UNIVETSITY OF SOUTHAMPTON

Machine learning research on humpback whale audio recognition based on noise reduction

Minzhou Wang, Student Number: 34087559

Supervisor: Professor Paul White

Msc Advanced Mechanical Engineering science

in Computational Engineering Design

Faculty of Engineering and Physical Science

september 2023

MSc AMES in Computational Engineering Design

September 2023

Declaration of Authorship

I, Minzhou Wang declare that this thesis and the work presented in it are my own and has been

generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a degree at this University;

2. Where any part of this thesis has previously been submitted for any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. None of this work has been published before submission.

Abstract

Audio data sets from passive ocean acoustic monitoring can help researchers gain insights into cetacean vocal patterns and behavior. We need an automated way to classify recordings based on the sounds present. Humpback whale vocalizations are uniquely complex among cetaceans, with a wide range of articulation frequencies that would be challenging for this approach. But using convolutional neural networks for this classification has shown promise, and the representation of input features is a very important step. We can take advantage of different time-frequency analysis methods, and then use multiple residual convolutional neural networks for comparison, Audio enhancement techniques can also be built to process the sound data, and the model can then be used to perform binary classification of sound recordings from Perth Canyon. Research on short-time Fourier transform (STFT), constant Q transform (CQT) by training and testing them using different time-frequency representations to explore methods of these representations. In the overall comparison, choose the slightly better STFT method and then use resnet18 and resnet50 for performance testing. In further research, audio enhancement technology is used on the originally selected model and time-frequency analysis method. Even if the number of training sets is not increased, its accuracy and performance will be greatly improved based on the previous ones. Two audio enhancement technologies, PCEN and median normalization both improve and significantly improve the recognition of unseen data, and PCEN's improvement performance is better.

Acknowledgements

I have to thank my family, my friends, for their endless support. Of course, the person I should most thank is my supervisor Paul. My speaking and expression skills were not very good at the beginning, but he has been very patient and humorous in promoting my learning process. I feel very happy to accept his supervision and listen to his explanations in speaking. I am also very grateful for this entire project, because the knowledge and skills I learned from this process are of great value and are worth remembering throughout my life.

Contents

List of figure

List of table

1. Introduction

As we all know, whales are one of the most well-known mammals on the planet, with a huge range of species and a large oceanic spread. Whales, dolphins, and porpoises are only a few of the 94 species of cetaceans. Although we frequently refer to whales, dolphins, and freshwater dolphins separately in everyday speech, in zoology they all belong to the same family. No whiskers, just teeth, typically conical, in the mouths of toothed whales. They primarily eat fish and squid,And they have only one nostril is present. From a 1-meter-long pygmy dolphin to an 18-meter-long sperm whale, the body size varies substantially. Typically, males are larger than females. Toothed whales come in a variety of varieties. Sperm whales and pygmy sperm whales, white whale, and small headed whale are among the approximately 75 species that now exist and are grouped into 9 families. River dolphins, dolphins, porpoises, narwhals, puffers (freshwater dolphins), beaked whales (sword-beaked whales), subspermidae, and one species of baiji.

The suborder of toothed whales, which includes sperm whales, killer whales, and other species, is called Delphinidae; the suborder of baleen whales has baleen in place of teeth. Baleen whales come in fewer varieties; there are 15 species total, which are further split into right whales and new baleen whales (or tiny whales). Baleen whales, grey whales, and right whales. Humpback whales and the blue whale, the biggest animal on Earth, are baleen whales, which filter plankton from the ocean to feed themselves. Despite the fact that whales resemble fish, they are actually a species of mammal. They primarily utilise their throats to produce noises, and by contracting their chest and throat, they can alter the sound's amplitude and frequency.There are many teeth in the toothed whale's mouth, which can be located by vibrating the lips to generate sounds. The low-frequency portion of the sound is primarily centred in the 20-200Hz region. Between 30 Hz and 2000 Hz, conspecifics communicate at low frequencies.

Despite the fact that whales resemble fish, they are actually a species of mammal. They primarily utilise their throats to produce voices, and by contracting their chest and throat, they can alter the sound's amplitude and frequency.There are many teeth in the toothed whale's mouth, which can be located by vibrating the lips to generate sounds. The low-frequency portion of the sound is primarily centred in the 20-200Hz region. Between 30 Hz and 2000 Hz, conspecifics communicate at low frequencies.

The Megaptera novaeangliae family of baleen whales, which includes the humpback whale, is known for its distinctive and sophisticated vocalisations. Songs and calls are all that they make. Only male humpback whales can write the songs, while both sexes can make the calls. The songs are made up of call-like sound kinds but are organised into patterns. Humpback whale cries, which are layered periodic sounds made up of intricate layered repeating patterns that distinguish humpback whales from other species, contain numerous complicated and diverse signals. The sound has a unique quality.

Humpback whale songs are composed only by male, with calls produced by both sexes. The songs are composed of sound types similar to calls, but arranged into patterns. It can allow us to study whale classification while understanding humpback whale habits, and in addition, the data collected from PAM has some challenges of its own. Because the background of the collected sounds is the real ocean, they may have a variety of sound sources, which can be broadly categorised as biosound (biological), anthropogenic (man-made), and natural sound (climatic and geological, etc.). Collected data on humpback whale vocalisations are often prone to low signal-to-noise ratios(In the case of low signal-to-noise ratio, the impact of noise is greater and may interfere with or obscure the useful information of the signal.), which may be due to sound attenuation and scattering, which is also related to the dive depth of the hydrophone, in addition to interference from the sources mentioned above, and even electrical noise from the recording equipment, in addition to the complex ocean environment. So successful classification of sound often requires robust classification methods as well as effective noise cancellation techniques so that sound classification can be performed successfully in the presence of significant background noise.

We typically learn data in which sounds are turned into visuals using time-frequency conversion techniques before supplying input to the machine learning model. The spectrogram offers a time-frequency representation of the signal that differs from the initial time domain representation as it transforms from time domain to frequency domain representation. Frequency domain data may be more useful or simpler to read than time domain data for some purposes, such as audio processing or communications. This improves model performance and makes it easier to grasp the programme. Selecting various transformation techniques also allows us to preprocess audio files into the frequency domain and process noisy sound data. Along with translating the data into the frequency domain, preprocessing interference needs to be kept to a minimum. The cause is that the performance (accuracy) of the model may be impacted by the raw audio data's potential for high noise and irrelevant sound information. Of course, the machine learning algorithm model is the most crucial component for processing data. Different neural networks can be used in specific contexts.

2. Literature review

In order to succinctly discuss all the important components of this dissertation, I will summarise the pertinent literature in this section. I start by outlining machine learning applications. The use in noise interference reduction study thus draws specific emphasis to the necessity of translating the signal from the original time domain representation to the time-frequency domain representation[1]. There is discussion of the convolutional neural network model. Two efficient audio processing technologies are based on this. The definitions of data sources and sub-projects are presented, along with an explanation of their significance. Additionally, some information pertinent to this subject is given, along with its sources.

2.1 Role of Convolutional Neural Networks in Classification

Machine learning has demonstrated amazing computational speed, an ability that gives them an advantage in many complex and varied scenarios, and it has also demonstrated its potential in classification tasks. While traditional classification methods often rely on hand-designed features and fixed rules, automatic detection and classification algorithms for whale calls have been studied quite extensively because machine learning algorithms are able to automatically learn features and rules from large amounts of data.

However, with the very large datasets and increased computational power available today, deep learning has become widely available, showing success especially in classification tasks. Convolutional Neural Networks (CNN) are one of the most commonly used type of deep learning. There are many significant advantage when i use CNN: for example i can automatically identify relevant features by CNNS without any human supervision. CNNs are widely used in image recognition. The traditional convolutional neural network its is composed of three sets of layered structures: a convolutional layer, a rectified linear unit layer and a pooling layer. This structure typically needs to be repeated two to four times. The convolutional layer detects the input whale call data by using a filter or convolutional kernel and extracts features in a hierarchical fashion.ReLU (Rectified Linear Unit) is an activation function commonly used in deep learning and neural networks. Its functional form is very simple and is defined as follows: when x is greater than 0, return the value of x; when x is less than 0, return the value of 0. ReLu is not required, but the introduction of ReLU will allow the neural network to learn and approximate complex non-linear linear function. Without a nonlinear activation function, no matter how many layers there are, the neural network is always a linear transformation of the input data, which greatly limits its representation ability. Although there is more than one activation function ReLU, using ReLU to learn and identify local features in images and represent more complex functions and decision boundaries is a very popular way in academic area. Finally, through dimensionality reduction, dimensionality reduction aims to reduce The dimensionality of the data, while trying to retain the important information or structure in the original data, and the pooling layer is a good tool for reducing the spatial size of the data. Backpropagation can use the chain rule to calculate the gradient of each layer to calculate the gradient of the loss function with respect to the network weights to update the grid weights. The combination described above allows the model to be adapted to complex and multi-dimensional sound data.

Machine learning neural networks that are often used for sound recognition are ResNet (Residual Network) It is a very popular convolutional neural network structure for deep learning, and is often seen especially in image classification and object detection tasks.

A ResNet is composed of a multicomponent layer structure, which includes a convolutional layer, a nonlinear activation function (e.g., ReLU), and special "jump" connections or "residual" connections. Although pooling layers are present in many deep network architectures, in ResNet, especially in deep ResNet, the convolution with a stride of 2 is usually used to reduce the size, instead of using pooling layers(2). Meanwhile, the main innovation of ResNet is residual connections. These connections allow information to "skip" one or more layers directly from a lower layer to a higher layer. This approach effectively solves the problem of gradient vanishing in deep neural network training. In summary, the main contribution of ResNet is to avoid the problem of gradient vanishing in deep networks by introducing "residual blocks", so that deeper network structures can be trained.

2.2 Time-frequency representation and spectrograms

Even though CNNS models have proven to be very powerful, machine learning is still a method that uses data to drive it. The correctness and quality of the data is critical for the model to be able to learn and make accurate predictions about unseen data. It becomes especially critical to find a robust way to make the data passed to the machine learning model more accurate. We have the option of converting to a frequency domain representation, and much of the literature suggests that digital time domain audio signals recorded from PAMs often need to undergo some processing before they can be analysed and imported. Generally the acquired direct data is represented in the time domain, and the time domain features aim at extracting recognisable information from the original radiated signal of the target. There are various time-domain based features, such as the energy, position and amplitude of spectral peaks in the time domain. On the basis of these features above, a series of new features can be obtained through statistical methods, such as the maximum, minimum, mean, and standard deviation of the relevant features to pave the way for identification. However, time-domain representations may not be suitable for capturing certain features in audio because time-domain representation of audio is a direct representation of the waveforms that change over time. This is because many audio signals are non-stationary, meaning their statistical properties change over time. In the time domain, this change may not be obvious, whereas in a time-frequency representation (such as STFT or CQT)(3), this change may be more easily observed. When waveforms of multiple frequencies are superimposed, they can be very complex and difficult to distinguish in the time domain. In the frequency domain, on the other hand, each individual frequency can be clearly represented. In the time domain, different audio components may also be mixed together, making them difficult to distinguish. Whereas in the frequency domain, each component in an audio signal has its own specific frequency, which makes differentiation and interpretation easy. In summary, this representation (time domain) does not intuitively reveal all the characteristics of sound. As for the frequency domain representation: it is the Fourier transform that establishes the link between the time and frequency domains. A complicated time series signal may become a simple and regular frequency domain distribution after Fourier transform. In the field of audio recognition, often the frequency domain better reflects the characteristics of the original signal nature. On the basis of the Fourier transform can be further power spectrum, this method is fast, simple and effective to implement, and the power spectrum characteristics also confirm its widely used in the identification of the status of the ship's radiation noise.

