CLASSIFICATION OF WHALE CALLS BASED ON TRANSFER LEARNING AND CONVOLUTIONAL NEURAL NETWORK

Hao Yue[a], Dezhi Wang[a*], Lilun Zhang[a], Yanqun Wu[a], and Changchun Bao[a]

[a]Academy of Marine Science and Engineering, National University of Defense Technology, Changsha, China

[*]Dezhi Wang, National University of Defense Technology, Changsha, China, 410073
wang_dezhi@hotmail.com

Abstract: *As a primary communication method, whale vocal calls contain valuable information and abundant characteristics that are important for recognition and classification. However, subject to large variations of call types, there is still a great challenge to accurately categorize the different whale species or subpopulations. In this study, an effective method of transfer learning based on a data-driven machine-learning approach i.e. Convolutional Neural Network (CNN) is developed to extract the significant features of whale calls and classify the different whale categories from a large open-source acoustic dataset recorded by audio sensors carried by whales .The results show that the proposed method can achieve 97.04% and 91.47% in accuracy respectively to categorize the calls into the two whale species and the four whale subpopulations. The phylogeny graph is also produced to illustrate the similarities between the whale subpopulations. Moreover, all the results are carefully compared with those obtained by using the Wndchrm scheme and the Fisher discriminant scores on the same dataset.*

Keywords: *Convolutional Neural Network (CNN)，transfer learning, classification of whale calls, similarity analysis*

1. INTRODUCTION

As the common species of whales in the ocean, Killer whales (Orcinus orca) and pilot whales (Globicephala) have been continually studied for over three decades. The two types of whales are both gregarious living within socially stable family units known as 'pods'. Within a pod, whales share a unique repertoire (also known as dialect) of stereotyped calls, which are comprised of a complex pattern of pulsed and tonal elements [1].

With the continuous development of devices such as hydrophones deployed from ships, or digital acoustic recording tags (DTAGs) placed on marine mammals, large datasets of whale sound sample are acquired increasingly. In 2014, Lior Shamir etc. [2] proposed an automatic method for analyzing large acoustic datasets from the Whale FM project [3] and studied the differences between sounds of different subpopulations of whales. In their study, the significant features of whale calls were extracted by using Wndchrm [4] for biological image analysis and Fisher Discriminant Scores algorithm. These features are used to classify or evaluate the similarity between the different populations of samples without expert-based auditing. Though this work has already made a progress in the unsupervised classification and similarity analysis of large acoustic datasets of whale calls, it still highly relies on the effectiveness of different polynomial decomposition techniques and the Fisher scores algorithm.

Nowadays, as a class of highly non-linear machine learning models, Convolutional Neural Networks (CNNs) become very popular after achieving state-of-the-art results for image recognition [5], for example, the Google Inception-v3 model [6]. The purpose of this study is to apply pre-trained Inception-v3 for transfer learning and efficiently extracting the informative features from large datasets of whale calls for classification and clustering. The approach is carefully described and the whale phylogeny is also produced. All the results are also compared with those presented in the work [2].

2. MATERIALS

The datasets are obtained from the Whale FM website [3] which is a citizen science project from Zooniverse and Scientific American. All the datasets were collected by the recording DTAG [7], which can be attached to individual whales to record the sounds the whale makes as well as calls from other animals nearby. It also has motion sensors that allow following the movement of the whale underwater. The dataset consists of about 10,000 MP3 audio files ranging between 1s to 8s in 16 separate recording events based on the sensors carried by 7 pilot whales and 9 killer whales close to the coasts of Norway, Iceland, and the Bahamas. The other description about the data is shown in [2].

3. APPROACH

Transfer learning is a machine learning method which utilizes a pre-trained neural network without training models from beginning. In this study, we use Inception-v3 model trained by Google to learn the spectrums converted from the audios of whale calls and achieve the purpose of classification or identification. It contains two main parts that are feature extraction part using CNN networks and classification part using fully-connected networks. A more detailed overview of Inception-v3 can be referred to [6]. The Inception-v3 used in our work could be downloaded from Google [8]. In addition, our implementation is carried out based on Tensorflow (Version 1.1.0 on CPU), which is an open-source software library for Machine Learning using data flow graphs. Besides classification, the similarity analysis is also carried out in terms of the phylogeny produced by the distances matrix among the abstract features of different pods.

The original audio files are firstly converted to 2D spectrograms shown as Fig. 1 for the subsequent analysis. In detail, the 2D spectrograms of the audio files were created using the STFT function in MATLAB (Hanning Window = 256; Overlap = 128; FFT size = 1024; FS = 20000). And all the images are RGB style with the resolution of 256*256*3. ('3' refer to RGB channels). The spectrograms are varied even if the calls are emitted by the same whale since a single whale could send a variety of sounds, which also increases the difficulties of classification. As a kind of deep learning method, CNN is more appropriate to process large datasets. It is necessary to prepare sufficient data to fully stimulate the CNN network, for which we increased the data amount artificially just by adjusting the image contrast with the function 'imadjust' in MATLAB.

