Technical Report: BioDCASE Baleen Whale Deep Learning Detection and Classification Network

Contributors:

- Gabriela Alongi[1]
- Liz Ferguson[1]
- Peter Sugarman[2]

[1] *Ocean Science Analytics LLC, San Diego, California, 92129, USA*
[2] *Acoustic Interactions, LLC, Bellevue, Washington, 98008, USA*

1) Tool Introduction

DeepAcoustics is a deep learning-based tool designed for the detection and classification of underwater acoustic signals, with a focus on marine mammal vocalizations. Developed with a modular architecture, DeepAcoustics allows users to configure, train, and apply custom neural network models for sound event detection. DeepAcoustics supports a range of deep learning architectures for acoustic detection and classification, including Tiny YOLO, ResNet, and Darknet-based networks, allowing users to select the model best suited for their specific data and computational needs.

This graphical user interface (GUI)-based tool was implemented in MATLAB (version R2024b, developed by MathWorks), leveraging the following MATLAB toolboxes: Computer Vision, Curve Fitting, Deep Learning, Image Processing, Parallel Computing, Signal Processing, and Statistics and Machine Learning. Over the past two years, DeepAcoustics has undergone continuous development and recently incorporated multiclass detection capabilities, all within a user-friendly framework that supports efficient spectrogram preprocessing, model training, testing, and deployment.

2) Tool Modifications Prior to Task

Participating in the BioDCASE challenge prompted several modifications to DeepAcoustics. Most changes were minor, including improving the utility of the call review dialog and adapting to the date/time format used by the files provided. However, one significant advancement was changing how validation data was implemented. Previously, DeepAcoustics would randomly subset the training data to be used as validation data (using a user-specified percentage). However, this approach biases the model and limits the generalizability of the final model since the validation data, coming from the same data pool, would be autocorrelated with and therefore extremely similar to the training data. Because the BioDCASE challenge provided validation data as already separated datasets, this gave

us the impetus to improve this functionality in DeepAcoustics and add the option for the user to select a pre-determined, separate pool of data to use for validation.

3) Network Details

For the baleen whale detection task, DeepAcoustics employed the Tiny YOLO network, a lightweight convolutional neural network consisting of 74 layers. Tiny YOLO offers a compact alternative to full-scale YOLO architectures, optimized for real-time detection without sacrificing significant accuracy. This network was selected due to its efficiency and proven performance in constrained computational environments.

The Tiny YOLO model was pre-trained on the COCO dataset, a comprehensive object detection dataset containing over 330,000 labeled images. Unlike datasets such as ImageNet that focus on bounding boxes, COCO provides object segmentation, which enhances the model's ability to recognize fine-grained structures in data. This pre-training served as the foundation for transfer learning, enabling the network to adapt effectively to spectrogram image inputs of baleen whale calls.

Training and validation images were generated from audio files using an image resolution of 288 × 288 pixels, with a duration of 10 seconds per image. This duration was chosen because 90% of calls in the dataset were under 8.8 seconds. Using a larger time window was avoided to minimize classification issues introduced by the high variability between blue and fin whale call durations. This decision reflects a trade-off between multi-class call representation and call separation fidelity. Spectrograms were generated with a window size and NFFT of 128, with an 80% overlap between windows. The resulting normalized power spectral densities were enhanced by applying a contrast-limited adaptive histogram equalization transform (CLAHE) using a Rayleigh distribution with each tile in the transform sized to about 50 Hz by 0.2 seconds, a contrast enhancement limit of 0.005, and a distribution parameter of 0.4.

The model architecture was based on the Tiny YOLO framework, which was adapted from previous dolphin whistle classification work (Sugarman et al., *In Press*). Training parameters were selected based on prior benchmarking and computational constraints, summarized below:

Parameter	Value
Image resolution	288 x 288 pixels
Maximum image duration	10 seconds
Window size & NFFT	128
Overlap	80%
Anchors	6 (based on previous dolphin whistle work)
Optimization algorithm	RMSprop
Learning rate	0.0005
Mini-batch size	16
Epochs	7
Training duration	39 hours

Hyperparameter selection for this training effort was guided by results Sugarman et al. (In Press), and a project involving neural network development for 40 Hz fin whale calls. These earlier efforts informed optimal choices for batch size, learning rate, and model architecture. Ideally, without time constraints, significant effort would have gone into tuning the hyperparameters to the call types present in this dataset, adding some data augmentation, and perhaps dividing the approach to accommodate the significant difference in character between, for example, the shorter D-calls and fin calls versus the longer B-calls.

Sugarman, P. C., Ferguson, E. L., Alongi, G. C., Schallert, J. P., & Lyn, H. (in press). Effects of network selection and acoustic environment on bounding-box object detection of delphinid whistles using a deep learning tool. Journal of the Acoustical Society of America.

4) Detection Output

The trained DeepAcoustics network was evaluated on two separate datasets. On the ddu2021 dataset, the network produced 2767 detections, while the kerguelen2020 dataset yielded 856 detections across multiple call classes. The model was evaluated on the validation data using the Performance Metrics tool in DeepAcoustics to get a more quantitative sense of performance, although since these were used to train the network, the results are likely biased. Even so, performance was uneven, with blue whale D-calls performing best (F-score = 0.5092 for the Casey 2017 dataset, 0.5150 for the Kerguelen 2014 dataset, and 0.6244 for the Kerguelen 2015 dataset, 2D IOU threshold = 0.3). This may suggest that the model would benefit from being divided and tuned to different call types, rather than attempting to detect seven call types with one model. In an ideal scenario we would incorporate this feedback into iterative training and model testing.

Due to the significant effort required to modify the DeepAcoustics tool to accommodate the bioCASE dataset structure and formatting, we were unable to perform a more comprehensive network training and evaluation with different frameworks. However, we are excited to observe how this lightweight network with limited training still demonstrated promising detection performance across evaluation datasets.

Enclosed with Submission Packet:

1. This report
2. Ferguson_OSA_task2_1.meta.yaml
3. Ferguson_OSA_task2_1.csv