The frequency domain is an effective method that allows us to gain a deeper understanding of the information in the signal. Humpback whale calls, singing and other sounds have their own specific patterns in frequency. These characteristics can be observed more clearly in the frequency domain, but are less obvious in the time domain. And using time-frequency representations like spectrograms, it is possible to capture changes in both time and frequency of a signal. For non-stationary signals (such as the sound of a humpback whale), such representations provide a powerful tool to see both how the sound changes over time and how it is characterised in frequency. As a result, time-frequency representations are widely used to analyse bioacoustic signals. It is common in spectrograms to represent signal quantities as colour intensities within an image, with boundaries depicting time and frequency. Spectrograms are complex humpback whale calls transformed, however, these signals are feature representations of the main candidates for image classification through transfer learning. As the transformations contain humpback whale signals, the features are encoded as image data and machine learning is performed, and the

CNN can learn to recognise the weights of high-level features in the spectrogram. The basis of the common theory for generating spectrograms is the short-time Fourier transform. The frequency domain can also provide a range of advanced feature extraction methods such as CQT (Constant-Q Transform) These features have been shown to be very effective in sound and music analysis.

2.3 alternative representation

Knowing how to analyze sound is critical. Time domain analysis focuses on the waveform of the sound signal, while frequency domain analysis focuses on the frequency components of the signal. Time-frequency analysis combines these two dimensions and not only provides richer frequency information, but is also extremely valuable for specific applications, such as the analysis of humpback whale sounds. Short-time Fourier transform (STFT) is a commonly used time-frequency analysis method, which provides a basic framework for expressing time-frequency characteristics. However, STFT has a fixed bandwidth and the same temporal resolution, which may not be ideal for all frequencies. On the other hand, the constant Q transform (CQT) is similar to the STFT, but its characteristic is that the bandwidth is proportional to the center frequency, so the center frequency separation of CQT is geometric, while that of the STFT is linear. This means that CQT provides higher frequency resolution at low frequencies and higher time resolution at high frequencies. For a specific application, such as humpback whale sound analysis, it may be necessary to try multiple sound representation methods to determine which method best captures the key features of the sound. Overall, time-frequency analysis provides researchers with a toolset to analyze and interpret sound data from multiple perspectives.

2.4 Audio Feature Enhancement

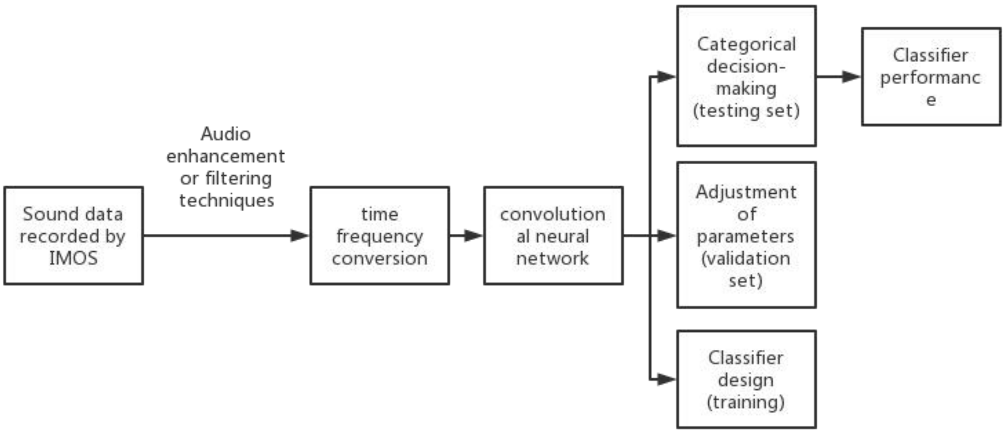
Whether in forests or in the ocean, the sound data we record often has a lot of noise, which not only interferes with the correct training of the algorithm, but also makes validation a major frustration. An elegant solution to this problem is to use feature normalization technology. Their purpose is to preprocess audio signals to enhance certain features or reduce noise and other unwanted components, where one or more features are deformed and applied to a set of annotated training samples, resulting in a lager datasets. One of the key concepts of enhanced features is that the tagged application data itself changes, but does not change the meaning of the tag. For example, taking computer vision as an example, no matter whether the rabbit is rotated, translated, mirrored or scaled, the deformed image of the rabbit is still a continuous image of the related rabbit. This leads to the fact that we can apply these techniques to deform itself in order to generate additional training data while still Keep label meanings valid. We can better generalize to unseen data by acquiring additional deformation data and continuing to train the network, hoping that the network's interpretation of the classification of these deformations remains unchanged. Many scholars in the audio field have also proposed label-preserving deformations and shown that this operation can effectively improve model accuracy in sound classification tasks. However, in the case of classifying sounds from a specific environment, the use of feature extraction and enhancement is relatively limited, so the correct combination of audio data can improve the efficiency of processing unknown audio. In the ocean, where humpback whales typically reside around 200 to 250 meters underwater, common techniques for enhancing audio signatures include spectral subtraction: this is one of the earliest noise suppression techniques and is designed to estimate and reduce the noise component of an audio signal. Median normalization: Background noise can be reduced using median normalization, where the spectrogram is normalized based on the median value of each frequency band. This normalization technique is often used, especially when the background noise is relatively constant and persistent. It is based on the idea that for a stable background noise it will remain approximately constant throughout the recording period and will typically appear as the median value in each frequency band. Therefore, subtracting the median of each frequency band from that band can effectively reduce or eliminate background noise. PCEN (Per-Channel Energy Normalization): has received a lot of attention in recent years, especially in acoustic event detection and bird song recognition. PCEN effectively suppresses background noise and highlights the main features of sound by modeling the nonlinear characteristics of auditory perception. Its original design intention was to offset non-smooth background noise and improve the signal-to-noise ratio.

2.5 PAM in Perth Canyon

The dataset we explore was collected as part of IMOS (Integrated Ocean Observing System). In particular, we focus on a subset of this dataset located in the Perth Gorge off the coast of Western Australia. The canyon facility manages multiple passive acoustic observatories, each consisting of two to four moored hydrophones and noise recorders. These noise recorders capture all sounds, including the natural sounds of the ocean and biological sounds, including those of fish and marine mammals, each of which has its own unique acoustic properties. IMOS conducted passive acoustic soundings in the Perth Canyon from 2007 to 2017, and this part of the dataset is referred to as PAPCA. The sounds of whales, especially humpback whales, are a major part of the soundscape of Perth's canyons. In addition, this site serves as a feeding ground for several species and a migration route for some species. Although there is some other noise here, such as human-made noise, the richness of the sound data from Perth Canyon makes it a more challenging dataset, especially for training and testing machine learning models.

2.6 Voice recognition system framework

The recognition of hydroacoustic signals, including humpback whale calls, used to rely on the original acoustic system to collect data for recognition, but with the increasing amount of information, this method has long been difficult to meet the current needs due to its low efficiency. The recognition process of humpback whale acoustic signals can be classified as a kind of pattern recognition, which is divided into four parts: audio signal acquisition, pre-processing, feature extraction and classifier classification decision, and its process steps are given by Figure 1.


Sound data recorded by IMOS
Audio enhancement or filtering techniques
time frequency conversion
convolution al neural network
Categorical decision-making (testing set)
Classifier performanc e
Adjustment of parameters (validation set)
Classifier design (training)

Figur1 Project flow chart

(1) Data collected by IMOS: refers to the use of different underwater signal sensors to convert signals in the water into processable digital signals.

(2) Sound enhancement or filtering techniques: These techniques can improve the signal-to-noise ratio, making sound analysis more accurate by reducing noise and enhancing target sounds (such as whale calls). Its purpose is to better divide, enhance or select the signal we need from the background. Common ones include filtering and denoising, pre-emphasis and windowing, etc.

(3) Convolutional neural network: CNN can automatically learn and extract meaningful features from the spectrogram, obtain the classification feature sequence that best highlights the signal characteristics, and train to obtain a good classification model.

(4) Classification decision: The main function of the classification decision module is to train data samples according to the mathematical proportions assigned

by the classifier, and then let the classifier learn the criteria for classification judgment by itself, and finally use this judgment rule to realize the identification of the objects to be identified. Sample function. In the above process, it is mainly necessary to allocate different proportions of samples to be used for training as the training set, another part to verify and adjust the parameters as the verification set, and the last part as the test set for testing.

3. Aims and Objectives

Once the relevant context is established, its overall goals emerge. The purpose of this project is to explore whether using different techniques to process audio signals (including time-frequency analysis methods, audio enhancement filtering techniques, convolutional neural network models, etc.) as feature inputs for machine learning models will have an impact on classifier performance. We use data from Perth Canyon for training and recognition classification. Different performance indicators should also be compared. Parameters related to performance metrics should be controlled constant to better understand how they affect the model. If performance changes are observed, we will also continue to explore the direct impact of these parameters and effectively validate these changes. These goals can help solidify the effectiveness of audio signal processing techniques and can highlight the strengths and weaknesses of these techniques for future research. The project includes the following goals:

1. Evaluate and compare the effects of different spectrogram conversion methods (such as STFT, CQT, etc.) in humpback whale sound classification.
2. Use Perth Canyon data to train and compare performance of different convolutional neural network models.
3. Explore and select the most effective audio enhancement filtering techniques (such as median normalization and PCEN) to reduce background noise and improve the signal-to-noise ratio of sound recognition.
4. Compare different performance indicators to determine which indicators are most critical to improving model performance and obtain the results for verification.

4. Method

4.1Initial Data Acquisition and Experimentation

The PAPCA dataset covers recordings from 2008 to 2016 (October). Measurements were taken at a depth of 200 meters and data were collected at a 6 kHz sampling rate (3 kHz bandwidth). The recording cycle is 15 minutes, with 5 minutes of recording and 10 minutes of rest. I used two sets of data from the papca dataset, 3444 and 3376, which were collected at different times, of which there exist three smaller subsets containing wav files and dat files, which were collected in binary format from the audio recorder Raw data. A wav file is a 20 second clip from a larger 10 minute recording. But the most important are the wav format files in 3444 and 3376; 3444 has 427 HW (humpback whale sounds present) wav files and 578 NOHW (mainly ambient noise, no humpback whale sounds) wav files, 3376 has 455 HW wav files and 410 NOHW wav files. Groups 3444 and 3376 are used to perform machine learning and then output its accuracy for comparison with later techniques for removing background noise. The main goal is to compare the performance of models trained on different spectral representations of test data. I conducted several experiments on this to compare the performance of models trained on different spectral representations. This leads to the optimal spectrum that can be used to train whale sound classification. Python libraries for creating spectrograms are matplotlib.pyplot and Librosa. Together, these two libraries are used to load WAV data, the matplotlib.pyplot library creates STFT spectrograms, and the Librosa library contains processing for creating surrogate representations of sound signal functions.

Several experiments were conducted to compare the performance of models trained on different sound spectral representations. The first experiment compares the performance of models trained on linear scale STFT and CQT spectrograms(4).

In order to better choose the time-frequency analysis method to train the model, I conducted many experiments to compare the performance of the classification model with different versions of the model trained on different method representations. I wrote a spectrum generation channel that uploads spectrograms to the model's training and test sets. The first experiment compares the performance of models trained on STFT spectra, CQT (two models in total). These were tested on episodes 3444 and 3376.

The second experiment is to select the time-frequency analysis method with better performance, and then test two residual network models and compare their performance.

The third model fixes the first two variables and then attempts to explore the impact on model performance by combining different audio signal enhancement processing techniques and comparing their performance.

