DEEP LEARNING MODEL ADAPTATION FOR DETECTION OF WHALE CALLS

Daniela Rose Garcia Valencia

ABSTRACT

Whales produce some of the most distinctive and complex vocalizations in the marine environment. As highly vocal species, they rely on sound for critical behavior such as navigation, reproduction, and foraging. In that sense, Passive Acoustic Monitoring (PAM), which involves monitoring wildlife using acoustic sensors, has emerged as a promising technique for studying whale vocalizations. At the same time, deep learning, consisting of machine learning algorithms capable of learning features from data, enables the automated and accurate identification of species-specific vocalizations [7]. The combination of PAM and deep learning provides a powerful tool for ecology, conservation, and biodiversity monitoring.

This study focuses on the detection and classification of seven distinct call types from two whale species: the Antarctic blue whale (Balaenoptera musculus intermedia) and the fin whale (Balaenoptera physalus). We will analyze the AcousticTrends_BlueFinLibrary (ATBFL) dataset [5] using the Vision Transformer (ViT) pre-trained deep learning model [3], originally trained for image classification.

In this work, we aim to fine-tune the ViT pre-trained model for the classification and detection of whale calls by representing the audio as spectrograms, thereby studying their viability for application to whale conservation. We also compare our methods and results with those presented in the BioDCASE challenge [2], evaluating metrics like F-score, recall, and precision.

Keywords: Passive Acoustic Monitoring; Deep Learning; Whale Vocalization Classification; Antarctic Blue Whale; Fin Whale; Automated Detection.

1. INTRODUCTION

Antarctic blue whales and fin whales were severely affected by industrial whaling and remain vulnerable, making the monitoring of their vocalizations crucial for conservation. Automated methods for detecting and classifying whale calls are therefore relevant for supporting their conservation efforts.

To address this, several approaches have been proposed in previous works for whale-call detection. CNN-based models combined with BiLSTM layers have been applied within voice activity detection frameworks [13], while domain-adapted frameworks such as Voxaboxen [12] and modified YOLOv1 models with enhanced spectrogram processing [11] have improved detection performance in BioDCASE 2025 Task 2 [2]. Despite these advances, detecting and classifying whale calls remains a challenging task.

To overcome this, we adapt a transformer-based model [3] and fine-tune it using LoRA [4], which reduces computational cost.

2. METHODS

Dataset

This project utilizes the AcousticTrends_BlueFinLibrary [5], which contains 6,591 audio files totaling 1,880 hours of recordings, sampled at 250 Hz. It is organized into 11 site-year deployments across Antarctica from 2005 to 2017. The training set includes 6,007 files from eight site-years, while 587 files from Kerguelen 2014, Kerguelen 2015, and Casey 2017 are reserved for validation.

Site-year	Vocalization categories						
	A	B	Z	D	20Hz	20Plus	Dswp
Maud Rise 2014	2191	37	28	70	23	5	6
Greenwich 2015	827	157	29	66	2	1	46
Kerguelen 2005	812	237	166	435	788	78	444
Kerguelen 2014	2557	1177	563	435	1920	1826	344
Kerguelen 2015	1970	542	236	1180	552	718	344
Casey 2014	3681	1398	1091	679	17	0	0
Casey 2017	1741	558	119	553	78	214	0
Ross Sea 2014	104	0	0	0	0	0	0
Balleny Islands 2015	923	44	31	46	951	148	78
Elephant Island 2013	2447	1672	141	10 600	3266	1599	965
Elephant Island 2014	6934	967	100	1034	4940	2912	4077
Total	24 189	6791	2506	15 100	12 539	7503	6306

Table 1. Dataset composition

Whales' vocalizations are categorized into seven call types as follows:

Antarctic blue whale:

- Z-Call (BmZ): Three-part call (A, B, C) with a smooth transition from 27 to 16 Hz.
- A-Call (BmA): Only the A part of the Z-Call.
- B-Call (BmB): A and B parts of the Z-Call.
- D-Call (BmD): Downsweeping component from 20–120 Hz; may include additional modulations.
- 20 Hz Pulse without overtone (Bp20): Downsweep from 30 to 15 Hz.
- 20 Hz Pulse with overtone (Bp20plus): 20 Hz pulse with secondary energy at 80–120 Hz.
- 40 Hz downsweep (BpD): Downsweep ending around 40 Hz, usually 30–90 Hz.

Model

This work employs the Vision Transformer (ViT) model [3], which splits images into fixed-size patches and treats them as a sequence of tokens. ViT was pre-trained on ImageNet [14], and its capacity to capture complex visual patterns could be advantageous when applied to spectrograms.

Moreover, audio was converted into 15s spectrograms (sampled at 250 Hz, bandpass 5–124 Hz) and split into fixed-size patches. Furthermore, Low-Rank Adaptation (LoRA) [4] will be used to fine-tune the ViT model. LoRA adds lightweight trainable components, meaning it introduces a small number of additional parameters rather than retraining all the weights of the original model, facilitating its adaptation to specific tasks such as whale vocalization classification.

