural Networks." NIPS Curran Associates Inc. 2012.
[8] "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." (2016).