Springboard

Data Science Career Track

Capstone Project 3

Whale Detection with Audio Recordings

Emily Bocim

July 2020

Table of Contents

Introduction

Detection of whales is a topic that can mutually benefit whale populations, researchers and industries that involve oceanic travel . The increase of human presence in the oceans comes with an increase in noise pollution that interferes with whale communication. There also is a higher risk of physical injury to the whales from collisions or damaging sound frequencies. For researchers, this can assist with tracking whale populations. Whale collisions can also be costly and dangerous for those operating the vessels.

Client

NOAA tracks North Atlantic right whales so that they can communicate to those operating in that region, the best practices for avoiding whale strikes. The main focus is on speed restrictions in designated areas. These designated areas change based on the season and whale migration habits. If need be, certain areas may not be open for shipping and ship routing may need to be addressed. In order to optimize these regions and time frames, NOAA must have reliable ways of tracking the whales.

Database

The data comes from the “Real-time Monitoring System for Detecting North Atlantic Right Whales.” Through the efforts of Marinexplore and Cornell University, the data was made available on Kaggle as 2 second labeled clips. The labeled data can be found at:

https://www.kaggle.com/c/whale-detection-challenge/data

Data Wrangling

Data Summary:

The dataset contains 30,000 hydrophone clips that are identified as either containing the sound made by a North Atlantic right whale or not. Each clip is in the format of an AIFF file. Of these files, there are 7,027 that are identified as containing a whale sound.

Missing Data:

Missing data was not an issue for this project, however it is suspected that some files were mislabeled. This is something that could be addressed by going back through and verifying labels. However, verifying labels was not done for this project since it is thought that mislabels are minimal and the gain would not be significant enough for the amount of time it would take to verify all labels.

Whale Audio
Emily Bocim

Data Exploration

Sound Files:

Each sound file is a 2 second clip with the channel being mono, and having a framerate of 2000. When we listen to a file we can generally comprehend basic ocean noise. The problem is being able to define what a North Atlantic right whale sounds like. Many of the files that contain a whale call, it is barely audible through water sounds.

Wave Plots:

Using the Librosa package, the audio files could be loaded for visualizations. The most basic of these is a wave plot, which plots amplitude versus time. The amplitude of a sound wave is measured in decibels (dB) or is more commonly referred to how loud the sound is. Figure 1 is an example of a wave plot of an audio that does not contain a whale call and figure 2 is an example of one containing a whale call. Unless it is known right when the whale call is occurring, it is hard to say what the difference is between the two clips.

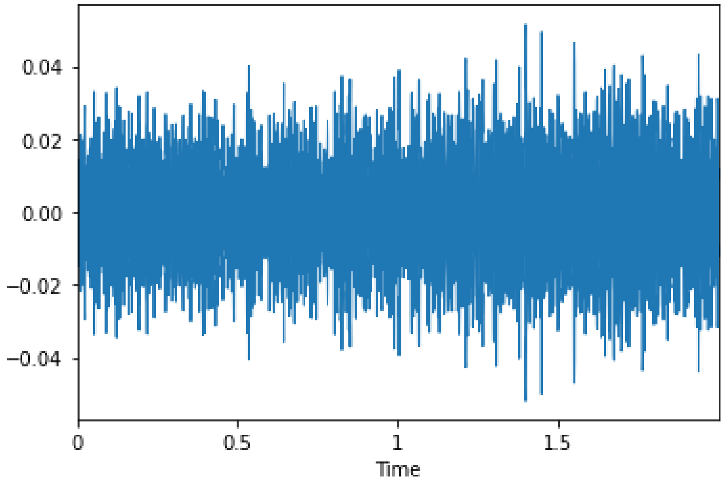


Figure 1: Wave plot with dB versus seconds for an audio clip that does not contain a whale call.