4.2 Time-frequency analysis methods

4.2.1 Short-Time Fourier Transform

Prior to creating the spectrogram pipeline, various representations were explored by testing them on a sample file first provided that contained humpback whale vocalisations. This process was instructive in understanding the characteristics of the representations and constructing the pipeline. The STFT method was first tested. A function in the matplotlib.pyplot library was used to generate spectrograms from WAV files. It uses the STFT algorithm, which is expressed in the general form. where m and k are the time frame index and discrete frequency index, and R is the jump factor(5).

$$\mathrm{X[m,k]} = \sum_{n=-\infty}^{N=\infty} w[\mathrm{n} - \mathrm{mR}]\mathrm{x[n]}e^{-j2\pi kn/N} , \qquad \mathrm{k,n} = 0,1,\dots \mathrm{N}-1 \qquad (1)$$

$$\mathrm{m} = 0,1,\dots \mathrm{M}-1$$

In the specgram function of matplotlib.pyplot, STFT (Short Time Fourier Transform) first multiplies the input signal x by the specified window function, and then performs a discrete Fourier transform (DFT) on the signal within the window. In order to obtain a continuous spectrum over the entire duration of the signal, the window function is moved by a specified step size R to calculate the next frame. This approach allows STFT to be viewed as a filter bank technique. R (also called jump or stride) determines the number of samples that overlap between adjacent windows. The result of the DFT is a matrix where the rows represent frequency indices and the columns represent time frame indices. I chose a Hann window of length 512 as the window function. The Hann window is a smooth cosine window whose shape is defined by a cosine waveform from 0 to 1. It is itself a cosine window. In addition, the specgram function provides a parameter called noverlap to specify the number of samples that overlap between windows.

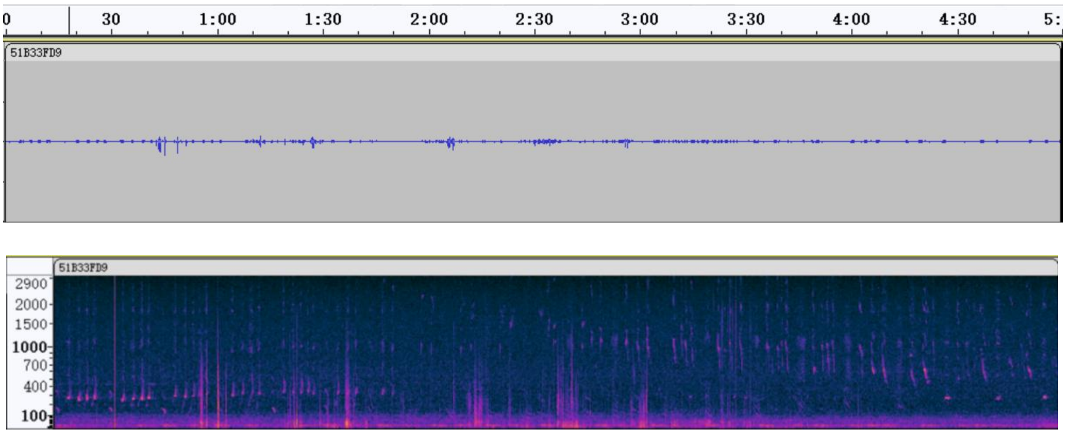


Figure 2 The specgram of audio file

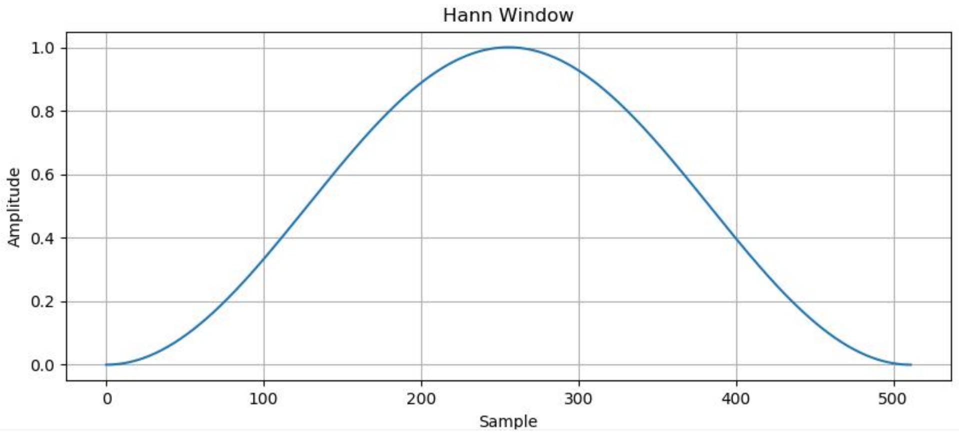


Figure 3 Functional diagram of the time-frequency image and spectrogram hann window about audio file

In my feature implementation, I converted the WAV audio file from time domain to frequency domain by STFT (Short-Time Fourier Transform), and I used the specgram method in the matplotlib library to generate the spectrogram, using an NFFT of 512 and an overlap of 500 to ensure that there is sufficient resolution on the frequency axis. And the linear frequency axis used by default is suitable for applications that need to look at the high frequency range, as the resolution of high frequencies is the same as the resolution of low frequencies on the linear axis(6). Although it is less intuitive because the human auditory system is more sensitive in the low frequency range than in the high frequency range, we convert the spectrograms for machine learning without considering the human ear, so the linear axis can be chosen because according to the literature depiction, the main subject of our project: humpback whales have a frequency distribution of around 20 to 4000 HZ, which is a high-frequency sound. I chose the linear frequency axis to convert for training purposes and also to analyse the data uniformly over the entire frequency range of the measured depth.

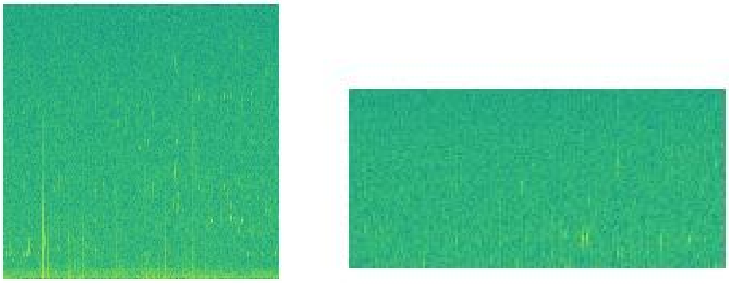


Figure 4 The image by STFT

4.2.2 Constant and variable Q transform spectrograms

Python's speech processing librosa library contains a range of functions for signal processing and Music Information Retrieval, and is used to generate spectral spectrograms for CQT.CQT is based on a geometric distribution of frequencies, where each octave contains the same number of frequency filters, so that the bandwidth of each frequency filter is a constant multiple of the previous one. constant multiple of the previous one. and the minimum centre frequency f. These are expressed as eq. . The general CQT algorithm is similar to a modified form of the STFT as shown in eq.

$$\mathrm{Q} = \mathrm{fk} / \Delta\mathrm{k} , \mathrm{fk} = f_0 2^{k/B} \tag{2}$$

$$X^{cqt}[\mathrm{m}, \mathrm{k}] = \sum_{n=-\infty}^{N=\infty} w_k[\mathrm{n} - \mathrm{mR}]\mathrm{x}[\mathrm{n}]e^{-j2\pi Qn/N[k]} \qquad \mathrm{k, n} = 0, 1, \ldots \mathrm{N} - 1 \tag{3}$$

$$\mathrm{m} = 0, 1, \ldots \mathrm{M} - 1$$

$$\mathrm{N[k]} = \mathrm{Qfs} / \mathrm{fk} \tag{4}$$

From a filter perspective, CQT can pass a signal x through a bank of filters that correspond to complex exponential terms multiplied by a window function. This is a common description of time-frequency representation in signal processing. But looking at the comparison between STFT and CQT shows that STFT's window function is fixed in size, meaning that its filters are equal in size for each frequency. The key difference with CQT is that its filter size varies with frequency, which allows it to have higher frequency resolution at low frequencies and higher time resolution at high frequencies. On computational efficiency it can indeed be seen that CQT is more computationally intensive than STFT. The traditional STFT uses a fixed-length window function whose computation makes efficient use of the Fast Fourier Transform (FFT.) The computational complexity of the STFT is roughly O(NlogN), where N is the length of the window. In contrast, CQT, especially the direct implementation of CQT, requires the computation of an FFT for each frequency filter, and each filter may have a different length. This may lead to a higher computational complexity of CQT than STFT.

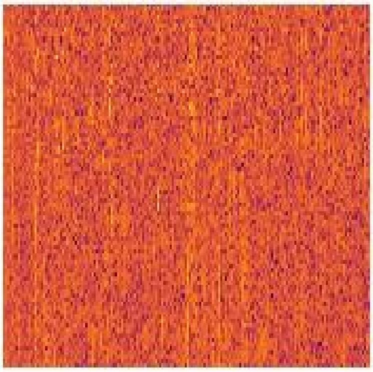


Figure 5 The image by CQT

4.3 Enhancement methods after signal processing

Here, we introduce two powerful preprocessing methods: median analysis and perceptual contrast enhancement (PCEN).

Median normalisation: short-time outliers are removed by calculating the median value in each frequency band. This method effectively removes transient noise and other disturbances, thus making the whale's acoustic signature more prominent.

Perceptual Contrast Enhancement (PCEN): This is a method based on modelling the human auditory system, and its technique of action is used to enhance the contrast of sounds. It utilises techniques that emphasise those parts of the sound that are distinct from noise and other disturbances to make the humpback whale's sound more distinctive on the spectrogram.

4.3.1 median normalization

Digital images are often corrupted by impulse noise (i.e. sensor noise) during the processing of sound data or during transmission over open communication channels. This noise can severely impact certain image processing, such as edge detection, image segmentation, data compression, and object recognition. Therefore, noise filtering (image restoration process) is an important part of many image processing systems before image processing is performed on the image. To this end, relevant researchers have proposed a variety of filters. For example, the median filter is a well-known nonlinear filter used to suppress impulse noise. Although the median filter is effective at smoothing noise, it tends to blur fine details and often destroys edges. In recent years, some improved median-based filters such as weighted median (WM) filter and detrended median (MD) filter are two methods. First, the WM filter is an extension of the median filter that assigns more weight to certain values in the filter window(7).

The weights assigned by the WM filter indicate the influence of the input samples and their importance in determining the final filter output. Therefore, if you want to have the best noise reduction capability while retaining the most signal details, the optimization of the WM filter is the most important thing. This method is very sensitive to setting appropriate weights, and this method is likely to cause filtering effects. Not good unless the weights are set correctly. For median detrending of sound files, the essence of this method is to calculate the median of each frequency point of the spectrogram, and then subtract this median from the corresponding frequency point. Therefore, MD is a median detrend. The trend approach is in the frequency domain. This processing can help remove average energy or background noise within a specific frequency range, making features in the spectrogram more obvious. It is simpler, more convenient and more reliable than WM. A comparison of the two methods shows that the expectation of our study is to reduce impulse noise: the median is a robust statistic, while MD is more robust to impulse noise or sudden outliers. Additionally, since detrending can help remove inherent background or average energy, this makes other changes or features more obvious. Since the main goal of our project is to remove the background or average energy to better observe certain features or changes. Therefore, we believe that detrending is more appropriate as a comparison (median detrending). The principle of median detrending is to remove slower trends or baseline drifts in the data while retaining faster changes. This method is often used in time series data and image processing. Its principle can be as follows: for a given data sequence X=$\{X_1,X_2,X_3,\ldots,X_n\}$

1.For each data point X_i, consider a local window of it, say of size M, such that the subset is

$$W_i=\{X_i-\frac{M-1}{2},\ldots,X_i,\ldots,X_i+\frac{M-1}{2}\} \tag{5}$$

2.computational subset $W_i$upper quartile M_i 。

3.From the original data point X_i Subtract the corresponding median M_i from，Get de-trended data $Y_i=X_i-M_i$。

where $i=\frac{M+1}{2},\ldots,N\frac{M-1}{2}$

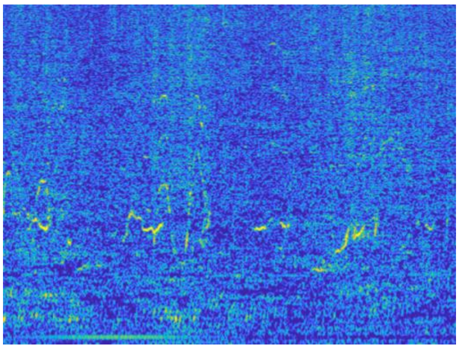


Figure 6 The image by Median normalization

4.3.2 Per-Channel Energy Normalization

In speech classification tasks, frequency shift is the main factor of intra-class variation. Frequency shift refers to the change of pitch, and in some audio classification tasks, this change is the main factor of intra-class variation. This means that different samples from the same class (e.g. calls from the same species of whale) may appear spectrally completely different because of their pitch. Includes bioacoustic species classification which tunes the auditory filter to the perceptual mel scale to provide a time-frequency representation, called the Mel-spectrogram, which takes into account the nonlinear frequency sensitivity of human hearing, shifts in pitch (or shifts in frequency transitions) appear as vertical shifts on the Mel Spectrogram.

