Biodiversity Classifier

Miguel Lopez, Mario Hernandez

July 2024

Abstract

This paper presents a comprehensive study on the classification of bird calls from Puerto Rican bird species using advanced machine learning techniques. We explore the efficacy of Convolutional Neural Networks (CNNs), attention mechanisms, and the Wav2Vec model in capturing the intricate acoustic patterns in bird calls. The performance of these models is evaluated using metrics such as precision, recall, and the Simpson Biodiversity Index to quantify the diversity captured by the classifier. Our results indicate a high accuracy in identifying bird species, demonstrating the potential of these techniques in bioacoustic monitoring and biodiversity assessment.

1 Introduction

Bird call classification is a vital tool in the field of biodiversity monitoring, allowing researchers to track and study avian populations and their behaviors. With the growing concern over the impact of environmental changes on wildlife, accurate and efficient methods for identifying bird species through their calls have become increasingly important. Puerto Rico, known for its rich avian diversity, presents a unique opportunity to develop and test these methods.

This study aims to build a robust classifier [1] for identifying Puerto Rican bird species from their calls. Leveraging advancements in machine learning, specifically Convolutional Neural Networks (CNNs) and the Wav2Vec model, we aim to capture the nuanced acoustic features of bird calls. Additionally, we incorporate attention mechanisms to enhance the model's ability to focus on critical segments of the audio data.

We evaluate the performance of our classifier using standard metrics such as precision, recall, and F1-score, and introduce the Simpson Biodiversity Index as a measure of the model's capability to reflect the ecological diversity present in the dataset. By analyzing the results, we aim to provide insights into the strengths and limitations of these approaches in real-world biodiversity assessment scenarios.

[1] All code can be found at: `https://github.com/MiguelALopez1/BirdCallClassifier`

2 Related Works

In the development of bird call classification models, several influential works have laid the foundation for both feature extraction and model architectures. Baevski et al. [1] introduced wav2vec 2.0, a framework for self-supervised learning of speech representations. This model has shown significant potential in various audio classification tasks due to its ability to learn high-quality features from raw audio waveforms without extensive labeled data.

The application of attention mechanisms in sequence modeling, as demonstrated by Vaswani et al. [2], further enhances the performance of classification models. The "Attention is all you need" paper presents a transformer model that has revolutionized natural language processing and has been adapted for use in audio processing tasks, providing state-of-the-art results in several domains.

Building on these advancements, Devlin et al. [3] developed BERT, a deep bidirectional transformer model pre-trained for language understanding. Although primarily focused on text, the underlying principles of BERT's pre-training and fine-tuning have inspired similar approaches in the audio domain, particularly in leveraging large-scale unsupervised pre-training followed by supervised fine-tuning.

Moreover, Mitrovic et al. [4] explored unsupervised models for biodiversity assessment, highlighting the importance of robust feature extraction and clustering techniques in analyzing bioacoustic data. Their investigation into various unsupervised methods provides valuable insights into the potential of these techniques for improving the accuracy and efficiency of bird call classification systems.

3 Background

3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep neural networks primarily used for analyzing visual imagery. CNNs consist of multiple layers including convolutional layers, pooling layers, and fully connected layers.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

where I is the input image, K is the kernel, i and j are indices that denote the position in the output feature map S. The convolution operation slides the kernel across the input image and computes the dot product at each position.

3.2 Attention

Attention mechanisms in neural networks enable the model to focus on important parts of the input sequence, significantly improving the performance in

tasks such as machine translation and speech recognition. The attention mechanism computes a weighted sum of values (V), where the weights (a) are derived from a compatibility function of the query (Q) with the corresponding key (K).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key vectors.

3.3 Fourier Transform

The Fourier Transform decomposes a function (signal) into its constituent frequencies. It transforms a time-domain signal into a frequency-domain representation.

$$\mathcal{F}\{f(t)\} = F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt$$

where $f(t)$ is the time-domain signal, $F(\omega)$ is the frequency-domain representation, and ω is the angular frequency.