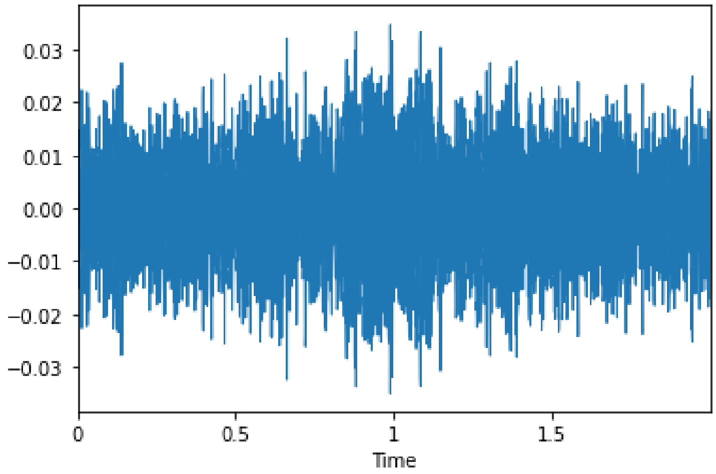


Figure 2: Wave plot with dB versus seconds for an audio clip that does contain a whale call. Visually, it does appear that something different is happening around 1 second, but it is not clear enough to confirm without further exploration.

Noise Reduction:

Using the noisereduce package, the whale call can be isolated for clearer visualizations. The base idea of this package is that it takes audio that is identified as noise and removes that type of noise from the audio that is identified as containing the sound to be kept. It is important to note, this method was attempted on all files, but even using one noise file against all whale files, there was too much loss of information. For the purpose of this project, noisereduce was only used to visually understand what the model needed to find, it was not used on the data that was used to train the model. Figure 3 shows an example of what was found after removing the noise. It appears as if there were two sounds, both with decreasing amplitudes. Something to also note is the difference in the range of scales. The amplitude scale is a quarter of what it was before noise reduction.

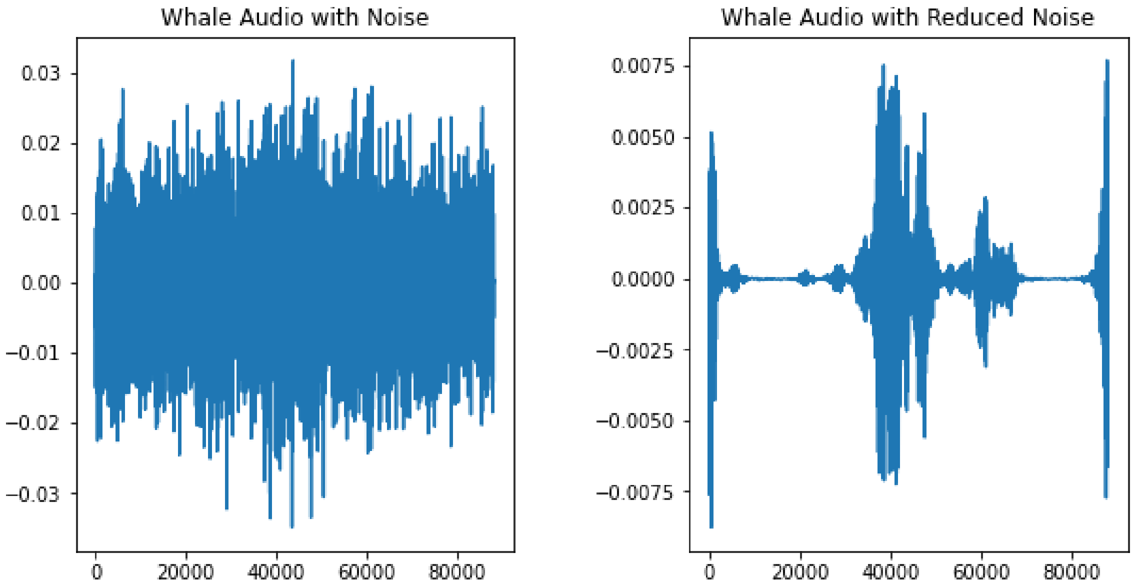


Figure 3: A comparison of wave plots for an audio that contains a whale signal before and after noise reduction.

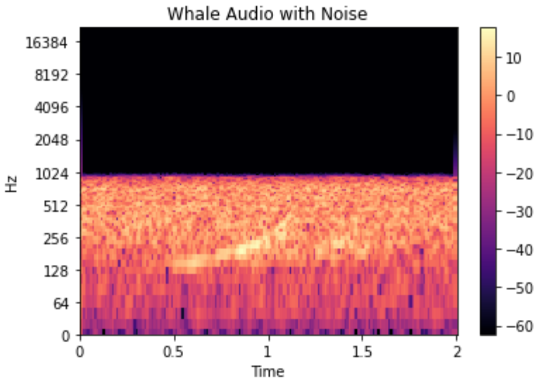


Figure 4: Mel spectrogram of whale audio without noise reduction, where the lightest section between 0.5 and 1 second is the upcall.