An audio representation that takes into consideration the nonlinear frequency sensitivity of human hearing is called a Mel-spectrogram. This property enables convolutional operators in the time-frequency domain, such as convolutional neural networks and time-frequency scattering, to extract pitch contours as spectrotemporal patterns in the context of this representation and in the presence of a single source. This characteristic is known as equal variance. However, in the real world, there are frequently multiple active sources(8).

In real world recordings, however, there is often more than one active source, especially outdoors. Even if the classification task is narrowed down to only identify the most salient source (which can also be referred to as foreground sound), the presence of background noise is detrimental to classification. In fact, on the one hand, intra-class variability leads to frequency shifts in the foreground independent of the background. On the other hand, equal variance is only possible if the foreground and background are swapped at exactly the same time. In this context, in the context of automatic speech recognition and acoustic event detection, the main role is briefly to normalize the energy of each channel of the audio signal. It was originally designed to simulate the compression effect of the mammalian ear, allowing it to better simulate human auditory perception when processing audio signals.

PCEN has been shown to be particularly useful for a variety of mammalian calls, as well as for speech recognition tasks in noisy environments. We apply PCEN to datasets of various natural acoustic environments, and find that PCEN can convert a large number of real-world sound scenes into additive white Gaussian noise (AWGN), which is a computationally efficient front-end for robust detection and classification. Compared with traditional audio feature extraction methods such as Mel Spectral Coefficient (MFCC) and Short-Time Fourier Transform (STFT), logarithmic compression is used in practical applications to simulate the nonlinear frequency perception of the human ear. However, these methods alone are not effective enough to model the response of the human auditory system to long- and short-duration sound events.

So PCEN tries to solve this problem by dynamically normalizing the energy of each channel, which takes into account the past audio energy. In this way, PCEN can adaptively reduce noise, echoes, and other persistent background sounds while emphasizing transient sound events. The generation of PCEN is a technology proposed to improve the robustness to channel distortion. It combines dynamic range compression and adaptive gain control with temporal integration(9).

Mathematically, for a given channel, PCEN can be calculated by the following formula:

$$\mathrm{PCEN}(\mathrm{y}(\mathrm{t})) = \frac{\mathrm{y}(\mathrm{t})}{(\varepsilon + (\mathrm{E}\ (\mathrm{y}))^{\alpha})^{\beta}} \qquad (6)$$

where

y[t] is the signal amplitude at time t.

E[y] is the local mean of the signal y, usually calculated through a smoothing filter (such as an exponential moving average).

α is a parameter between 0 and 1 that determines how smooth the local mean is.

ε is a small constant in case the denominator is zero.

β is a constant that adjusts the degree of non-linearity of the compression.

In the formula, E[y] is the local mean of the signal y. This is to ensure that currently, when calculating the signal value at a certain moment, it can consider not only the signal value itself at this moment, but also the signal value in the period before it. In other words, the current signal value is calculated and interpreted based on its recent past value, which can be designed to make the algorithm more sensitive to short-term sound changes and less sensitive to long-term background noise or continuous sound . In noisy environments, the features of speech or other target sounds may be masked or distorted by noise. PCEN can help enhance these features, making them easier to detect and recognize by subsequent algorithms or models. The β parameter in the formula allows nonlinear compression of the signal. This can further simulate the non-linear perceptual characteristics of the human ear, especially when processing sounds of varying loudness. The ϵ in the formula is added for stability to prevent numerical problems caused by the denominator being close to or equal to 0.

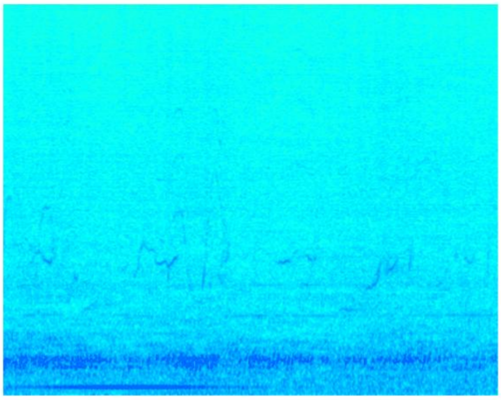


Figure 7 The image by PCEN

4.4 Deep Convolutional Neural Networks

4.4.1 Overview of Convolutional Neural Networks

In 1958, the concept of the perceptron was first proposed. This structure is the prototype of the current neural network, with two layers of neurons responsible for calculating each parameter. Perceptrons were considered to have clear advantages at the time. Because of its learnable properties, it is also considered the first learnable artificial neural network. Since then, the U.S. government has strongly funded the research of neural networks, and believes that neural networks are one of the most important scientific and technological development directions, The picture shows a single-channel convolutional neural network, and now multi-channel network neural networks are commonly used.

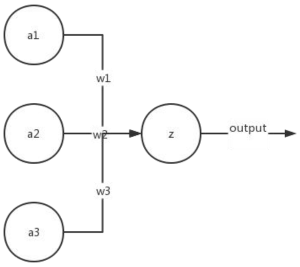


Figure 8 The monolayer neural network

The convolutional neural network is also a feedforward hierarchical neural network like the conventional neural network, The following figure gives an overview of the multilayer neural network.

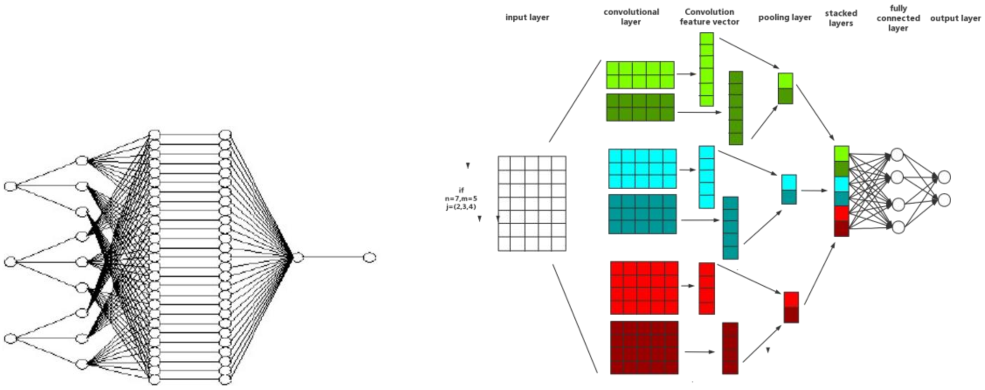


Figure 9 multilayer neural network

However, the form of the neural network layer has changed. Each convolutional network layer is composed of several convolution kernels and pooling layers(10). Generally, convolutional neural networks are sequentially composed of convolutional layers, activation layers, pooling layers, and fully connected neural networks. Layer composition, Figure 9 shows the structure of a convolutional neural network.

The convolutional neural network has the advantages of sharing weights, reducing parameter volume, and avoiding overfitting. Its properties are determined by its structure. Assume that the input data dimension is 224×224×3, and the output feature data size is guaranteed to be constant(11). The number of parameters required by the fully connected neural network layer is about 2×10^{10}. Assuming that the number of output feature layers of the convolutional layer is 256, the number of parameters required is about 7000, which is a huge difference. As shown in the figure, the size of the input feature map is After the combination of the convolutional layer and the pooling layer, it will be reduced, and the number of layers of the feature map will gradually increase, and the total number of parameters will gradually decrease, which reduces the complexity of the neural network, which is far less than that required for all fully connected neural network layers. parameter amount. The structural design and function of each part will be described in detail below.

4.4.2 Convolutional and pooling layers

The convolution kernel is the constituent element of the convolution layer. Each convolution kernel is composed of several n-dimensional matrices. When n is 1, the convolution kernel is a 1-dimensional vector. When n is 2, the convolution kernel is 2 dimensional single-channel matrix. The feature extraction methods mentioned above are all 2-dimensional, so the 2-dimensional convolution kernel is used for calculation, and the calculation method is shown in Figure 10 .

Convolution kernels are the constituent elements of the convolution layer. Each convolution kernel consists of several n-dimensional matrices. When n is 1, the convolution kernel is a one-dimensional vector. When n is 2, the convolution kernel is a two-dimensional single-channel matrix, and it is calculated in two dimensions as shown in Fig.

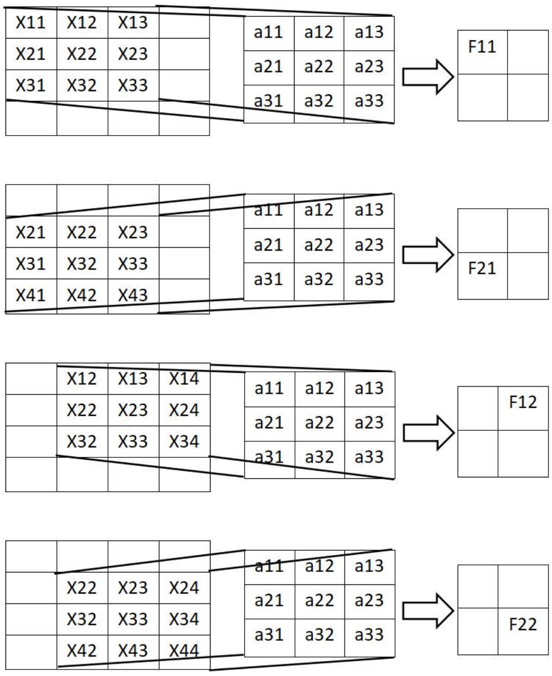


Figure 10 Operation of the single-channel convolution kernel

As can be seen from the figure, taking the case of a single-channel convolution kernel with a calculation step of 1 as an example, the size of the convolution kernel is 3×3, and the parameter value is a matrix. The single-channel convolution kernel is calculated by sliding from the upper left corner of the feature map to In the lower right corner, calculate the F_{11},F_{12},$\mathrm{F}_{21,}\mathrm{F}_{22}$ where 2×2 The matrix is the output feature calculated by the channel convolution kernel. The calculation method of the convolution kernel is(12)：

$$F_{m,n} = \sum_{i=1}^{3} \sum_{j=1}^{3} x_{m+1,n+j} a_{ij} \tag{7}$$

The convolution kernel is generally composed of multiple channels, and the number of output features is determined by the number of channels of the convolution kernel, and the two are always equal. In order to extract richer sample features, the convolution kernel is generally provided with a larger number of channels, and as the neural network layer increases, the number of convolution kernel channels gradually increases, and the size of the single-channel feature map is reduced to reduce the number of parameters. as the picture shows.