LoRA consists of two low-rank matrices A $\in R^{hxr}$ and B $\in R^{rxh}$, with r $\ll$ min(r, h) being a user-defined rank parameter. During fine-tuning, the pre-trained weights W_0 remain frozen while the LoRA weights (A and B) are updated, thereby significantly reducing computational cost. Consequently, the modified forward pass for layer l becomes:

$$X^{(l+1)} = W_0 x^{(l)} + \mathrm{AB} x^{(l)}$$

Metrics

We follow standard classification metrics (precision, recall, and F1-score) to evaluate model performance, focusing on both correctly detecting whale calls and minimizing misclassifications caused by noise.

3. RESULTS

Previous work established a baseline using ResNet18 and YOLOv11 [6], with results summarized in Table 2.

Overall, the results show an improvement when using our method for the detection and classification of whale calls. ViT model, fine-tuned with LoRA at rank r=4, achieves the highest recall (0.552), meaning it misses fewer real calls, and it maintains a solid precision (0.477) as well as the highest F1-score (0.496) of the three models.

Compared to the ResNet18 baseline, the ViT model captures whale calls more reliably, with a noticeable gain in recall and a more balanced precision–recall trade-off.

On the other hand, YOLOv1 shows high precision but misses many true calls, making it less suitable when comprehensive detection is needed. In contrast, ViT maintains both sensitivity and precision, leading to the best general performance.

Model	Recall	Precision	F1-score
YOLOv11	0.32	0.67	0.43
ResNet18	0.36	0.29	0.32
ViT	0.552	0.477	0.496

Table 2: Comparison of the results obtained with Vit model and the baseline. Scores are average across all call types and validation sets.

Table 3 compares the ViT results with other approaches from BioDCASE Task 2. The ViT model outperforms MFCCs+BiLSTM, HuBERT+BiLSTM, and AVA-VAD [13] across all metrics, achieving the highest recall, precision, and F1-score. This highlights the advantage of adapting a pretrained ViT with LoRA for this task.

Model	Recall	Precision	F1-score
MFCCs+BiLSTM [13]	0.409	0.226	0.291
HuBERT+BiLSTM [13]	0.064	0.104	0.079
AVA-VAD [13]	0.310	0.219	0.245
ViT	0.552	0.477	0.496

Table 3: Comparison of the ViT model with other BioDCASE Task 2 approaches. Scores are averaged across all call types and validation sets.

4. CONCLUSION

We successfully adapted Vit with Lora for BioDCASE 2025 Task 2. The results show that the Vision Transformer (ViT) outperforms the ResNet18 and YOLOv11 baselines as well as MFCCs+BiLSTM, HuBERT+BiLSTM, and AVA-VAD models. ViT achieves the highest recall and F1-score while maintaining balanced precision. These findings suggest the ViT model, combined with parameter-efficient fine-tuning using LoRA, provides an effective and reliable approach for whale-call detection and classification.

5. REFERENCES

[1] Castro, et al. (2024). Beyond counting calls: estimating detection probability for Antarctic blue whales reveals biological trends in seasonal calling. Front. Mar. Sci. doi:10.3389/fmars.2024.1406678

[2] Biodcase 2025 task 2: Supervised detection of strongly-labelled Antarctic blue and fin whale calls https://biodcase.github.io/challenge2025/task2, accessed December 2025.

[3] Dosovitskiy, et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[4] LoRA (Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.)

[5] Miller, et al. (2020). An annotated library of underwater acoustic recordings for testing and training automated algorithms for detecting Antarctic blue and fin whale sounds. doi: 10.26179/5e6056035c01b

[6] Miller, et al. (2021). An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors. Sci. Rep., 11, 806. doi:10.1038/s41598-020-78995-8

[7] Schall, et al. (2024). Deep learning in marine bioacoustics: a benchmark for baleen whale detection. Remote Sens Ecol Conserv, 10: 642-654. https://doi.org/10.1002/rse2.392

[8] Dubus, et al. (2024). Improving automatic detection with supervised contrastive learning: application with low-frequency vocalizations. Workshop DCLDE (2024)

[9] Fonseca, et al. (2022). FSD50K: An Open Dataset of Human-Labeled Sound Events. IEEE/ACM TALP, vol. 30(1)

[10] Schall, Parcerisas (2022). A Robust Method to Automatically Detect Fin Whale Acoustic Presence in Large and Diverse Passive Acoustic Datasets. J. Mar. Sci. Eng., 10(12), 1831. https://doi.org/10.3390/jmse10121831

[11] Amin, R., Yen, B., Ashizawa, T., & Nakadai, K. (2025). Enhanced Spectrogram Processing with Temporal Sequences for Antarctic Whale Call Detection Using YOLOv1. Technical Report

[12] Deep Voice Below the Surface: Improved Whale Call Detection via Voxaboxen Refinement (2025). BioDCASE 2025 Task 2 Technical Report

[13] Geldenhuys, C. M., Tonitz, G., & Niesler, T. R. (2025). Whale-VAD: Whale Vocalisation Activity Detection. DCASE2025 Challenge, Technical Report

[14] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.