UNDERWATER SPECIES CLASSIFICATION USING TRANSFORMER MODEL

Pratyaksh Raj, 244156005 under the supervision of Dr. Sonali Chauhan

CICPS, IIT Guwahati

Abstract

Monitoring marine ecosystems requires reliable identification of underwater species from complex acoustic signals. Recent models utilizing RNN / LSTM are prone to vanishing gradient problem for long audio signals, also they process input sequentially therefore slow.

Here we used Watkins marine mammals sound dataset. The audio recordings were preprocessed into two Spectro-temporal representations—Mel-spectrograms and MFCCs. The proposed model integrates a lightweight CNN-4 backbone to capture localized time–frequency patterns and then cross attention + encoder to model global temporal relationships. For a balanced subset of five classes, the system achieves approximately 96% accuracy, demonstrating strong generalization under noisy underwater conditions

The system demonstrates the effectiveness of Spectro-temporal features for species-level classification and real-time edge deployment due to fast inference by utilizing parallelizability of sequential data and learning long-range dependencies using CNN and attention mechanism.

Introduction

Here we used Watkins marine mammals sound dataset. it consist of 15000 audio files of 55 species of whales, dolphins and seals. The audio recordings were preprocessed into two Spectro-temporal representations—Mel-spectrograms and 36 MFCCs. The proposed model integrates a lightweight CNN-4 backbone and cross attention + encoder.

The dataset is highly imbalanced. For full dataset validation accuracy is 89% but F1 score of 0.7. Therefore, for testing and comparison purposes we tested it on 5 classes with almost equal size, showing 96.3% accuracy.

Preprocessing

- Resampling to a fixed sampling rate,16 kHz
- Mel-spectrogram generation, using 128 Mel bands
- MFCC extraction, comprising of 36 coefficients

Theoretical Aspects / Key feature of this model

Marine mammal calls often contain:

- narrow-band harmonic structures
- broadband impulsive clicks
- frequency sweeps

A CNN over the spectrogram extract these local Spectro-temporal motifs, such as:

- rising/falling pitch contours
- peak energy bands

However, CNN features alone may still be sensitive to background noise and overlapping calls

MFCCs are more robust to additive noise because:

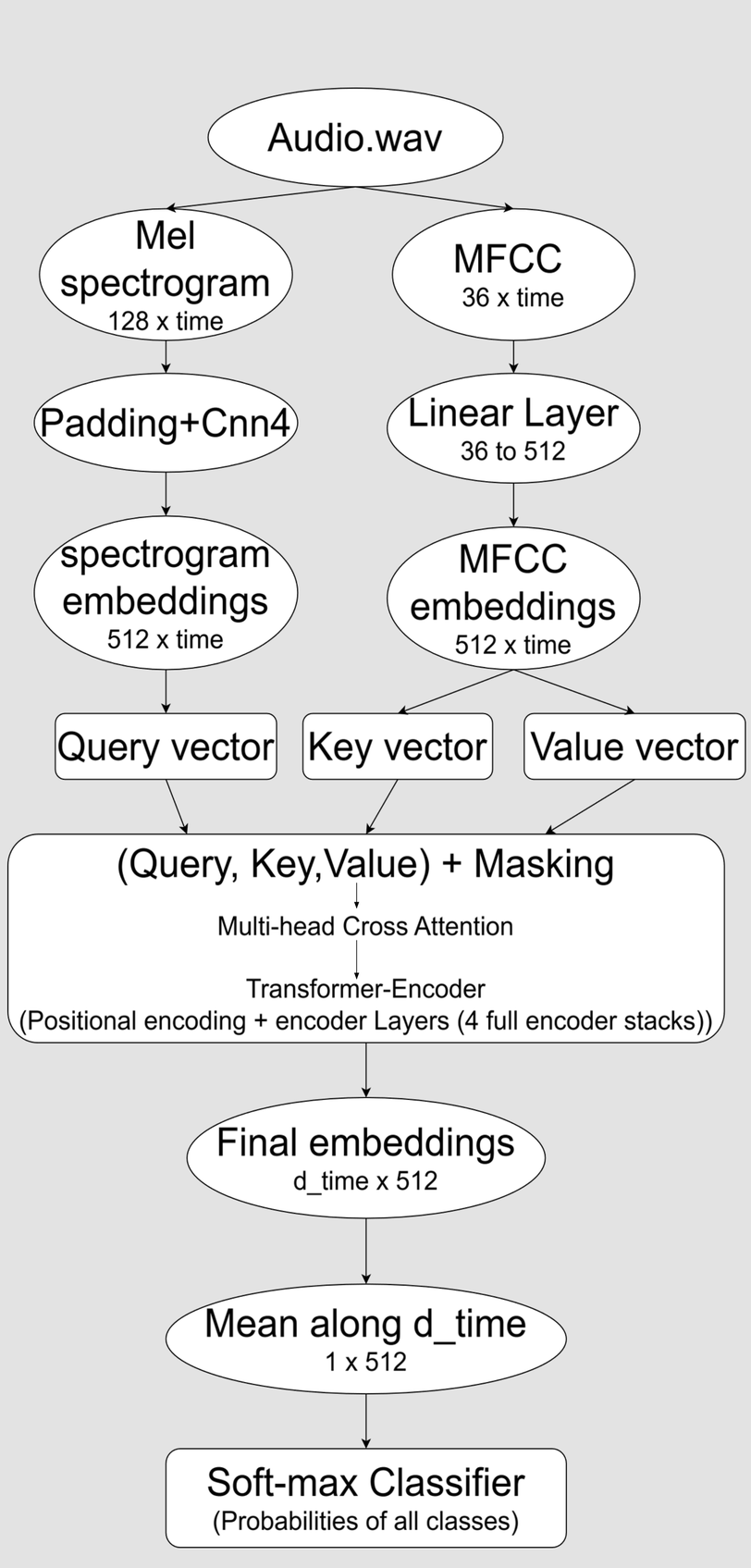
- they capture the shape of the spectral envelope
- they compress frequency information into smooth, lower-dimensional cepstral coefficients
- random noise tends to affect high-frequency coefficients but not the lower cepstral structure

