Detecting Endangered Baleen Whales within Acoustic Recordings using R-CNNs

Mark Thomas[1,2,*], Bruce Martin[2], Katie Kowarski[2], Briand Gaudet[2], and Stan Matwin[1,3]
[1] Faculty of Computer Science, Dalhousie University, Halifax, Canada
[2] JASCO Applied Sciences, Dartmouth, Canada
[3] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[*]mark.thomas@dal.ca

Abstract

Research and development into automated systems that can detect the vocalizations of endangered species of whales within acoustic recordings is a difficult yet important task. Over the past several years, hundreds of deceased whales have washed ashore along the coasts of North America. In many cases the primary cause of death of these species has been directly linked to human activity including vessel collisions and entanglement in fishing gear. In this work, we introduce preliminary work towards developing an end-to-end detection system using a Region-based Convolutional Neural Network (R-CNN) trained on spectrogram representations of acoustic recordings and labelled bounding boxes around the vocalizations of three species of endangered baleen whales: blue, fin, and sei whales. In this way, the R-CNN can detect vocalizations in terms of both time and frequency against a background of ambient noise and other non-biological sources. The R-CNN can be used by stakeholders and policy makers to mitigate the risk of collisions and entanglements when the aforementioned species are detected in a given area.

1 Introduction

Conservation studies have shown that human activity has been detrimental to over 40 percent of the ocean's ecosystems [3]. In some cases, human activity has been directly linked to population decline for various species of marine life [15]. A large contributor to the more recent (i.e., 21st century) decline in population of baleen whales is vessel collisions in busy shipping channels as well as entanglement in fishing gear. In order to reduce the risk of collisions and entanglement, the Canadian government has started to impose speed restrictions and temporary fishing bans when endangered whales are present in protected areas. However, determining whether a given species is present or not can be a difficult task and often the aforementioned restrictions are imposed after an incident or death has occurred rather than as a safe measure.

One of the more robust techniques used to determine presence/absence of endangered whales is through the analysis of acoustic recordings. Such analysis is often referred to by marine biologists as Passive Acoustic Monitoring (PAM) and is preferred to GPS tagging and visual surveys as it is non-invasive and less susceptible to poor weather conditions [16]. Often, the necessary acoustic recordings used for PAM are collected using specialized hardware fitted with hydrophones and several terabytes of storage. These devices can be left static for several months at a time and produce very large quantities of data. PAM can also be carried out in real-time using ocean gliders or an array of hydrophones towed behind a vessel. In this work, we employ a Region-based Convolutional Neural Network (R-CNN) trained on the former category of acoustic recordings, i.e.: a large corpus of recordings that were collected using moored devices off the coast of Atlantic Canada. In particular, the R-CNN is trained in the frequency domain using spectrograms to detect bounding boxes around

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

the vocalizations of three species of endangered baleen whales: blue whales (*Balaenoptera musculus*, BW), fin whales, (*Balaenoptera physalus*, FW), and sei whales (*Balaenoptera borealis*, SW).

We propose that the R-CNN can be used in almost real-time on board a vessel, an ocean glider, or a moored device with telecommunication capabilities to determine presence/absence of the aforementioned species of endangered whales. By continuously monitoring for the presence of endangered species in close to real-time, policy makers can enforce restrictions on vessels and fishing more effectively.

1.1 Related Work

Research and development into automated systems for detecting marine mammals within acoustic recordings has been a topic of interest for many years [8, 11]. Much of the research that has taken place over the past decade or so has been focused on designing specialized detection algorithms using templates or hand-engineered features pertaining to different types of vocalizations produced by a specific species [1, 9, 12, 14]. These systems are often not able to generalize to new sources of noise or additional species of marine mammals that were not considered during the design of the algorithm. More recently, several researchers have used CNNs to develop more generalizable systems that are capable of determining whether particular species are present or absent within samples of a full recording [7, 13]. While the results of these systems are quite promising, they are only capable of determining presence/absence in terms of time. Additionally, they are limited to detecting one species per sample even if multiple vocalizations from different species are present within the same sample.