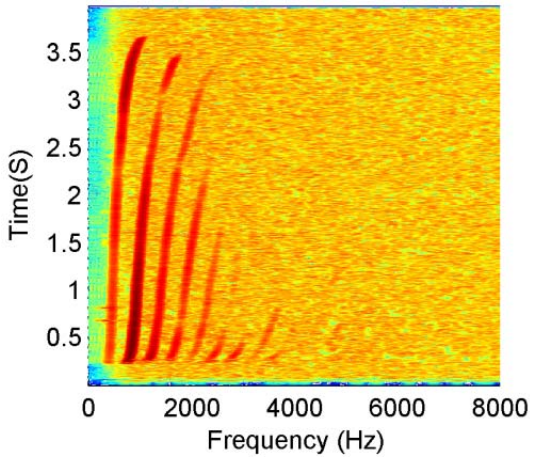

Fig. 1: Spectrogram of a sample of whale calls

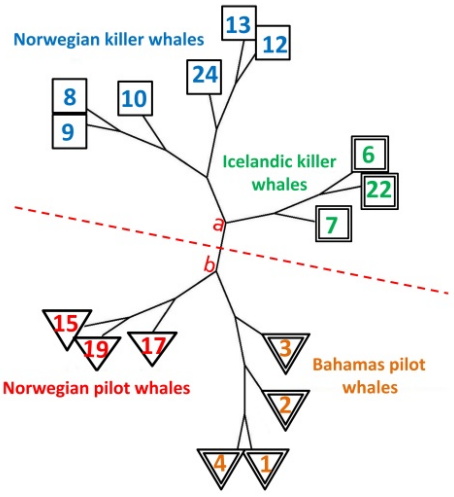

Fig. 2: The phylogeny created based on the CNN feature vectors

4. IMPLEMENTATION

4.1 Experiment 1: Classification into the pilot and killer whales

In order to classify the whale calls into the two species as the pilot and killer whales, all the MP3 audios are divided into two groups. One group consists of 5112 samples while another one consists of 4188 samples. Randomly, we choose 10% samples for validation, and 30% for testing, the rest samples are all used to retrain model. During the training process, the learning rate is set to 0.01 and the batch is set to 100. Finally, the training process will finish when the step achieves 10000.

4.2 Experiment 2: Classification into four subpopulations

This experiment aimed at separating dataset into four subpopulations by different species and habitats, they are, Norwegian killer whales, Iceland killer whales, Bahamas short-finned whales and Norwegian long-finned whales. Since the number (about 1000) of MP3 files of some pods is significantly smaller than that of the others (about 3000), we expand their amounts artificially by adjusting the image contrast of the 2D spectrograms in Matlab. After the adjustment, the number of spectrograms of each subpopulation reaches 3000. Thus totally we have 12000 samples for the four subpopulations. Similarly, 10% samples are selected randomly for validation, 60% samples for training and the rest 30% for testing. The CNN structure is as the same as one used in the last section.

4.3 Experiment 3: Classification into 16 different pods

Similarly, the image contrast of the spectrograms of each pod is adjusted so that the amounts of samples in each pod could reach 800. Thus the total number of samples of the 16 pods is 12800. Then 10% samples are selected for validation, 60% for training and the rest 25% for testing.

4.4 Experiment 4: Similarity analysis and the phylogeny

The classifier of Inception-v3 used in this study is a normal full-connected neural network, which is a classifier based on the Softmax function. When we have 16 classes (pods), a probability vector with the size of 1×16 would be the output for each sample. Each element (within a range [0, 1]) in the probability vector can be considered as a measure for the likelihood of the sample classified into the corresponding class or, in other words, the similarity between the sample and the class. Thus the probability vectors are reasonable to be used to evaluate the similarities of the whale calls in different pods. Totally 4, 300 samples are selected randomly from each pod to achieve the corresponding probability vectors. The row elements of all the probability vectors of each pod are summed up to achieve a 1×16 general probability vector, which is employed to indicate the similarities between one pod and all the other pods. Then, we regard the probability vectors as the abstract feature of corresponding classes, and calculate their Square Euclidean distance each other to describe their similarity each other.

In Lior Shamir's study [2], in order to visualize the relations between the different subpopulations of whales, the phylogeny inferred by the Phylip software [9] is obtained based on the Minkowski distance between the selected polynomial decomposition descriptors of different pods, as shown in [2]. In our experiment, the phylogeny is also obtained by using Phylip but on the basis of the Square Euclidean distance of the extracted general probability vectors. In detail, we mainly use Fitch-Margoliash method and n-Body algorithm to create and improve the outtree respectively, the other parameters are defaults.

