Deep Learning Based Detection and Classification of Marine Mammals

Bora Aydin [1,2]

[1]University of Toulon [2]Norwegian University of Science and Technology - NTNU

Co-funded by the Erasmus+ Programme of the European Union

Introdcution

Marine mammals play a crucial role in marine ecosystems and contribute to the balance and health of the oceans. Changes in their populations and migration patterns may indicate shifts in their environment, including variations in food availability, water temperature, or the presence of pollutants. Bioacoustic research investigates underwater soundscapes to identify different animal species. The result of these investigations allows to develop strategies to preserve the biodiversity of marine ecosystems and to protect endangered species.

The overall goal of this thesis is to develop a deep learning-based system for detecting and classifying typical vocalizations of marine mammals based on recordings of passive acoustic monitoring (PAM) devices like hydrophones.

Objectives

1. Conduct a literature review on marine mammal detection and classification.
2. Review and analyze available research datasets.
3. Preprocess datasets for model training.
4. Study state-of-the-art sound event detection models.
5. Explore data augmentation methods.
6. Test transfer-learning approaches.
7. Document results and compile the thesis.

Methods

Data Collection

- Watkins Marine Mammal Sound Database [1]:

Collection of Marine Mammal Sound Recordings consists of recordings of various marine mammal species collected over seven decades

- BEANS: The Benchmark of Animal Sounds [3]:

A collection of bioacoustics tasks and public datasets, specifically designed to measure the performance of machine learning algorithms in the field of bioacoustics. The benchmark consists of two tasks in bioacoustics: classification and detection.

Preprocessing

- Noise Reduction:

Methods to eliminate or reduce background noise in audio recordings, such as spectral gating, adaptive filtering, and noise profiling.

- De-noising using Hydrophone Arrays:

Utilizing multiple hydrophone arrays to spatially filter out noise and enhance signal-to-noise ratio through beamforming techniques.

Models

- Slow-Fast Auditory Streams [4]:

The architecture fuses the two streams at multiple representation levels, inspired by the human auditory system, which includes a slow and a fast stream to process different aspects of sound. It enhances the model's ability to recognize a wide range of audio activities, from momentary sounds to repetitive actions.

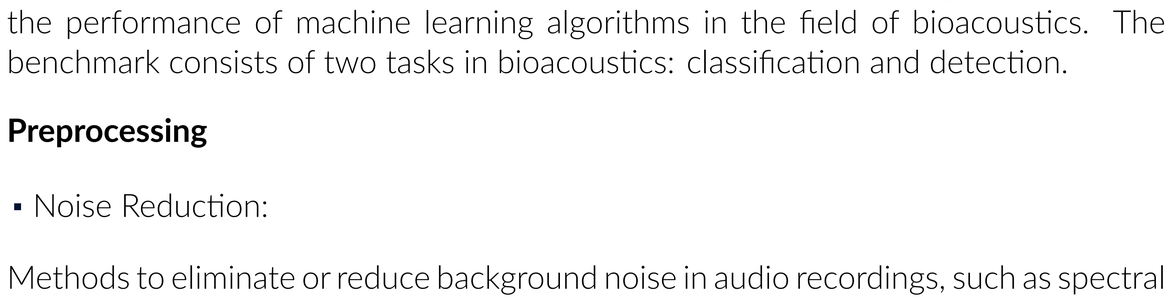

Figure 1. Slow-Fast Auditory Streams Model Architecture

The Slow stream captures high channel capacity and operates at a low sampling rate, focusing on frequency semantics. The Fast stream, with a higher temporal resolution, captures temporal patterns

- Audio Spectrogram Transformer(AST) [2]:

The paper introduces the Audio Spectrogram Transformer (AST), a novel model for audio classification that is purely based on attention mechanisms, without using convolutional neural networks (CNNs). The AST model directly processes audio spectrograms to classify audio events, achieving state-of-the-art results on various benchmarks.

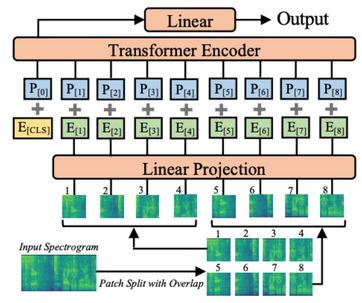

Figure 2. Audio Spectrogram Transformer Model Architecture

Data Augmentation

- MixUp

Mixup is a data augmentation technique that generates a weighted combination of random image pairs from the training data. Given two images and their ground truth labels: $(x_i, y_i), (x_j, y_j)$, a synthetic training example $(\hat{x}, \hat{y})$ is generated as:

$$\hat{x} = \lambda x_i + (1 - \lambda) x_j \quad \text{and} \quad \hat{y} = \lambda y_i + (1 - \lambda) y_j$$

where $\lambda \sim \text{Beta}(\alpha = 0.2)$ is independently sampled for each augmented example.

- SpecAugment

SpecAugment is a data augmentation technique specifically designed for speech data, particularly for training neural networks on automatic speech recognition (ASR) tasks. It involves modifying the input spectrogram of audio data to make the neural network more robust to variations and improve its generalization capability.

Transfer Learning

- Utilizing pre-trained models on AudioSet [5] and ImageNet datasets.

- AudioSet dataset

AudioSet consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. The ontology is specified as a hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds.

- ImageNet dataset

The ImageNet project is a large visual database designed for use in visual object recognition software research. More than 14 million images have been hand-annotated by the project to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided. ImageNet contains more than 20,000 categories, with a typical category, such as "balloon" or "strawberry", consisting of several hundred images.

Expected Outcomes

- Development of a robust deep learning model for detecting and classifying marine mammal vocalizations.
- Improved understanding of marine mammal behaviors and environmental impacts.
- Contribution to bioacoustic research and conservation efforts.

References

[1] Watkins Marine Mammal Sound Database.
Watkins marine mammal sound database, n.d.
Accessed: 2024-06-01.

[2] Yuan Gong, Yu-An Chung, and James Glass.
Ast: Audio spectrogram transformer.
In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 571–575, Brno, Czech Republic, 2021.

[3] Yuan Gong, Yu-An Chung, and James Glass.
Beans: The benchmark of animal sounds.
2022.

[4] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen.
Slow-fast auditory streams for audio recognition.
In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859, Toronto, ON, Canada, 2021.

[5] Google Research.
Audioset: An ontology and human-labeled dataset for audio events, n.d.
Accessed: 2024-06-01.