To the best of our knowledge, no work has previously been reported towards developing a marine mammal vocalization detection system using deep learning that is capable of handling multi-species detection in the both time and frequency.

2 Data Collection and Processing

2.1 Acoustic recordings

A large collection of acoustic recordings were captured off the coast of Atlantic Canada during the summer and fall of 2015 and 2016. The recording locations were situated along an area of biological interest known as the Scotian Shelf. The recordings were sampled at 8kHz and 250kHz, however, we restrict our training data to the lower sampling rate as the majority of endangered baleen whale vocalizations fall well below 1000Hz. A small percentage of the acoustic recordings were analyzed by expert marine biologists and subsequently annotated to produce bounding boxes around example vocalizations of the three previously mentioned species of baleen whales. As marine biologists are often only concerned with presence/absence of specific species, the acoustic recordings were only partially labelled. For example, if two vocalizations fell within several seconds of one another, it is likely that only one was annotated, as illustrated in Figure 1.

2.2 Spectrograms of acoustic signals

The examples shown in Figure 1 each depict a visual representation of an acoustic signal known as a spectrogram. Roughly speaking the spectrogram representation of a signal x can be obtained as the square of the absolute-value of the Short-time Fourier Transform expressed below:

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-j\omega m}, \tag{1}$$

where time (n) is discrete, frequency (ω) is continuous, and w is a windowing function.

In much of the related literature–both traditionally using convolved templates and classification of hand-engineered features, as well as more recently using deep learning–the algorithms developed to detect marine mammal vocalizations operate in the frequency domain (i.e., using spectrograms). One reason for the ongoing use of spectrograms is that they are the primary resource used by marine biologists during annotation as they allow for fast analysis of signals inside and outside of the human hearing range.

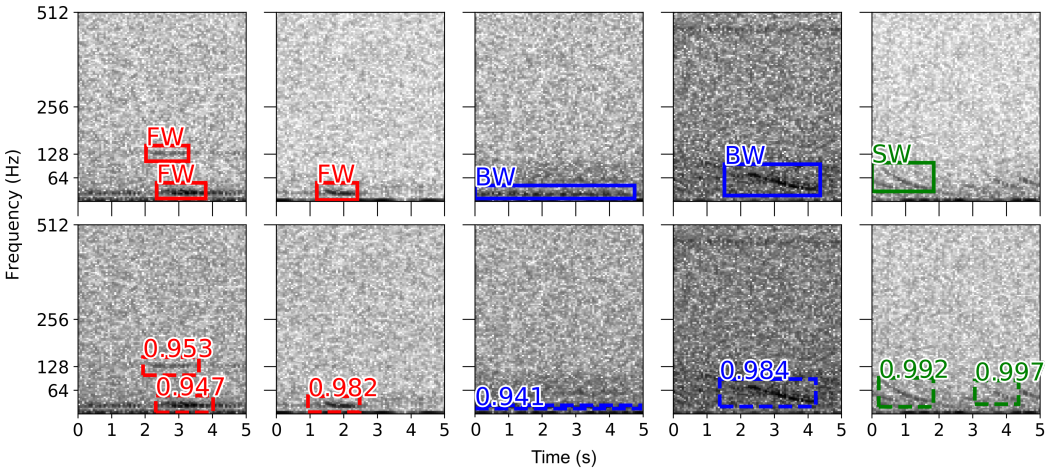

Figure 1: Example annotations (top row) and corresponding predictions made by the R-CNN (bottom row) for several example vocalizations produced by the three species of interest. As previously mentioned vocalizations have only been partially labelled, for example: in row 1 column 5, there appears to be several sei whale vocalizations occurring consecutively, however, only one has been annotated. Interestingly, the R-CNN has detected two vocalizations and therefore the reported metrics for this test instance would be artificially low.