5. RESULTS

5.1 Classification

As shown in Table 1, the classification accuracies corresponding to the two, four and 16 classes are listed and compared. It is worth noting that the reason why the 73.5% accuracy reached when classifying samples into 16 classes is that the samples from different pods might be the same whale subpopulations with very similar features. It results in a great difficulty to perfectly classify the samples. In this case, the similarity analysis or clustering analysis might be a more appropriate way to study the differences between the pods. About computational efficiency, we just use single Intel® Xeon® E7-4830 V2@2.20HZ CPU to retrain our dataset in our work. Averagely, it takes 20～40 min to deal with the dataset, and about 0.1s per step for training. So, it just needs about 15min when finishing the 10000 iterations.

Methods	Wndchrm	CNN
Input data	2D Spectrograms	2D Spectrograms
Feature extraction	Polynomial decomposition methods and Fisher Scores algorithm	Inception-v3
Classification to two species	92%	97.04%
Classification to four subpopulations	×	91.47%
Classification to 16 pods	44%～62%	73.50%
Similarity	Phylogeny shown in [2]	Phylogeny (Fig. 4)

Table 1: Comparison between Lior Shamir's study and our work

5.2 Similarity analysis

As shown in Fig. 2, the phylogeny graph can be clearly divided into pilot whales and killer whales along the middle dashed line. Also the influence of the geographic locations on the whale subpopulations is completely distinguished by our method as there are four distinct branches displayed. Specifically, there are two obvious points

which clearly indicate the bifurcation of different subpopulations. For example, the bifurcation point ‘a’ is highlighted for the killer whales in different regions, from where the left branches are Norwegian killer whales and the right are Icelandic killer whales. Compared with the phylogeny in [2], our result is apparently more distinct and informative. The detailed comparison is shown in Table 1.

6. CONCLUSION

The main purpose of this study is to use the Convolution Neural Network (CNN) to analyze large acoustic datasets and study differences between sounds of different whale subpopulations. The results show that the transfer model Inception-v3 is able to accurately classify these datasets into different categories in a supervised fashion. As shown in Table 1, the accuracy is obviously better than that in [2]. In general, in the case of sufficient dataset, the effect of transfer learning is not as good as complete re-training, but the data and time required for transfer learning are far less than complete re-training. Therefore, it is favorable to do the underwater target recognition with insufficient dataset. Moreover, transfer learning can present great efficiency although there are no high performance GPU devices for general users.

The other purpose of this study is to characterize the similarities between the whale calls of different subpopulations in an unsupervised fashion. When applying the CNN to the datasets, we do not need to tag these datasets to certain classes but only randomly label each dataset with an ID. In this way, the informative features of the datasets can be extracted by the CNN algorithm for the similarity analysis. In [2], only 15% to 20% polynomial decomposition descriptors are selected by Fisher scores criterion to do the similarity analysis. In contrast, the features obtained by CNN more concisely represent the original data. The phylogeny graph is also produced based on the abstract CNN features in order to achieve a better understanding of the relations between whale subpopulations, which is also more distinct than the one shown in [2]. Thus CNN algorithm is shown to be an effective solution for the classification and similarity analysis of large acoustic datasets.

7. ACKNOWLEDGEMENTS

This study was funded by the National Key Research and Development Project of China (No. 2016YFC1401800) and the Scientific Research Project of NUDT (No.ZK16-03-46, No.ZK16-03-31).

REFERENCES

[1]. Miller, P.J.O. and D.E. Bain, Within-pod variation in the sound production of a pod of killer whales, Orcinus orca. Animal Behaviour, 2000. 60(5): p. 617-628.

[2]. Shamir, L., et al, Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls. Journal of the Acoustical Society

of America, 2014. 135(2): p. 953-62.
[3]. See the project at https://whale.fm and obtain the supplementary whale dataset used in the study at: https://github.com/zooniverse/WhaleFM/blob/master/csv/whale_fm_anon_04-03-2015_assets.csv
[4]. Shamir, L., et al, Knee x-ray image analysis method for automated detection of osteoarthritis. Biomedical Engineering IEEE Transactions on, 2009. 56(2): p. 407.
[5]. Krizhevsky, A., I. Sutskever, and G.E. Hinton, ImageNet classification with deep convolutional neural networks. in International Conference on Neural Information Processing Systems. 2012.
[6]. Szegedy C, Vanhoucke V, Ioffe S, et al, Rethinking the Inception Architecture for Computer Vision[J]. Computer Science, 2015:2818-2826.
[7]. Johnson, M., et al, A digital acoustic recording tag for measuring the response of marine mammals to sound. Journal of the Acoustical Society of America, 2003. 108(5): p. 2582-2583.
[8]. Download the Inception-v3 at: https://storage.googleapis.com/download.tensorflow.org/models/inception_dec_2015.zip
[9]. Felsenstein, J, PHYLIP: phylogeny inference package. The Quarterly Review of Biology, 1989. 5(Volume 64, Number 4): p. 164-166.