Thus, MFCC provides a stable, noise-invariant representation complementary to spectrogram features.

Cross-attention performs feature-level alignment, ensuring that:

- Important spectrogram cues are *validated* by corresponding MFCC cues
- Noisy spectrogram regions receive low attention weights, reducing their influence
- Harmonically-relevant MFCC coefficients strengthen the network's confidence

Methodology

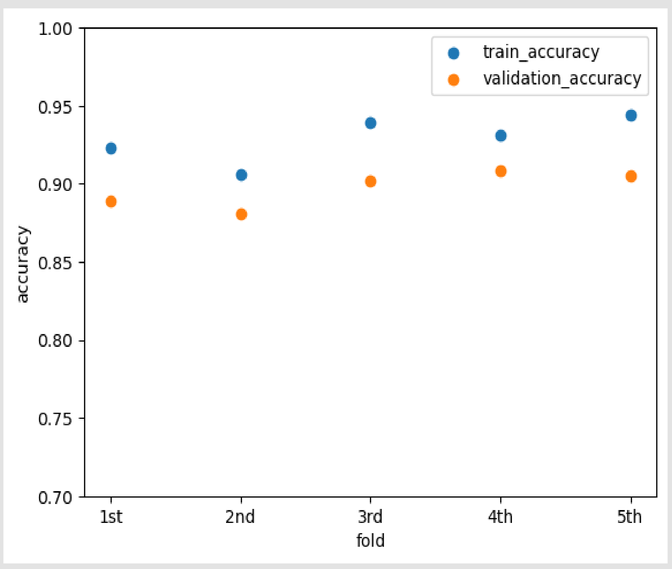


Trained for 25 epochs.

Results and Conclusion

Below fig shows the train and validation accuracy for 55 classes, Showing train and validation accuracy of ~ 92% and 89% on an average, on 5-fold CV. Average Macro F1 score ~0.7 on validation set, low due to imbalanced dataset.

Also, For a test dataset consisting of 355 audios of 5 different classes consisting of two species of dolphins, two of whales and one of seal (balanced dataset), the test accuracy is 96.3% and macro f1 score of 0.958.



Therefore, from above results we can conclude that our model is performing very well, without overfitting. Also, the training and inference time is quite fast compared to traditional techniques.

References

1. B Mishachandar, S Vairamuthu, An underwater cognitive acoustic network strategy for efficient spectrum utilization, Applied Acoustics, Volume 175, 2021, 107861, ISSN 0003-682X, https://doi.org/10.1016/j.apacoust.2020.107861
2. Muhammad Azeem Aslam, Lefang Zhang, Xin Liu, Muhammad Irfan, Yimei Xu, Na Li, Ping Zhang, Zheng Jiangbin, Li Yaan, Underwater sound classification using learning based methods: A review, Expert Systems with Applications, Volume 255, Part A, 2024, 124498, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2024.124498.
3. arXiv:1706.03762 [cs.CL] https://doi.org/10.48550/arXiv.1706.03762
4. Classification of Underwater Broadband Bio-acoustics Using Spectro-Temporal Features Navinda Kottege Autonomous Systems Laboratory CSIRO ICT Centre Pullenvale QLD 4069, Australia navinda.kottege@csiro.au Raja Jurdak Autonomous Systems Laboratory CSIRO ICT Centre Pullenvale QLD 4069, Australia raja.jurdak@csiro.au
5. Miao, Yongchun, Zakharov, Yury , Sun, Haixin et al. (2 more authors) (Accepted: 2020) Underwater acoustic signal classification based on sparse time-frequency representation and deep learning. IEEE Journal of Oceanic Engineering. pp. 1-14. ISSN: 0364-9059 (In Press)