2.3 Training data and experimental setup

Distinct data sets for training, validating, and testing the R-CNN were produced using a stratified sampling routine and a random split ratio of 70/15/15, respectively. A single training instance can be interpreted as a tensor with one channel corresponding to a spectrogram. In practice the spectrograms were not produced using Equation 1 directly, but rather a Fast Fourier Transform (FFT) algorithm. The FFT was used to produce a spectrogram of a signal five seconds in length using the Hann windowing function with a window length of 2048 samples and a window overlap of 512 samples. Due to the low-frequency nature of the vocalizations we are interested in, the spectrograms used during training were truncated using a maximum frequency of 512Hz.

We adopt the architecture proposed in [4] known as Mask R-CNN implemented in Python using the open-source deep learning framework PyTorch [10]. Specifically, we use ResNet-50 [5] for feature extraction coupled with a Feature Pyramid Network (FPN) [6] as a backbone. The 256 output features of the FPN are then handed to the standard Region Proposal Network (RPN) producing 1000 region of interest (RoI) proposals per training instance. The 1000 RoIs are then passed through the RoIAlign procedure and the head network composed of fully connected layers for classification and bounding box regression.

The R-CNN was trained for 100 epochs with early stopping being evaluated on the loss of the validation set. Four NVIDIA P100 Pascal GPUs each with 16GB of memory and a batch size of 4 were used for training. The initial learning rate of the stochastic gradient descent algorithm was set to 0.003 and decayed by a factor of 10 when the loss of the validation set stopped decreasing.

3 Preliminary results

The results detailed in this section depict the median of ten training runs using different random number generator seed values to produce distinct training, validation, and test data sets.

The R-CNN performs well when considering a low Intersect over Union (IoU) (e.g., AP/AR@.5) between the ground truth bounding boxes and the predictions, as reported in Table 1. The performance drops for IoU values larger than 0.7, which is reflected through the mAP/mAR@[.5:.95] columns of the same table. Notably, the R-CNN outperforms the current detection algorithms used in production at JASCO Applied Sciences for both fin and sei whales. These algorithms are implemented using a more traditional approach by first extracting candidate detections as contours within spectrograms [9]

and comparing specific features extracted from the contours against ground truth templates. A full comparison of the detection algorithms for blue whale vocalizations was not possible as the JASCO detectors require 30 second samples to detect blue whales and the R-CNN was trained using only five second samples.

Table 1: Median values of the average precision (AP) and average recall (AR) metrics evaluated over various IoU thresholds as described in the COCO Detection Challenge[†].

		AP@.5		**mAP@[.5:.95]**		**AR@.5**		**mAR@[.5:.95]**	
Species	**Label**	R-CNN	JASCO	R-CNN	JASCO	R-CNN	JASCO	R-CNN	JASCO
Overall	-	82.1	-	41.8	-	91.9	-	54.8	-
Blue whale	BW	85.7	-	52.8	-	96.2	-	70.9	-
Fin whale	FW	75.3	*65.0*	30.8	*27.4*	89.9	*62.6*	40.0	*35.0*
Sei whale	SW	85.4	*75.7*	41.9	*35.0*	89.7	*34.4*	49.4	*18.4*

4 Conclusion

The endangered baleen whales mentioned throughout this work are three of several cetaceans that make up the top of the aquatic food chain. Subsequently, their population levels play a major role in the lives of those lower in the food chain and the the long-term stability of the ocean's ecosystems [2]. Moreover, each of the species focused on in this work were primary targets during the commercial whaling industry. While, the populations of many species of cetaceans, including some pods of blue whales and fin whales, have been rising since commercial whaling was banned in the late 1980's, increased human activity at sea has presented another significant threat to the livelihood of these species. In order to reduce this threat governmental policy makers must continue to impose speed restrictions and temporarily suspend fishing activities in susceptible areas.

It is our belief that continuous monitoring of cetaceans and effective policy decisions are feasible through PAM and the R-CNN implementation outlined in this paper.