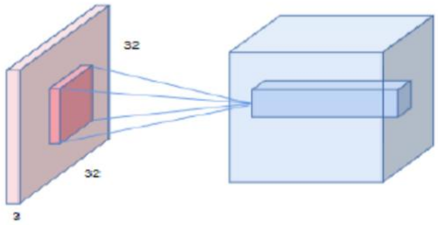
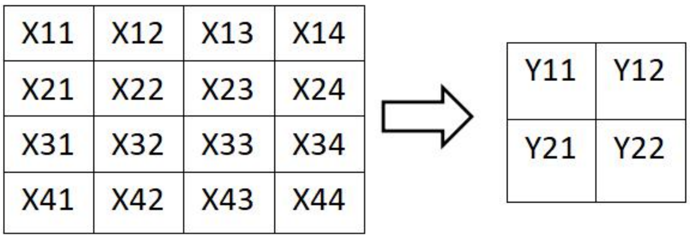


Figure 11 Operation of the multichannel convolution kernel

The reason why the convolutional neural network can share the parameter value of the convolution kernel is that for each layer of the convolutional neural network, the convolution kernel is unique, that is, the convolution kernel will completely slide through all the input feature data for calculate[(14)].

The number of channels is generally a multiple of 2 such as 32, 64, 128, etc. This setting method is conducive to parameter optimization of the underlying hardware such as GPU or CPU.

The size of the convolution kernel generally takes a multi-channel matrix of equal length and width, such as 1×1, 3×3, 5×5, 7×7. The larger the convolution kernel size, the larger the characteristic receptive field and the lower the recognition granularity . When the size of the convolution kernel is greater than 1, the length and width of the output feature will be reduced. When we do not want the size of the input feature to be inconsistent with the size of the output feature, we can pad zero elements around the input feature and increase the size of the input feature to maintain the input. Consistency of output. In order to reduce the number of parameters, in addition to adjusting the number of channels of the convolution kernel, the size of the input features can also be reduced through the pooling layer. Assuming that the pooling step size is 2, taking maximum pooling as an example, the working method of the pooling layer is shown in the figure12[(14)].



X11	X12	X13	X14
X21	X22	X23	X24
X31	X32	X33	X34
X41	X42	X43	X44

Y11	Y12
Y21	Y22

Figure 12 The neural network

Where x_{11} is max number of .$x_{21}x_{12}x_{22}$， By analogy, it can be seen that the pooling layer effectively reduces the size of the feature map, but does not modify the number of channels of the feature. Commonly used pooling methods include maximum pooling, random pooling, median pooling, and average pooling. Maximum pooling is more sensitive to the edge and texture structure of the image, and is suitable for image recognition problems. The advantage of mean pooling is that it can reduce the offset of the estimated mean, thereby improving the robustness of the model, so when dealing with underwater acoustic signals, the effect of mean pooling will be better.

It can be seen from the structure of the convolutional layer and the fully connected layer that the convolutional layer does not need to determine the size of the input feature map, but only needs to know the number of channels, while the fully connected layer requires a fixed input dimension to correctly match the parameters for calculation. The global average pooling layer can overcome this shortcoming. Its function is to average all the input values of each channel as the output result, and the dimension of the output feature must be consistent with the specified number of output channels. The result after the whale call is extracted from the spectrogram It is a two-dimensional single-channel feature map, but most deep convolutional neural networks accept three-channel data as input, and the input data needs to be converted into three channels to adapt to the structure of the neural network. The shape of the input feature can be changed through the 1×1 convolution kernel layer. As can be seen from the above, the number of channels output by the convolution neural network layer is determined by the number of channels of the convolution kernel, and the 1×1 size with a step size of 1 The convolution kernel can change the channel number of the spectrogram without changing the input feature dimension. Because the convolutional neural network expects the number of input channels to be 3, a 1×1×3 convolution kernel can be used as the first network layer. Compared with directly copying the input data to three copies as the input of the neural network, this method increases the nonlinear characteristics of the neural network.

4.4.3 Deep Residual Convolutional Networks ResNet

It is known that the deeper the neural network, the stronger the ability to extract nonlinear feature distribution and the richer the information, such as VGGNet.

The structure reaches 19 layers, and the GoogLeNet structure reaches 22 layers. However, with the gradual increase of the number of layers in the conventional neural network structure, the expressive ability of the model does not increase continuously, and the gradient will have a large attenuation when it is propagated layer by layer. , so that the problem of gradient disappearance occurs. When the number of neural network layers is about 20 layers, the phenomenon of overfitting and gradient disappearance can be alleviated through batch normalization processing of data and the Dropout method. When the number of layers exceeds the critical value, the The effect of the class method gradually weakens. The residual block proposed by ResNet is a network structure that effectively solves the problem of gradient disappearance. It makes an identity map for the input of each layer of neural network, so that the neural network learns to remove the residual part outside the identity map. The structure of this residual block makes the parameters of the neural network easier to optimize, and further deepens the neural network layer. Numerous behaviors are possible. Res Net alleviates the negative impact of the above problems through the idea of residual blocks. By directly transmitting input information to the output port of the network, a direct information transmission channel is established to protect the integrity of information. The entire residual block The network structure only needs to learn the part of the difference between input and output, which simplifies the goal and difficulty of learning. In order to allow deeper neural networks to achieve good results, He Kaiming proposed a deep convolutional neural network structure - Res Net. The idea of this neural network structure is mainly derived from VLAD (the source of the idea of residual) and Highway Network (the source of the idea of skip connection). The design method of the neural network structure breaks the limitation that the traditional neural network can only be transmitted layer by layer, and provides a new direction for the experiment of superimposing more layers of network parameters. With the deepening of the number of network layers, the learning goal of each layer of Res Net network parameters is changed to the learning process of identity mapping, which protects the integrity of information and reduces the loss of information in the transmission process, so as to maintain the accuracy of the subsequent neural network layers. There will be no decrease in accuracy. When the input and output channels of the residual block are different, the network structure adopts channel mapping, that is, the number of channels is modified through a 1×1 convolution kernel to make the dimensions of the input and output consistent.

All in all, ResNet is a popular network architecture in deep learning, especially in computer vision tasks such as image classification and object detection. It is a general term for residual networks, which have multiple layers of networks. The core idea of ResNet is to introduce so-called "residual blocks" (or "shortcut connections"), allowing activations to directly skip one or more layers. This design helps to train a deeper neural network, while reducing the problem of gradient disappearance and representation bottlenecks. The basic idea of each residual block is: if your original input is x, then after some convolution operations, i get the output F(x), then the output of the network is not just F(x), but F(x)+x(15). Thus, if F(x)=0 (or close to 0), then the output of the network approximates the original input x, which allows information to "skip" certain layers. Resnet 18 and 50 are residual convolutional neural network models with 18 and 50 layers respectively. Let it be trained on a database of image sets, and these learned weights can be used to measure and compare the performance of the neural network, thereby tuning its functions and parameters, and helping us determine appropriate methods and functions to remove background noise.

4.4.4 Optimizer of the neural network

In the process of optimizing the neural network, that is, when the error function value is minimized, after obtaining the gradient of each parameter, the algorithm that optimizes the parameters according to its numerical value is called the optimizer. The original gradient descent method is:

$$\theta' = \theta - \partial \nabla L(\theta) \tag{8}$$

θ is the original parameter of the neural network, ∂ which is the learning rate, which is a advance parameter of the neural network, namely the fixed constant value before the optimization of the neural network. Its value is generally between 0.0001 and 0.5. If the learning rate is large, it may lead to the neural network error.

The minimum value of the error function about the distribution parameters, as shown in the figure, when the value of the learning rate is small, the parameter optimization will be too slow, as shown in Figure13 (b). $\nabla L(\theta)$ is the gradient calculated by the back-propagation algorithm, and θ' is the optimized parameter value$^{(16)}$.

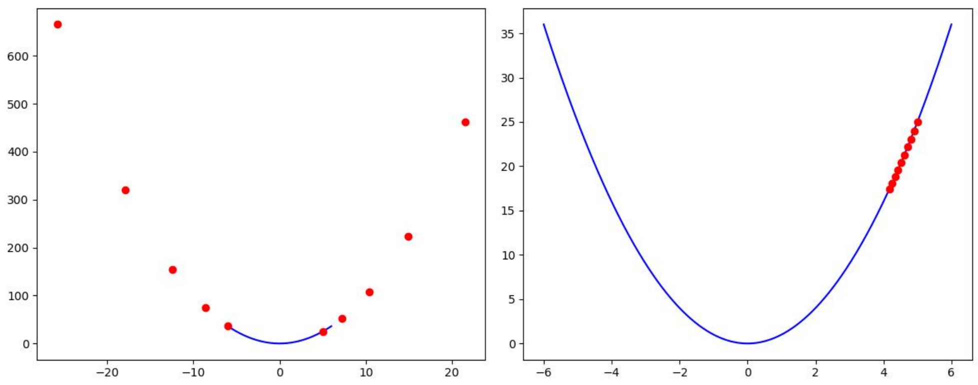


Figure 13 Learning rate is too large or too small(a.too big,b.too small)

In order to speed up the optimization of neural network, and solve the problem of saddle point, some parameter optimization algorithm, called stochastic gradient descent algorithm high neural network optimization efficiency, and set the learning rate of 0.001. When each group of sample number of 1, batch gradient algorithm is called stochastic gradient algorithm, namely each sample, the optimization of neural network parameters. This optimization method leads to the chaotic optimization direction of the parameters, the error function value will be larger, but the optimization times are more, the optimization speed of the neural network is faster.

4.4.5 Important parameters:

transforms.Resize((100, 100)): All images will be resized to 100x100 size.

Data Splitting:

train_test_split(...): Divide the data into training, validation and test sets, using a split ratio of 80%-10%-10%.

The SGD optimizer is used.

The learning rate lr is set to 0.001.

The momentum momentum is set to 0.9.

5. Results and analysis

5.1 Evaluation Principles and Standards

A.The confusion matrix is to combine the number of correctly classified samples and the number of incorrectly classified samples through a table, and present this result in this table. The purpose of the confusion matrix is to count the results of the classification model. For example, in the binary classification, category 1 is positive, category 2 is called negative example, false value (False) indicates that the classifier result is incorrectly predicted, and true value (True) indicates that the classifier result is correctly predicted. The four basic elements below the data are based on the combination of these four basic concepts（17）:

1. TP (True positive): a true example, the model predicts that it is a positive example, but it is actually a positive example.

2. FP (False positive): False positive example, the model predicts a positive example, but it is actually a negative example.

3. FN (False negative): False negative example, the model predicts a negative example, but it is actually a positive example.

4. TN (True negative): true negative example, the model predicts that it is a negative example, but it is actually a negative example.

The details are shown in the following table:

Table 1 confusion matrix diagram

Real situation	predicted results	
	Positive	negative
Positive	TP	FN
Negative	FP	TN

There are several formulas for the confusion matrix:

Accuracy

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Meaning: The correct proportion of the total predicted value in the predicted results.

B.Loss

Used to assess the difference between model predictions and actual results. When training a neural network, the loss function provides a measure of the model's current performance. The optimizer uses this loss value to update the model's weights and biases. For classification problems, a common loss function is the cross-entropy loss. Loss is a scalar value that measures how well the model predicts.

5.2 Experimental results on time-frequency

To determine model performance, we used several binary classification predictors obtained from various models. There are many indicators. For Experiment 1, we used resnet18 as the deep network to obtain the model, and then used STFT and CQT to train the training set 3444 respectively to obtain a statement of the confusion matrix of its training set for each training model. The confusion matrix shows the classifier's predicted threshold for a given decision, where the rows of the matrix correspond to the true labels of humpback whale sounds and the columns correspond to the model's predictions. The confusion matrices with a threshold of 0.5 for the two data sets are shown in Tables 2 and 3.