3.4 Fast Fourier Transform

The Fast Fourier Transform (FFT) is an efficient algorithm for computing the Discrete Fourier Transform (DFT) and its inverse. The DFT transforms a sequence of complex numbers $x_0, x_1, \ldots, x_{N-1}$ into another sequence of complex numbers $X_0, X_1, \ldots, X_{N-1}$, representing the frequency components of the original sequence.

3.4.1 Discrete Fourier Transform

The Discrete Fourier Transform is defined by the formula:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i\frac{2\pi}{N}kn}$$

for $k = 0, 1, \ldots, N-1$, where: - x_n is the n-th sample of the input sequence, - X_k is the k-th sample of the output sequence, - N is the total number of samples, - i is the imaginary unit, and - $e^{-i\frac{2\pi}{N}kn}$ represents the complex exponential basis function.

The DFT can be interpreted as a decomposition of the input sequence into a sum of sinusoidal components with different frequencies and amplitudes.

3.4.2 Fast Fourier Transform

The Fast Fourier Transform is an algorithm that reduces the computational complexity of calculating the DFT from $O(N^2)$ to $O(N \log N)$. This efficiency is achieved by recursively breaking down the DFT of a sequence of length N into

smaller DFTs of subsequences. The most common FFT algorithm is the Cooley-Tukey algorithm, which exploits the symmetry and periodicity properties of the complex exponential basis functions.

The FFT and its inverse are defined by the following algorithms:

$$\text{FFT}(x)_k = \sum_{n=0}^{N-1} x_n e^{-i\frac{2\pi}{N}kn}$$

$$\text{IFFT}(X)_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{i\frac{2\pi}{N}kn}$$

for $k, n = 0, 1, \ldots, N-1$.

The FFT is widely used in various applications, including signal processing, image processing, and solving partial differential equations, due to its ability to efficiently transform data between the time (or spatial) domain and the frequency domain.

3.5 MEL Spectrograms

A MEL spectrogram is a representation of the power spectrum of a sound signal, using a MEL scale for the frequency axis which better represents human auditory perception.

$$MFCC(n) = \sum_{m=1}^{M} \log(S_m) \cos\left[n\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right]$$

where S_m are the powers at each frequency bin in a Mel spectrogram, M is the number of Mel filters, and n is the number of MFCCs.

3.6 Simpson Biodiversity Index

The Simpson Biodiversity Index is a measure of diversity which considers both species richness and evenness. It is defined as:

$$D = 1 - \sum_{i=1}^{S}\left(\frac{n_i(n_i-1)}{N(N-1)}\right)$$

where S is the total number of species, n_i is the number of individuals of species i, and N is the total number of individuals.

4 Audio Processing

The audio processing pipeline in this study involves several key steps: loading audio files, preprocessing, feature extraction, and classification. Each step is mathematically formulated to ensure the accurate and efficient analysis of bird call recordings.

4.1 Loading and Preprocessing Audio Files

The audio files are first loaded and preprocessed to ensure consistency and quality of the input data.

4.1.1 Loading Audio Files

Audio files are loaded using the `librosa` library, which reads the audio signal $x(t)$ and converts it into a digital format. The audio signal is represented as a discrete sequence of amplitude values:

$$x[n] = x(t)\big|_{t=nT_s}$$

where T_s is the sampling period, and n is the sample index.

4.1.2 Preprocessing

The audio signal is preprocessed by normalizing and resampling. Normalization scales the signal to have zero mean and unit variance:

$$\hat{x}[n] = \frac{x[n] - \mu_x}{\sigma_x}$$

where μ_x is the mean and σ_x is the standard deviation of the signal. Resampling changes the sampling rate to a target rate f_s, ensuring consistency across all audio files.

4.2 Feature Extraction

Feature extraction transforms the raw audio signal into a set of features that capture the important characteristics of the bird calls.