4.1 Future work

The work outlined in this paper is preliminary and part of an ongoing project focused on detecting the vocalizations of various species of marine mammals–both endangered and non-endangered–within acoustic recordings. More specifically, larger models which include data that was collected in various locations around the world allow for the inclusion of species not present along the coast of Atlantic Canada as well as the ability to interpret various sources of ambient noise (i.e., soundscapes).

Recent work has demonstrated that CNNs trained to detect whale vocalizations are able to generalize to new species using transfer learning [13]. We intend to exploit transfer learning in order to fine-tune the model described in this work to detect the vocalizations of the North Atlantic right whale (*Eubalaena glacialis*) for which there is very limited available training data due to extremely low population levels (< 400 individuals).

A major limitation of the current implementation is partially annotated data. Research into using semi and/or weakly-supervised learning methods is currently in progress. Another possible solution which is being considered is to employ active learning in an expert environment to correct partial labels.

Finally, further work with respect to model compression is in progress such that the network described above can be used in real-time.

Acknowledgements

The acoustic recordings described in this paper were collected by JASCO Applied Sciences under a contribution agreement with the Environmental Studies Research Fund.

[†]Details of the COCO Detection Challenge: `http://cocodataset.org/#detection-eval`

References

[1] Mark F Baumgartner and Sarah E Mussoline. A generalized baleen whale call detection and classification system. *The Journal of the Acoustical Society of America*, 129(5):2889–2902, 2011.

[2] Pieter A Folkens and Randall R Reeves. *Guide to marine mammals of the world*. National Audubon Society., 2002.

[3] Benjamin S Halpern, Shaun Walbridge, Kimberly A Selkoe, Carrie V Kappel, Fiorenza Micheli, Caterina D'agrosa, John F Bruno, Kenneth S Casey, Colin Ebert, Helen E Fox, et al. A global map of human impact on marine ecosystems. *Science*, 319(5865):948–952, 2008.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[7] Wenyu Luo, Wuyi Yang, and Yu Zhang. Convolutional neural network for detecting odontocete echolocation clicks. *The Journal of the Acoustical Society of America*, 145(1):EL7–EL12, 2019.

[8] David K Mellinger. A comparison of methods for detecting right whale calls. *Canadian Acoustics*, 32(2):55–65, 2004.

[9] David K Mellinger, Stephen W Martin, Ronald P Morrissey, Len Thomas, and James J Yosco. A method for detecting whistles, moans, and other frequency contour sounds. *The Journal of the Acoustical Society of America*, 129(6):4055–4061, 2011.

[10] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[11] John R Potter, David K Mellinger, and Christopher W Clark. Marine mammal call discrimination using artificial neural networks. *The Journal of the Acoustical society of America*, 96(3):1255–1262, 1994.

[12] Marie A Roch, Holger Klinck, Simone Baumann-Pickering, David K Mellinger, Simon Qui, Melissa S Soldevilla, and John A Hildebrand. Classification of echolocation clicks from odontocetes in the southern california bight. *The Journal of the Acoustical Society of America*, 129(1):467–475, 2011.

[13] Mark Thomas, Bruce Martin, Katie Kowarski, Briand Gaudet, and Stan Matwin. Marine mammal species classification using convolutional neural networks and a novel acoustic representation. *arXiv preprint arXiv:1907.13188*, 2019.

[14] Ildar R Urazghildiiev, Christopher W Clark, Timothy P Krein, and Susan E Parks. Detection and recognition of north atlantic right whale contact calls in the presence of ambient noise. *IEEE Journal of Oceanic Engineering*, 34(3):358–368, 2009.

[15] Angelia SM Vanderlaan and Christopher T Taggart. Vessel collisions with whales: the probability of lethal injury based on vessel speed. *Marine mammal science*, 23(1):144–156, 2007.

[16] Walter MX Zimmer. *Passive acoustic monitoring of cetaceans*. Cambridge University Press, 2011.