For both groups, the confusion matrix shows that the model performs well on classification, and most predictions are on the diagonal, indicating that the predictions are correct. It can be seen that the results of the second model generally have more mispredictions. For the first model, fewer false predictions are represented by STFT and most predictions are positive. This is not the case for CQT, where the proportion of false labels is higher for false positives than for false negatives. We assume that HW is positive and NOHW is negative.

Table2　Use resnet18 to train set1 and set2 (STFT)

Predicted label　（train test）						
Train label	Set_3444	HW	NOHW	Set_3376	HW	NOHW
	HW	40.90%	1.59%	HW	50.87%	1.73%
	NOHW	0.70%	56.82%	NOHW	0.58%	46.82%

Table3　Use resnet18 to train set1 and set2(CQT)

Predicted label　（train test）						
Train label	Set_3444	HW	NOHW	Set_3376	HW	NOHW
	HW	40.80%	1.69%	HW	51.45%	1.16%
	NOHW	0.30%	57.21%	NOHW	1.16%	46.24%

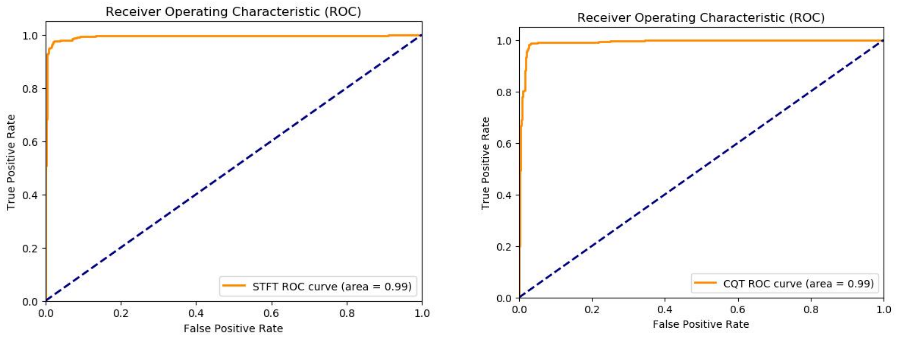


Figure 14 The ROC curve for the set_3376

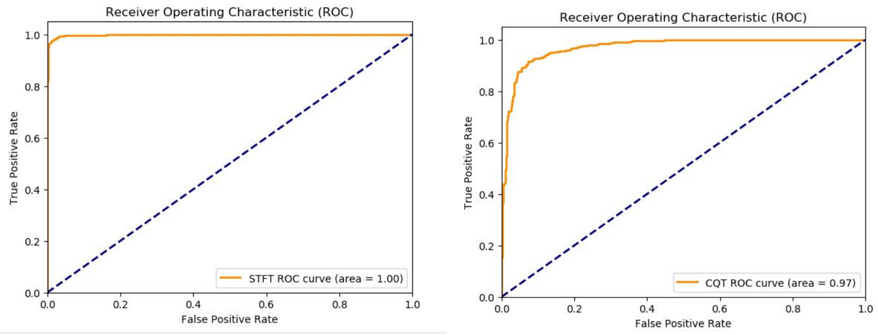


Figure 15 The ROC curve for the set_3444

In the context of the ROC curve, "area" usually refers to the area under the curve, fully known as "Area Under the Curve", or AUC for short [18].

AUC is the area under the ROC curve. The meaning of AUC for this value is as follows:

AUC = 0.5: The model is as predictive as random guessing. This means the model does not have any predictive value.

0.5 < AUC < 1: The predictive performance of the model is better than random guessing. The closer the AUC is to 1, the better the performance of the model. For example, AUC = 0.7 means that the performance of the model is moderate, while AUC = 0.9 means that the performance of the model is very good.

AUC = 1: The model's predictions are perfect on all positive and negative examples.

AUC < 0.5: The predictive performance of the model is worse than random guessing, but this usually indicates a problem with the model or data, since simply inverting the prediction results can give AUC values greater than 0.5.

In practice, AUC is widely used as a metric to evaluate the overall performance of a model without considering specific decision thresholds. This makes it a useful tool for evaluating and comparing different models, especially with imbalanced datasets. In the ROC curve, the dashed line with a slope of 1 represents the performance of a random classifier. This line is also known as the "line of indifference" or "the line of randomness". Simply put, since the humpback whale sounds I am dealing with only have two subsets, HW and NOHW, this is a binary classification problem, and when we guess the class of each sample completely randomly (regardless of any features or information), then We will have a 50% chance of guessing the positive example and a 50% chance of guessing the negative example. Therefore, my False Positive Rate (FPR) and True Positive Rate (TPR) will be equal, producing this line with a slope of 1. In an ROC curve, anything above the line indicates that the model is proving to perform better than random guessing, while anything below the line indicates that the model is performing worse than random guessing. This line is the reference point for evaluating model performance. Ideally, the ROC curve is as prominent as possible to the upper left corner, which means that the model has a high true positive rate and a low false positive rate.

As shown in Figure 14 above, this type of graph shows the thresholds of true positive rate and false positive rate for different classifications. The first experiment is to use resnet18 to train data containing two subsets of set_3376 and set_3444. From the confusion matrix, it can be seen that the two The performance of the two models trained is similar, and only a small part of the data fails to judge. Whereas the ROC curve (AUC) is the probability that a sample is labeled as the true value. When tested on subset 3376, the shape of all curves indicates that the classifier performs very well. Looking at the results of the resnet18 model, it can be found that the linear frequency STFT representation performs better, and the CQT performance is also better, but if it is worse than the former. When testing in subset 3444, the performance of CQT dropped significantly, and the performance of STFT was higher than that of CQT compared to the first time.

Table 4 The accuracy of two method be trained

Experiment	Set_3376	Set_3444	Set_3376	Set_3444
method	Accuracy（train）	Accuracy（train）	Accuracy (test)	Accuracy (test)
STFT	0.9769	0.9801	0.8736	0.9208
CQT	0.9544	0.9214	0.8517	0.8864

From the data of different experiments above, it can be concluded that for the recognition of humpback whale sounds, it can be found that the performance of STFT is better than that of CQT.

5.3 Experimental results for different residual models

Since the experimental results for time-frequency show that STFT is slightly better than CQT in this problem, so for the second experimental control variable problem, we choose fixed STFT as the method of time-frequency conversion, and then select resnet18 and resnet50 respectively. The subsets 3376 and 3444 are trained and studied, 80% of the data is randomly selected as the training set, and 20% is used as the verification set, and the accuracy and loss are obtained respectively.

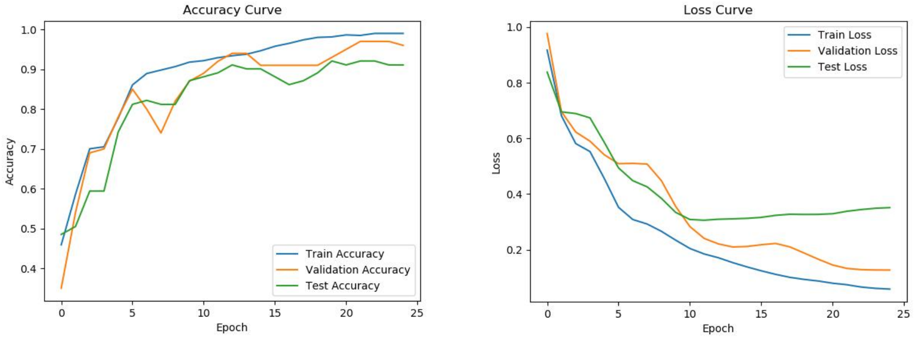


Figure 16 The training process of set 3444 resnet18

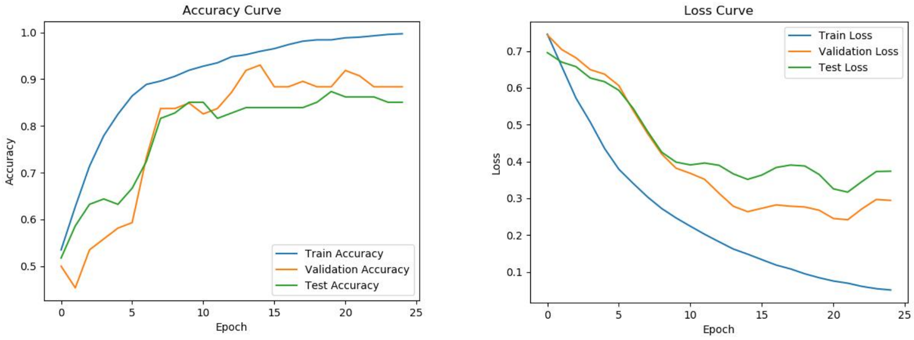


Figure 17 The training process of Set 3376 resent18

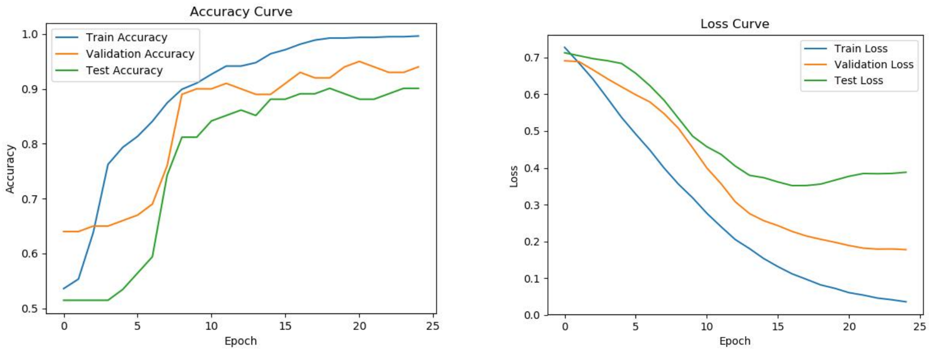


Figure 18 The training process of Set 3444 resnet50

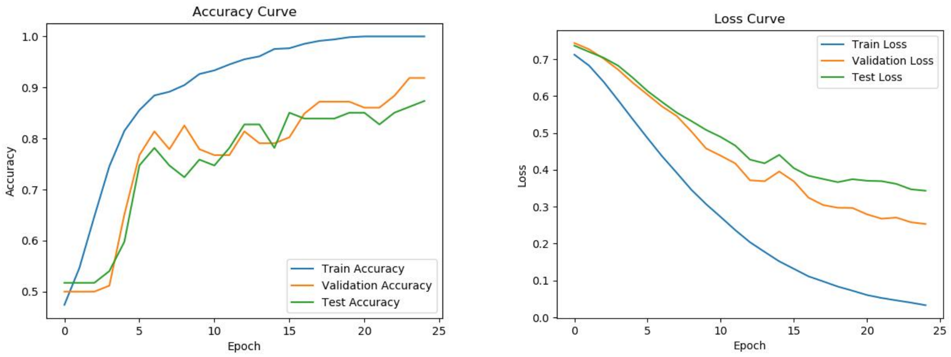


Figure 19 The training process of Set 3376 resent50

As shown in the graph of accuracy, we can find that if the accuracy of the two datasets trained with resnet18 is faster than that of resnet50, as shown in Figure 16,17 and 18,19 above, for the same training target, different convolutional neural network training verification The error curves of the sets are quite different. For classification models with similar accuracy, the deep convolutional neural network with faster convergence and less oscillation is preferred. Because we use ResNet to try to classify the test set, for the category of the training set, the average prediction accuracy rate during training is about 94.44%. The classification accuracies of sounds are 0.9971(18,76), 0.9900(18,44), 0.9963(50,76), 1.0000(50,44), respectively.