4.2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are widely used in audio processing as they represent the short-term power spectrum of the sound signal on a non-linear mel scale of frequency. The computation involves several steps:

1. **Short-Time Fourier Transform (STFT)**:

$$X_m[k] = \sum_{n=0}^{N-1} x[n + mH] \cdot w[n] \cdot e^{-i\frac{2\pi}{N}kn}$$

where $w[n]$ is the window function, H is the hop size, and m is the frame index.

2. **Mel-Filterbank**:

$$M[f] = \sum_{k=0}^{N/2} |X_m[k]|^2 \cdot h_{m,k}$$

where $h_{m,k}$ represents the mel-filter weights.

3. **Logarithm of Mel-Spectrum**:

$$\log M[f]$$

4. **Discrete Cosine Transform (DCT)**:

$$\text{MFCC}[m] = \sum_{n=0}^{N_m-1} \log M[n] \cdot \cos\left(\frac{\pi m(n+0.5)}{N_m}\right)$$

where N_m is the number of mel coefficients.

4.2.2 Delta and Delta-Delta Coefficients

Delta and delta-delta coefficients capture the temporal dynamics of the MFCC features. They are calculated as the first and second-order differences of the MFCCs:

$$\Delta\text{MFCC}[t] = \text{MFCC}[t+1] - \text{MFCC}[t-1]$$

$$\Delta^2\text{MFCC}[t] = \Delta\text{MFCC}[t+1] - \Delta\text{MFCC}[t-1]$$

5 Model Architecture

5.1 Wav2Vec

The Wav2Vec model is a state-of-the-art framework for self-supervised learning of speech representations, as detailed in the paper by Baevski et al. (2020) [1]. This model is designed to learn high-quality features from raw audio waveforms without the need for extensive labeled data. Raw audio waveforms, represented as $\mathbf{x}$, are continuous signals that capture variations in air pressure caused by sound waves. In digital form, these waveforms are typically sampled at a specific rate (e.g., 16 kHz or 44.1 kHz) and quantized into discrete amplitude values. Each point in $\mathbf{x}$ corresponds to the amplitude of the sound wave at a given time step. This sequence of amplitude values serves as the input to the Wav2Vec model, which processes these raw signals to extract meaningful representations. The specific model used in this study is `"dima806/bird_sounds_classification"`, which is a fine-tuned version of the `"facebook/wav2vec2-base-960h"` model tailored for bird sound classification.

5.1.1 Input Audio Preprocessing

The raw audio waveform $\mathbf{x}$ is first preprocessed by normalizing and resampling to a target sampling rate. The preprocessed audio is then fed into the feature extractor:

$$\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$$

5.1.2 Feature Extraction

The Wav2Vec model employs a convolutional feature encoder to process the raw audio waveform into latent speech representations. This involves several convolutional layers that transform the audio signal into a sequence of latent representations:

$$\mathbf{z} = f(\mathbf{x}) = \{z_1, z_2, \ldots, z_L\}$$

where f is the feature extraction function and L is the length of the output sequence. Here, $\mathbf{z}$ represents the latent speech representations. Mathematically, the convolution operation can be represented as:

$$z_t = \sum_{i=0}^{k-1} x_{t+i} \cdot w_i$$

where w_i are the weights of the convolution filter and k is the kernel size.

5.1.3 Quantization Module

The latent representations $\mathbf{z}$ are quantized into $\mathbf{q}$ using product quantization, where $\mathbf{z}$ is mapped to one of the entries in the codebook. The Gumbel softmax is used for differentiable quantization to generate a probability distribution over the codebook entries:

$$p_{g,v} = \frac{\exp((l_{g,v} + n_v)/\tau)}{\sum_{k=1}^{V} \exp((l_{g,k} + n_k)/\tau)}$$

where: - $l_{g,v}$ are the logits (raw, unnormalized predictions of the previous feature extraction/convolutional layer) for the v-th entry in the g-th group. - n_v are the Gumbel noise samples, introducing randomness for exploration. - τ is the temperature parameter, controlling the smoothness of the output distribution.