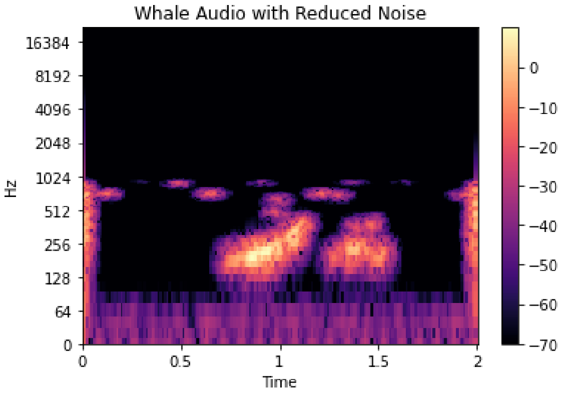


Figure 5: Mel spectrogram of whale audio with noise reduction to clarify where the upcall is located. As previously discussed, noise reduction appears to remove part of the upcall as it does not extend all the way between 0.5 and 1 second.

Mel Spectrogram:

Mel spectrograms are one of the most commonly used features for audio analysis. Part of that being because, visually, they better represent what is being heard. Mel spectrograms plot frequency (Hz) versus time, with an added color scale representing decibels. In Figures 4 and 5, a more distinct visual of what the North Atlantic right whale call looks like can be seen. This is what is called an upcall or whoop. It is a relatively short sound with increasing frequency and, as suggested in the wave plot, decreasing loudness.

Chroma:

A chroma feature or vector is typically a 12-element feature vector indicating how much energy of each pitch class, {C, C#, D, D#, E, …, B}, is present in the signal (Chauhan, 2020). It is similar to frequency, so it is not surprising that the upcall can be seen through pitches as well, as in figure 6. Each whale did not seem to have the same pitches, but instead there was the trend of increasing pitch.

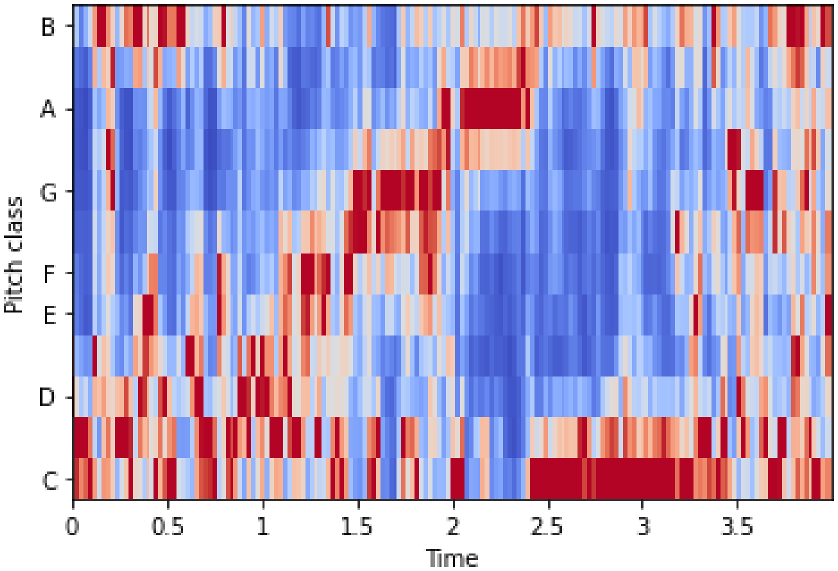


Figure 6: Chroma of a North Atlantic right whale upcall. Each tick is a quarter of a second, so time is still 2 seconds. In this case, red represents higher energies and blue represents lower energies.

Mel-frequency Cepstral Coefficients (MFCC):

Like Mel Spectrograms, this is another feature that is commonly used in audio analysis. Comparatively, it is more computationally expensive than many of the other features, as it is commonly described as a process of transforming the wave plot to a spectrum, then taking the log of the magnitude of the spectrum, and ending with a cepstrum. Cepstrum is the rate of change in spectral bands (Nair, 2018). Visually, this comes out to something between a mel spectrogram and chromagram and does not necessarily help with understanding the sound, but still contains a lot of information that can be useful to building a model.