Using the trained model to test the verification set, the average prediction accuracy rate of the verification is 89.32%. The classification accuracies of resnet18 and resnet50 algorithms for identifying subset 3376 and subset 3444 for humpback calls are 0.9419 (18, 76), 0.9700 (18, 44), 0.9186 (50, 76), 0.9500 (50,44).

Then use the accuracy of the trained model on the test set (including 3376 and 3444) (also the most important accuracy)

Table 5 Test set accuracy

Experiment	Set_3376	Set_3444
method	Accuracy	Accuracy
Resnet18	0.8736	0.9208
Resnet50	0.8736	0.9010

Among them, the humpback whale call test set of data set 3376 for resnet18 and resnet50 has a relatively low accuracy rate, only 87.36%.

Through the above experiments, we can find that the general accuracy rate of 3376 sets is lower than that of 3444 sets. Through the analysis of the Perth Canyon dataset trained with resent50, I think that the background noise in the 3376 dataset used should be significantly more than that of the dataset 3444, because we Only the STFT transformation is used without any processing, and the noise is the biggest factor affecting the judgment. In addition, I think the reason why the accuracy of resnet18 seems to be slightly higher than resnet50 is that the number of samples of the humpback whale call data used is relatively small compared to the number of other research original data.

Resnet50 may cause overfitting problems, resulting in a decrease in accuracy, which can be avoided by further expanding the data set. Overall, the whale call classification algorithm based on the deep convolutional neural network described in this paper is feasible, and the optimal accuracy rate of the test set through Res Net18 and Res Net50 after many experiments is 92.08%, which is higher than the achieved A good classification effect is obtained, which proves that the deep convolutional neural network is more effective in solving the underwater whale call classification task. At the same time, it can be proved that when using a relatively small data set (about 500 data per classification), a deeper network (such as ResNet50) may lead to overfitting and the accuracy is slightly lower than resnet18.

When a model is over Overfitting occurs when the details and noise in the training data are accurately learned while ignoring the general trend of the data. This leads to the model performing very well on the training data but not performing well on new unseen data (test data) because it doesn't generalize well and since it has more parameters we can also It is found that resnet50 performs well in the training set, but the accuracy rate drops sharply in the verification set, which is an obvious overfitting performance. In contrast, ResNet18 may be more suitable for such a dataset because of its lower model complexity, and under limited computing resources, deep networks cannot be fully trained, while shallow networks may converge more easily.

5.4 Experimental results of different enhancement or filtering methods after signal processing

We can find that the first experiment determined the method of converting the audio source file into a spectrum in machine learning, the second experiment determined the type of residual network, and for the third experiment, a suitable signal processing method was selected. The enhancement or filtering method is particularly important. Here, two technologies, median analysis and per channel energy Normalization (PCEN), are mainly used and compared. At the same time, because the noise of the 3376 set is obviously more than that of the 3444 set, the 3376 set is selected as the training data set of the third experiment to test the optimization effect of the enhancement or filtering technology on the high-noise background, which can also better highlight the effect of the filtering method on the model. Improve the situation.

The third experiment is to apply the median analysis and per channel energy Normalization (PCEN) two enhanced filtering techniques to the 3376 subsets and compare their accuracy to judge the effect.

The parameters for PCEN and median normalization are: window length: 1024, Hanning window function, FFTsize of 4096, overlap of 768. In addition, the way the spectrograms are displayed remains consistent between the two methods. My MATLAB implementation uses: The spectrogram amplitude is plotted on a logarithmic (dB) scale over a fixed color range of -5 to +20 dB, the spectrogram is plotted on a mel frequency scale, and the frequency is plotted as log10(1+f/700).

The first is the respective accuracy of the models trained 10 times using median normalization training,The PCEN and median normalization accuracy were also compared to the model without any addition (using STFT only):

Table 6 Accuracy without added audio enhancement technology

Nopre_acc	Best_train	Best_val	Test_acc
1	0.9971	0.9419	0.8506
2	0.9957	0.9070	0.8506
3	0.9986	0.8721	0.8966
4	0.9971	0.9302	0.8736
5	0.9986	0.9070	0.8851
6	0.9986	0.9186	0.8621
7	0.9957	0.9186	0.8851
8	0.9971	0.9302	0.7931
9	1.0000	0.9535	0.8851
10	0.9986	0.9070	0.8506

Table 7 Accuracy after adding median normalization

Med	Best_train	Best_val	Test_acc
1	1.0000	0.9535	0.9195
2	1.0000	0.9651	0.8851
3	0.9986	0.9535	0.8621
4	0.9986	0.9186	0.8966
5	0.9986	0.9535	0.8506
6	1.0000	0.9651	0.8966
7	0.9942	0.9302	0.8621
8	0.9986	0.9651	0.8966
9	1.0000	1.0000	0.8966
10	1.0000	0.9651	0.8736

Table 8 Accuracy after adding per channel energy normalization

PCEN	Best_train	Best_val	Test_acc
1	0.9957	0.9651	0.9425
2	0.9928	0.9419	0.8736
3	0.9942	0.9651	0.9310
4	0.9986	0.9651	0.9425
5	0.9986	0.9651	0.9310
6	0.9986	0.9535	0.9310
7	0.9986	0.9535	0.9425
8	0.9942	0.9884	0.9310
9	0.9942	0.9651	0.9540
10	0.9971	0.9651	0.9310

Next, use per channel energy Normalization training to repeat the accuracy of the model for 10 times of training:

For these two sets of data, we can see that whether the model using median normalization or pcen is used, the accuracy rate is greatly improved compared with the resnet18 model that only uses STFT conversion, which proves the accuracy rate. Raising is effective. In order to explore the effect of the two techniques on eliminating background noise, the accuracy rate of the two sets of data will be calculated based on this assumption, since we are processing the same audio file on the same data set, but the processing method is different (One uses PCEN, the other uses median normalization), and I want to compare the effect of these two treatments under a certain evaluation standard, I should use paired sample t-test rather than independent t-test.

This is because each audio file gets an accuracy rate under both processing methods, and these two accuracy rates are "paired" or "correlated". Specifically, they are related because they originate from the same audio file. Therefore, in this case, it is appropriate to use a paired sample t-test.

5.4.1 Comparison without audio enhancement and median normalization

Perform a paired-sample t-test:

nopre and median:

Calculate the mean and standard deviation for each group：

For the first group :

$\mathrm{X}_1$0.8506,0.8506,0.8966,0.8736,0.8851,0.8621,0.8851,0.7931,0.8851,0.8506

$$\overline{\mathrm{X}}_1\text{(average) = 0.8671}$$

$$\mathrm{S}_{\mathrm{x}1}\ \text{(standard deviation)} \approx 0.0305$$

For the second group X_2:

0.9195,0.8851,0.8621,0.8966,0.8506,0.8966,0.8621,0.8966,0.8966,0.8736

$$\overline{\mathrm{X}}_2\text{(average) = 0.8885}$$

$$S_{x2} \text{ (standard deviation) } \approx 0.0205$$

$d_i=x_i-y_i$ here ， x_i and y_i is the observed value of the ith pair of paired data.

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n} d_i=-0.0178 \quad （10）$$

where n is the number of pairs

$$S_d=\sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}}=0.0132 \quad （11）$$

Compute the t-statistic :

$$t=\frac{\bar{d}}{s_d/\sqrt{n}}=-4.84 \quad （12）$$

use formula：

where n is the sample size of the two groups, which is 10 in this example.

Find the p-value: for a t-value of −4.84 and degrees of freedom

df=n-1=9

We can find the corresponding p-value. This p-value is significantly less than 0.05, indicating that the difference between the two sets of data is statistically significant.

5.4.2 Comparison median normalization and per channel energy normalization

median and pcen:

Calculate the mean and standard deviation for each group：

For the first group :

$X_1$0.9195,0.8851,0.8621,0.8966,0.8506,0.8966,0.8621,0.8966,0.8966,0.8736

$$\bar{X}_1\text{(average)} = 0.8885$$

$$S_{x1} \text{ (standard deviation) } \approx 0.0205$$

For the second group X_2:

0.9425,0.8736,0.9310,0.9425,0.9310,0.9310,0.9425,0.9310,0.9540,0.9310

$$\bar{X}_2\text{(average)} = 0.9321$$

$$S_{x2} \text{ (standard deviation) } \approx 0.0247$$

$d_i=x_i-y_i$ here ， x_i and y_i is the observed value of the ith pair of paired data.

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n} d_i=-0.0461 \qquad （13）$$

where n is the number of pairs

$$S_d=\sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}}=0.0217 \qquad （14）$$

Compute the t-statistic :

$$t=\frac{\bar{d}}{s_d/\sqrt{n}}=\frac{-0.0461}{0.0217/\sqrt{10}}=-6.71 \qquad （15）$$

Use the formula:

where n is the sample size of the two groups, in this case 10.

Find the p-value: for a — value of 6.71 and degrees of freedom

df=n-1=9

We can find the corresponding p-value. The p value is significantly less than 0.05, indicating that the difference between the two sets of data is statistically significant.

We can find that the accuracy of using median normalization is higher and significantly higher than that without audio enhancement technology, and pcen is more accurate and significant than median normalization. Finally, I think it is positive that there is a significant difference in accuracy between the data from these two experiments.

At the same time, the average accuracy of pcen is higher than that of median normalization (0.8885 vs. 0.9321). Therefore, it can be concluded that the recognition model using pcen is improved by 3444 sets (noisier) than the model using median normalization. subset), the accuracy is higher, and the anti-noise ability is significantly improved.

Finally, the processed model accuracy conclusion can be obtained:

Nopre : 86.71% +/- 3.05%

median normalisation：88.85% +/- 2.05%

per channel energy Normalization：93.21% +/- 2.47%

5.4.3 Comparing the accuracy of two audio enhancement techniques when they encounter unseen audio files

In summary, the model processed by pcen is better than the model processed by median normalization in terms of standard deviation and accuracy.

In order to further verify the performance of applying the median Normalization or pcen model to other humpback whale sound recognition classifications, I also tried to test the recognition accuracy of the 3444 sets (relatively less noisy) using median Normalization or pcen.

Table 9 Add median normalization and PCEN and test the accuracy of 3444 sets on the model trained on 3376 sets.

times	Med_acc	Pcen_acc
1	0.9183	0.9313
2	0.9034	0.8856
3	0.9124	0.9184
4	0.9004	0.9254
5	0.9124	0.9294
6	0.9104	0.9234
7	0.9173	0.9343
8	0.9173	0.9284
9	0.8994	0.9144
10	0.9124	0.9204

For the first group ：

$$\overline{\mathrm{X}}_1 \text{（average）} = 0.9110$$

$$\mathrm{S}_{x1} \text{ (standard deviation) } \approx 0.0068$$

For the second group ：

$$\overline{\mathrm{X}}_2\text{(average)} = 0.9211$$

$$S_{x2}\ \text{(standard deviation)}\ \approx 0.0154$$

$d_i = x_i - y_i$ here， x_i and y_i is the observed value of the ith pair of paired data.

In the same way as above, there can be：

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n} d_i=0.0107 \tag{16}$$

where n is the number of pairs

$$S_d=\sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}}=0.0126 \tag{17}$$

Calculate t-statistic：

$$t=\frac{\bar{d}}{s_d/\sqrt{n}}=\frac{0.0107}{0.0126}=2.675 \tag{18}$$

df=9

We can find the corresponding p-value, for a two-tailed test and 9 degrees of freedom, the critical t-value is about 2.262. Our calculated t-value of 2.675 is greater than the critical value, so we reject the null hypothesis that there is a significant difference between the two groups of data.

in conclusion:

There is a significant difference in the accuracy of the two sets of data.

