CS230

Learning Blue Whale Calls from Underwater Audio Streams (Sound Recognition)

Kanoe Aiu, Kendall Ota
kanoeaiu@stanford.edu, kcota@stanford.edu

Abstract

Our project sought to develop an improved computer vision algorithm to identify blue whale D calls in acoustics data. The nearby Monterey Bay is part of an essential blue whale migration corridor, and often an important foraging stop along the migratory root. D calls are not well understood, but thought to be associated with foraging behavior. We were motivated to create an accurate and efficient detection algorithm to allow for real time identification of D calls, increasing biological insight into the behavior and movement of the increasingly threatened species and informing dynamic protection efforts. In doing so, we utilized both a pre-existing convolutional neural network and our original transformer, which operate on spectrograms of acoustic data and raw audio, respectively. Our algorithm then outputs a score from 0 to 1 based on whether it thinks an image or sound contains a blue whale D call. We evaluated it's performance via comparison with true values, including false positives and false negative rates, as well as failure patterns by examining temporal patterns in accuracy and sounds that generally confuse the detector.

1 Introduction

Blue whales are a migratory marine species found globally, with shifts in calling behavior correlated with migratory onset. Blue whale song is incredibly varied, and relatively simple to capture through hydrophones. However, blue whale D calls are some of the most cryptic calls, in part because a number of other species exhibit very similar call types, making it challenging to process and identify such acoustics. We will develop an improved computer vision algorithm to identify blue whale D calls in acoustics data. The input to our algorithm is acoustics data from a hydrophone deployed in Monterey Bay, and we will then use either CNNs or transformers to process and identify D calls. Identifying D calls is increasingly important, as D calls are thought to be associated with foraging, a behavior threatened by climate induced shifts in krill populations.

2 Related work

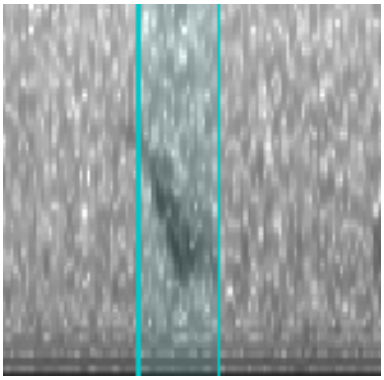
Passive acoustic monitoring (PAM) is revolutionizing our ability to study marine species, allowing scientists to probe into the relationship between vocalizations and behavior and obtain high resolution time series of presence/absence data (Kowarski, 2021; Lewis, 2018). For example PAM was instrumental in revealing habitat use patterns of threatened orca populations (riera), social cueing behavior in migrating blue whales (Oestreich, 2020), and population density of beaked and sperm whales in the Bahamas (Thomas, 2012).

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Recently, advances in machine learning and cloud based computational resources have contributed to the growing popularity of machine learning algorithms in processing acoustic data. Efforts to apply machine learning to Blue whale vocalizations include the development of a clustering algorithm to differentiate between different types of whale vocalizations (Bahoura, 2010), and recognition algorithms focused on populations in the Southern Ocean (Rasmussen, 2021). However, these attempts have not yet focused on north Pacific whale populations, and have not achieved enough accuracy to be used in construction of biologically significant data sets. Our work will incorporate the successes and failures of previous researchers to build a detector that is simple, efficient, and accurate. We will incorporate the well-developed clustering methods into a neural network architecture, and include biological parameters with known correlations with whale presence as well. With an accurate and efficient detector, the Monterey Bay Aquarium Research Institute (MBARI) will be able to run live detection on their hydrophone data and expand biological understanding of blue whale vocalizations.

3 Dataset and Features

Out dataset was provided publicly by MBARI via the Pacific Sound Database. There are over 10 years of data on which to test the algorithm. A small labeled dataset was provided, and we examined 48 hours of additional audio data to identify more examples. Examples were selected from a variety of years, times of year, and biological/contextual significance of false calls (i.e. boat noises, other whale calls, etc.). For the initial model, the labeled data set was preprocessed by running a band limited energy detector to identify potentially interesting time stamps, computing the mel spectrogram of the audio at the times identified. Some filtering blurring and color shifting was applied. The dataset contains 2000 total samples with a train-dev-test split of 90, 5, 5. For the transformer, which takes raw audio, data augmentation steps included adding noise and distorting the audio, as well as overlaying audio from different time stamps.



4 Methods

We began with using an existing convolutional neural network (CNN) to lay a foundation. We used a pre-trained res-net-50 which ran transfer learning. This CNN had used stochastic gradient descent to train the weights for each neuron, and operated on spectrograms of acoustic data, on which data augmentation (including adding noise and blurring) had already been performed. The CNN worked by taking in filters and using them to find relationships between adjacent pixels. Since this data has a known temporal component and the D call spans known frequencies, any transformations would erode the foundational assumptions about the data set. The following loss function was used:

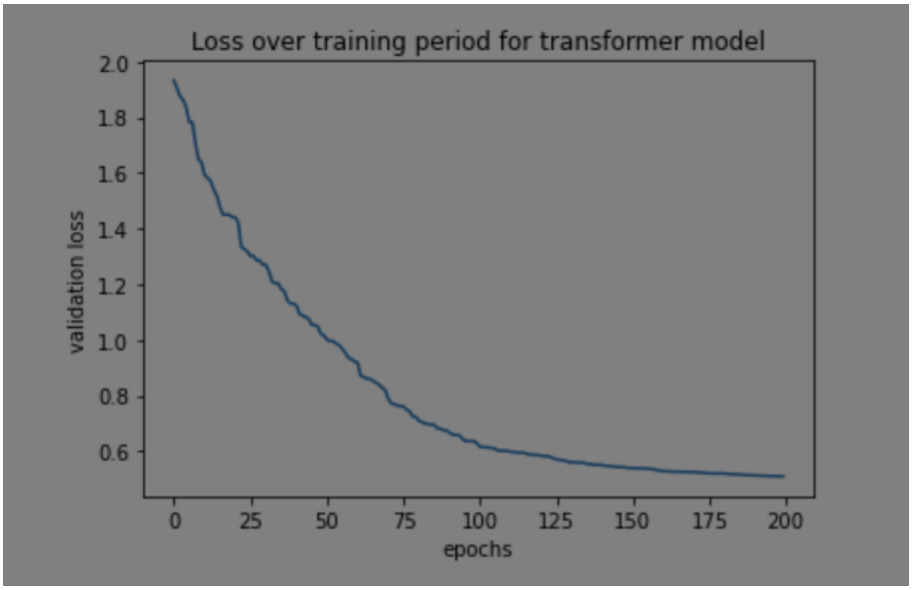
$$-(ylog(p) + (1-y)log(1-p)).$$

We also utilized a transformer which operated on raw audio, for which we performed data augmentation. To do so, we used transfer learning from Wav2vec, a pre-existing model that recognizes speech from raw audio data. The transformer worked by encoding relationships between time steps, and using this information to essential run a neural network. We opted to employ a transformer to allow for broader context use as compared to CNNs by encoding memory across time. We continued to use a cross entropy loss function.

5 Experiments/Results/Discussion

Transformer performance:
For hyperparameters, we used an initial learning rate of 10^-2 decaying exponentially. We trained for 100 epochs with a batch size of 32. We evaluated our model on accuracy. Our model rarely detected false positives, but observed a significant number of false negatives.

Graph of loss:

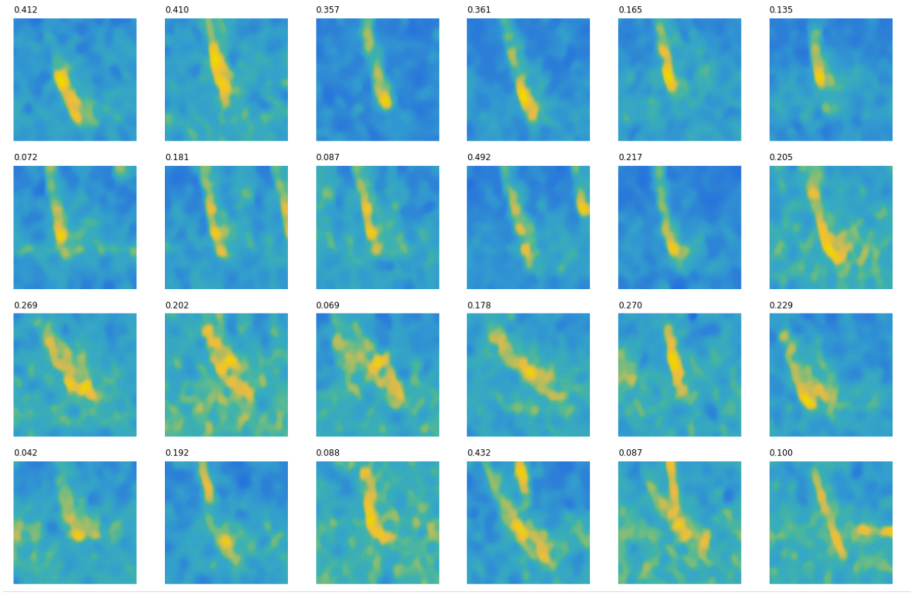


Training loss: 0.33
Dev loss: 0.41

Confusion matrix:

	label true	label false
BDT	872	102
BDF	3	1023

Precision: 99%
Recall: 90%
Accuracy: 95%



CNN performance:

Given the training loss and dev loss, we overfit to the training set a bit. We used a mini batch with a relatively small batch size to prevent overfitting.

6 Conclusion/Future Work

Due to time and computational limitations, we were not able to vary our batch size or training rate. Varying batch size, as well as running experiments to understand how more data impacts training and evaluate the upper limit of value for increasing dataset size would be interesting. Additionally, with more time, we would experiment with transformers pretrained on non-human audio data would.

7 Contributions

We used pair programming for all work and training done with our CNN and transformer. We also collaborated entirely on the project proposal, milestone, and final project. Any sections not directly written by a partner were profusely edited and proofread by the,.

8 References

[1] Bahoura, M., & Simard, Y. (2010). Blue whale calls classification using short-time Fourier and wavelet packet transforms and artificial neural network. Digital Signal Processing, 20(4), 1256-1263.

[2] Kowarski, K. A., & Moors-Murphy, H. (2021). A review of big data analysis methods for baleen whale passive acoustic monitoring. Marine Mammal Science, 37(2), 652-673.

[3]Lewis, L. A., Calambokidis, J., Stimpert, A. K., Fahlbusch, J., Friedlaender, A. S., McKenna, M. F., ... Širović, A. (2018). Context-dependent variability in blue whale acoustic behaviour. Royal Society open science, 5(8), 180241.

[4] Oestreich, W., Cline, D. E., Cade, D., Calambokidis, J., Fahlbusch, J., Joseph, J., ... & Ryan, J. P. (2020, February). Temporal variations in blue whale call types in the northeast Pacific at diel, seasonal, and interannual time scales with tag-derived behavioral context. In Ocean Sciences Meeting 2020. Agu.

[5] Rasmussen, J. H., & 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), 3635-3644.

[6] Thomas, L., Marques, T. A. (2012). Passive acoustic monitoring for estimating animal density. Acoustics Today, 8(3), 35-44.