Other Features:

Features that were explored, but did not show any distinct trends included spectral centroids and spectral rolloff. The spectral centroid is a measure used to characterise an audio spectrum by finding its center of mass. It is also connected to the brightness of a sound, which refers to the higher mid and treble parts of the frequency (Kaggle, 2020). Spectral rolloff is a measure of the shape of the signal. It represents the frequency at which high frequencies decline to 0. To obtain it, we have to calculate the fraction of bins in the power spectrum where 85% of its power is at lower frequencies (Chauhan, 2020).

Data Modeling

Feature Extraction:

The features that were chosen for training models included Mel Spectrograms, Chroma, and MFCC. Using Librosa, each of these features were generated as arrays.

Model Type:

The chosen features, though being represented numerically to save space and processing time, are types of images. This means that a convolutional neural network will be trained for detecting the presence of audio signals.

Preprocessing Data:

A CNN expects the features to have certain types of shapes. For features, the main issue was defining channels. Greyscale was chosen, because color itself is not important to sound. Issues with the target shape were resolved with label encoding.

Training CNN:

Mel Spectrogram CNN Architecture:

1. Input
2. Two Hidden Layers
3. Pooling Layer
4. Output

Multi Feature CNN Architecture:

Three models with input, two hidden layers and a pooling layer. The three models are concatenated and then enter the output layer.

Model Metrics:

The multi feature model performed the same as the model with only the Mel Spectrogram, so there will only be a focus on the metrics for the single feature CNN.

1. **Accuracy:**
   a. Training Accuracy: 0.93
   b. Testing Accuracy: 0.90
2. **Loss:**
   a. The lowest loss was 0.23 at 9 epochs (Figure 7).

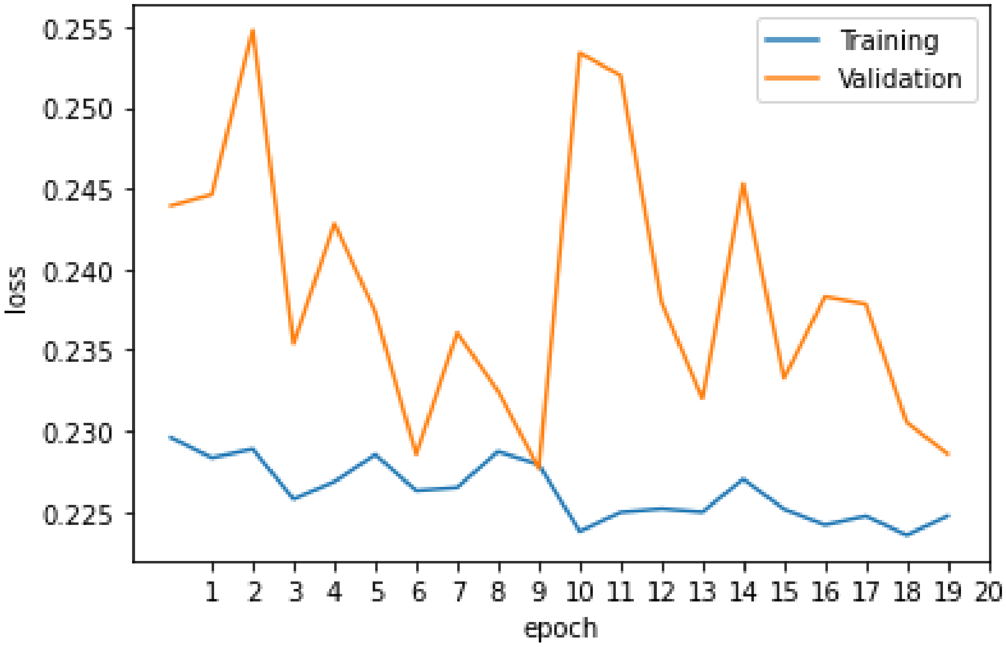


Figure 7: Mel Spec CNN model loss.

3. **Confusion Matrix:**

    a. For this project, a label of 1 indicated that the audio file contained a whale call.

    b. Almost 7% of the data was classified as a false negative, and 3% was classified as a false positive, for a total of 10% being incorrectly classified. Figure 8 shows all the percentages.