The average accuracy of the second group is higher (0.9110 vs. 0.9211) and the statistical conclusion is significant, so that it can be concluded that the recognition model trained with pcen on 3444 sets (noisier subset) is better than 3376 sets (not many The recognition ability of a subset of noise) is better than that of the model using median normalization, and this anti-noise ability is significantly superior.

Finally, the model accuracy conclusion after processing can be obtained:

median normalisation: 91.10% +/- 0.68%

per channel energy Normalization: 92.11% +/- 1.54%

In summary, the accuracy of the model processed by pcen is not only on the test set but also on the unlearned 3444 set. When processing audio, the verified effect is better than the model processed by median normalization.

6. Conclusion

This article analyzes the accuracy of the humpback whale call classification task, uses the deep convolutional neural network structure and audio filtering or enhancement technology that have been widely used in recent years to solve this problem, and explains and verifies the superiority of this method. . Based on this direction, the advantages and disadvantages of several commonly used design methods are implemented, compared and analyzed, and taolun, which combines the above two aspects and the depth of deep neural networks, is introduced to improve the classification performance of neural networks. The main work of the paper is summarized as follows:

(1) Processing of humpback whale sounds: The accuracy of two commonly used and effective audio time-frequency analysis methods, namely STFT and CQT, was compared. Among them, STFT has the same frequency details in the high-frequency and low-frequency ranges. Although the information is too complicated, it is a more suitable choice for humpback whales, which have very complex and diverse sound frequencies. CQT has higher frequency resolution in the low-frequency range and higher time resolution in the high-frequency range, which allows CQT to capture information in the low-frequency range in more detail, which results in high-frequency humpback whales. sound may be missed. Since the main purpose is to eliminate background noise, although STFT results in too much redundant information, it also retains the most effective information. In this case, the increase in dimension will not increase the computational complexity, so STFT is chosen as the time frequency. Analytical method.

(2) Comparison and analysis of deep convolutional neural networks: analyze the advantages of convolutional neural networks and deep networks, and compare the performance differences caused by network differences of widely

used deep convolutional neural network models, mainly including: Res net18 and Res Net50 . Through experiments to compare the classification performance of these two types of neural networks, the results show that Res Net18 has better classification ability than 50, with a classification accuracy of 92.1%. At the same time, this article compares various performance indicators, such as ROC curves and confusion matrices, to verify the accuracy of the test set in the experiment, the convergence speed and accuracy. It also verifies that when the purpose is to eliminate background noise, we Using the Res net18 algorithm is more effective for the training optimization process when processing small data sets.

(3) Comparison and verification of different audio filtering and enhancement technologies: Two common audio enhancement and filtering technologies are compared. Median normalization is simpler and can eliminate the influence of many outliers, but it may also lose many important values. Information; PCEN is a technology with more complex calculations that can effectively reduce the noise in background sounds. Among them, PCEN has higher confidence in the classification of humpback whales, which have a wide range of sound frequencies. Therefore, in the context of this purpose, PCEN Obviously more suitable for the classification task we want to achieve.

7. Discussion

7.1 Datasets and time-frequency analysis

For humpback whale sound recognition tasks, the quantity and quality of data are key. Humpback whale sounds can be obscured by various ocean background noises, making detection more difficult. In addition, diversity in the data is also an important factor, given that whales may make a wide variety of sounds, from low-frequency calls to high-frequency hits. Time-frequency analysis methods played a key role in the identification of humpback whale sounds. Traditional spectral features, such as short-time Fourier transform and constant Q transform, have been widely used. These features allow us to capture the frequency domain characteristics of the sound signal, allowing the machine learning model to identify specific whale sound patterns.

Even though we compared the two methods and learned that STFT is more suitable for whales like humpback whales, whose sounds have a wide range of frequency distribution, humpback whale sounds can also contain many different modulations, from rapid calls to long songs. . For fast sound events, high temporal resolution is required, while for long duration sound patterns, high frequency resolution is required. With STFT, it's difficult to get both. Although CQT can provide high resolution for low-frequency content, humpback whale songs generally contain important high-frequency content, and CQT has low resolution at high frequencies, so the time-frequency method also has its own limitations. .

7.2 Selected models and methods

Convolutional neural network (CNN) has excellent performance on images and spectrograms, so converting sound data into spectrograms and then processing them as image recognition is the best choice for this engineering task. Because CNN can not only automatically learn and extract features in the spectrogram, it does not require any complicated training process. However, it is difficult to choose the appropriate network structure and parameters, because they directly affect the generalization ability and performance of the model. For example, the learning rate of resnet18, the value I chose is 0.01, which is also a value determined after many attempts. Different The residual network may have different suitable learning rates, so I think the limitation of my selected model is that there may be a learning rate that is more suitable for my selected model, but I have not found it.

7.3 Interpretation of results for residual network models

Judging from the experimental results, the two residual network models perform very well on the training set, but their performance on the validation and test sets may decline. Although resnet50 has more layers and is more complex, the reason why its performance is not as good as resnet18 may be due to the complexity of overfitting or insufficient data enhancement, because more network layers mean that the model may have learned enough noise. In order to solve this problem, in addition to selecting an appropriate network structure, more data enhancement techniques need to be used.

7.4 Audio enhancement technology

I compared the model performance of multiple groups adding pcen and median normalization techniques and conducted a t test to see if there was a significant difference. Although in the end I found that using PCEN and median normalization both provide effective tools for audio processing. Moreover, PCEN is more suitable for noisy environments and can provide robust recognition of the sounds of whales and other creatures. Median normalization, on the other hand, provides a simple, computationally efficient, and robust to outliers approach to data processing. However, choosing the most appropriate technique depends on the specific application and data. Under appropriate circumstances, it may not require so many steps. For example, when the audio file does not have a lot of background noise, median normalization can be used.

7.6 future work

For future work, semi-supervised or unsupervised learning methods could be considered to leverage unlabeled data. In addition, introducing transfer learning or meta-learning methods may also help the model's performance in different environments and situations. Finally, in order to detect humpback whale sounds in real time, the real-time nature and deployment of the model is also an important aspect to consider.

7.7 Overall

Machine learning-based sound recognition for humpback whales has made encouraging progress, but further research and improvements are still needed to improve its effectiveness in practical applications. For the sound recognition task of humpback whales, it can be optimized by three aspects: time frequency analysis method, convolutional neural network model and sound enhancement technology, because the sound of humpback whales may be masked in various ocean background noises, thus increasing the the difficulty of detection. In addition, diversity in the data is also an important factor, given that whales may make a wide variety of sounds, from low-frequency calls to high-frequency hits.

Reference

1. P. J. Dugan, A. N. Rice, I. R. Urazghildiiev and C. W. Clark, "*North Atlantic Right Whale acoustic signal processing: Part I. comparison of machine learning recognition algorithms,*" 2010 IEEE Long Island Systems, Applications and Technology Conference, Farmingdale, NY, USA, 2010, pp. 1-6, doi: 10.1109/LISAT.2010.5478268.

2.White, E.L. *et al.* (2022) 'More than a whistle: Automated detection of marine sound sources with a convolutional neural network', *Frontiers in Marine Science*, 9, p. 879145. Available at: https://doi.org/10.3389/fmars.2022.879145.

2. Li, Maopeng *et al.* (2022) 'Multi-scale Sparse Network with Cross-Attention Mechanism for image-based butterflies fine-grained classification', *Applied Soft Computing*, 117, p. 108419. Available at: https://doi.org/10.1016/j.asoc.2022.108419.

3. Debnath, L. (1998) 'Brief historical introduction to wavelet transforms', *International Journal of Mathematical Education in Science and Technology*, 29(5), pp. 677–688. Available at: https://doi.org/10.1080/0020739980290504.

4. Li, Maopeng *et al.* (2022) 'Multi-scale Sparse Network with Cross-Attention Mechanism for image-based butterflies fine-grained classification', *Applied Soft Computing*, 117, p. 108419. Available at: https://doi.org/10.1016/j.asoc.2022.108419.

5. Haralick, R.M., Shanmugam, K. and Dinstein, I. (1973) 'Textural Features for Image Classification', *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6), pp. 610–621. Available at: https://doi.org/10.1109/TSMC.1973.4309314.

6. Lin, T.-C. (2007) 'A new adaptive center weighted median filter for suppressing impulsive noise in images', *Information Sciences*, 177(4), pp. 1073–1087. Available at: https://doi.org/10.1016/j.ins.2006.07.030.

7. Lostanlen, V. *et al.* (2019) 'Per-Channel Energy Normalization: Why and How', *IEEE Signal Processing Letters*, 26(1), pp. 39–43. Available at: https://doi.org/10.1109/LSP.2018.2878620.

8. Cartwright, M. *et al.* (2019) 'Tricycle: Audio Representation Learning from Sensor Network Data Using Self-Supervision', in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA: IEEE, pp. 278–282. Available at: https://doi.org/10.1109/WASPAA.2019.8937265.

9. LeCun, Y. *et al.* (1989) 'Backpropagation Applied to Handwritten Zip Code Recognition', *Neural Computation*, 1(4), pp. 541–551. Available at: https://doi.org/10.1162/neco.1989.1.4.541.

10. Wu, S. *et al.* (2018) 'A Deep Residual convolutional neural network for facial keypoint detection with missing labels', *Signal Processing*, 144, pp. 384–391. Available at: https://doi.org/10.1016/j.sigpro.2017.11.003.

11. He, K. *et al.* (2015) 'Deep Residual Learning for Image Recognition'. arXiv. Available at: http://arxiv.org/abs/1512.03385 (Accessed: 13 September 2023).

12. Lyu, R. *et al.* (2021) 'Network Intrusion Detection Based on an Efficient Neural Architecture Search', *Symmetry*, 13(8), p. 1453. Available at: https://doi.org/10.3390/sym13081453.

13.Zhang, M. *et al.* (2018) 'Recurrent attention network using spatial-temporal relations for action recognition', *Signal Processing*, 145, pp. 137–145. Available at: https://doi.org/10.1016/j.sigpro.2017.12.008.

14. Hu, Z., Cui, J. and Lin, A. (2023) 'Identifying potentially excellent publications using a citation-based machine learning approach', *Information Processing & Management*, 60(3), p. 103323. Available at: https://doi.org/10.1016/j.ipm.2023.103323.

15. Arubi, S.L. *et al.* (2020) 'WELL TEST ANALYSIS AND INTERPRETATION: THE USE OF ARTIFICIAL NEURAL NETWORK', *International Journal of Engineering Applied Sciences and Technology*, 04(11), pp. 438–446. Available at: https://doi.org/10.33564/IJEAST.2020.v04i11.079.

16. Rasmussen, J.H. and Širović, A. (2021) 'Automatic detection and classification of baleen whale social calls using convolutional neural networks', *The Journal of the Acoustical Society of America*, 149(5), pp. 3635–3644. Available at: https://doi.org/10.1121/10.0005047.

17. Hevner, A.R. *et al.* (no date) 'Design Science in Information Systems Research'.

18. Dandan Zhu and Yan Cui (2017) 'Understanding random guessing line in ROC curve', in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, China: IEEE, pp. 1156–1159. Available at: https://doi.org/10.1109/ICIVC.2017.7984735.

Appendix

An online storage system was used during the project to track modifications to all files and to maintain backups of simulation data and results. The software uses Git's online repository service. After the project is implemented, a new repository is created to share all scripts and data. There is also a copy with the thesis added. In order to clone the repository, use the following command:

URL MW8U22/code-of-dissertation (github.com)

Version:

Python 3.5.2

NumPy 2.0.0, SciPy 0.19.0 and Matplotlib 2.0.0 and librosa 2.0.0.