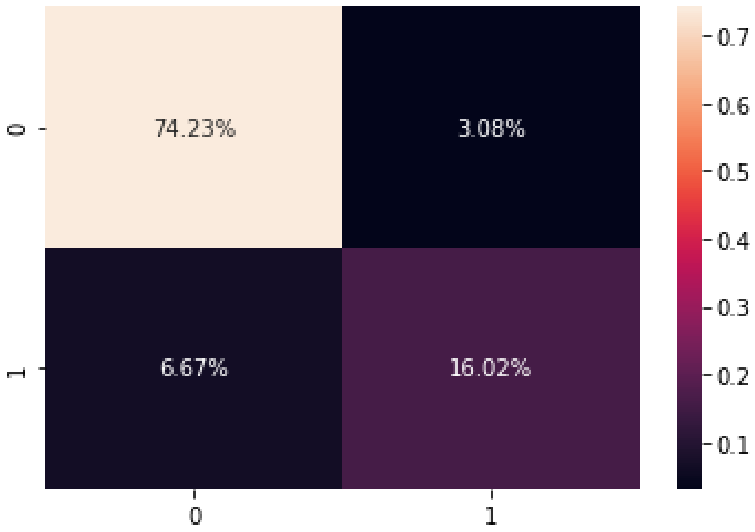
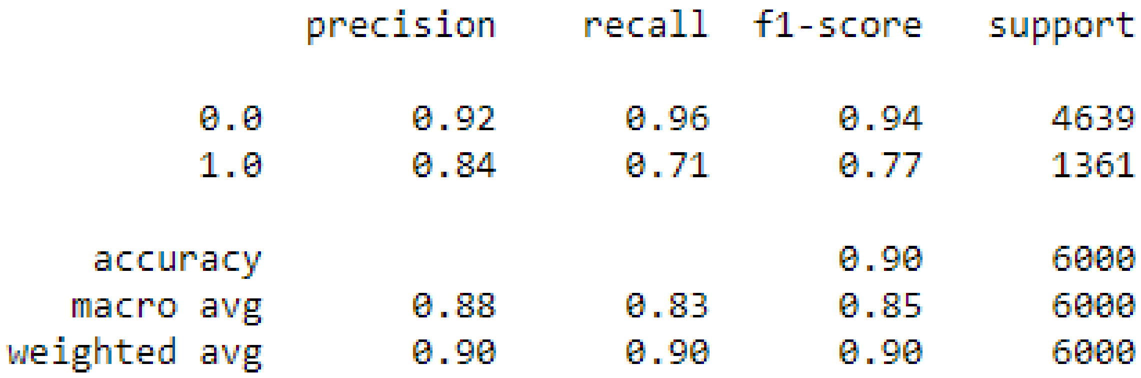


Figure 8: Confusion Matrix for CNN classification.

4. **Precision, Recall, F1:**

    a. Figure 9 shows scores for precision, recall, and F1.



```
              precision    recall  f1-score   support

         0.0       0.92      0.96      0.94      4639
         1.0       0.84      0.71      0.77      1361

    accuracy                           0.90      6000
   macro avg       0.88      0.83      0.85      6000
weighted avg       0.90      0.90      0.90      6000
```

Figure 9: Metric Scores for CNN classification.

Assumptions and Limitations

The model was trained under the assumption that all the labels were correct, though it is suspected that some of them were mislabeled. Audio is collected continuously through hydrophones. Due to being submerged in water and continuous collection, the audio quality is limited.

Recommendations for Future Work

It would be beneficial to increase the categories. The audio files that are indicated as not containing North Atlantic right whales do contain the sounds of different types of ships and other whales. The main issue with the model was false negatives. A possibility is that the North Atlantic right whale upcall is similar to another type of whale or ship sound. By adding in the other categories, the model becomes useful to more groups and potentially more accurate as it can learn to distinguish between the different types of sounds.

Conclusions

A convolutional neural network using only the Mel Spectrogram feature was trained to categorize whether an audio file contained a North Atlantic right whale upcall or not. Other features were explored, but it was found that this feature performed as well on its own as it did when trained with other features. The predominant type error was type 2, but the model still had an accuracy of 90% on the training data. In the future it could be beneficial to train the model with more labeled categories that include ships and different types of whales.

References

Chauhan, Nagesh. “Audio Data Analysis Using Deep Learning with Python”. 2020. <https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-2.html>

Kaggle. “The Marineexplore and Cornell University Whale Detection Challenge”. 2013. <https://www.kaggle.com/c/whale-detection-challenge/data?select=whale_data.zip>

Nair, Pratheeksha. “The dummy’s guide to MFCC”. 2018. <https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd>

NOAA. “Reducing Vessel Strikes to North Atlantic Right Whales”. 2021. <https://www.fisheries.noaa.gov/national/endangered-species-conservation/reducing-vessel-strikes-north-atlantic-right-whales>