The Gumbel softmax enables differentiable sampling, allowing gradients to flow through the quantization process during backpropagation. The probabilities $p_{g,v}$ are used to select the entries in each codebook, resulting in the quantized representations $\mathbf{q}$.

The quantized vector $\mathbf{q}$ is obtained by selecting the entries in each group according to the probabilities $p_{g,v}$. This discrete selection is made differentiable by the Gumbel softmax trick, allowing gradients to flow through the quantization process during backpropagation.

5.1.4 Transformer Network

The quantized representations $\mathbf{q}$ are fed into a transformer network to produce context representations $\mathbf{c}$:

$$\mathbf{c} = \text{Transformer}(\mathbf{q})$$

The self-attention mechanism in the transformer is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are the query, key, and value matrices derived from $\mathbf{q}$, and d_k is the dimension of the key vectors. Here, $\mathbf{c}$ represents the context representations. These context representations encapsulate both the local and global structure of the input sequence, as the self-attention mechanism allows each position in the output to attend to all positions in the input.

5.1.5 Contrastive Loss

The self-supervised training objective uses contrastive loss, which maximizes the similarity between the correct future latent representation and its prediction while minimizing the similarity to negative samples:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(\mathbf{c}_t, \mathbf{z}_t^+))}{\sum_{z \in \mathcal{N}} \exp(\text{sim}(\mathbf{c}_t, z))}$$

where $\text{sim}(\cdot)$ is a similarity measure (e.g., dot product), $\mathbf{z}_t^+$ is the positive sample, and $\mathcal{N}$ is the set of negative samples. The loss function trains the model to distinguish the true future latent representation from negative samples, refining the model's ability to represent audio data effectively.

5.1.6 Implementation

The model implementation for bird sound classification uses the Hugging Face 'transformers' library. The following code snippet illustrates the creation of the feature extractor and the audio classification model:

```
1
2     # "facebook/wav2vec2-base-960h"
3     model_str = "dima806/bird_sounds_classification"
4
5     # Create an instance of the feature extractor for audio.
6     feature_extractor = AutoFeatureExtractor.from_pretrained
          (model_str)
7
8     # Create an instance of the audio classification model.
9     # The 'num_labels' parameter is set to the number of
          labels in your 'labels_list'.
10    model = AutoModelForAudioClassification.from_pretrained(
          model_str, num_labels=len(labels_list))
```

Listing 1: Code for Creating the Feature Extractor and Audio Classification Model

Predicted label

Classification report:

	precision	recall	f1-score	support
AmericanOystercatcher	0.9375	0.9677	0.9524	31
Bananaquit	0.9770	0.9808	0.9789	260
NorthernMockingbird	0.9877	0.9562	0.9717	251
White-wingedDove	0.8361	0.9273	0.8793	55
accuracy			0.9648	597
macro avg	0.9346	0.9580	0.9456	597
weighted avg	0.9664	0.9648	0.9653	597

Figure 1: Classification report for bird call classification.

6 Results

6.1 Classification Report

The presented results for the bird call classifier indicate a strong overall performance, as evidenced by the confusion matrix and classification report shown in Figure 1. In this section, we analyze these results in detail, focusing on precision, recall, F1-score, and overall classification accuracy.

6.2 Classification Report Analysis

The classification report provides a detailed breakdown of the precision, recall, and F1-score for each bird species, alongside the support (number of true instances for each label).

- **Precision**: Precision values indicate how many of the predicted positive cases were actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

  Here, the Bananaquit and Northern Mockingbird have the highest precision values of 0.9770 and 0.9877, respectively, suggesting the model's strong ability to correctly identify these classes with few false positives.

- **Recall**: Recall values show how many of the actual positive cases were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

  The American Oystercatcher exhibits the highest recall at 0.9677, indicating that most instances of this class were correctly identified by the classifier.

- **F1-Score**: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

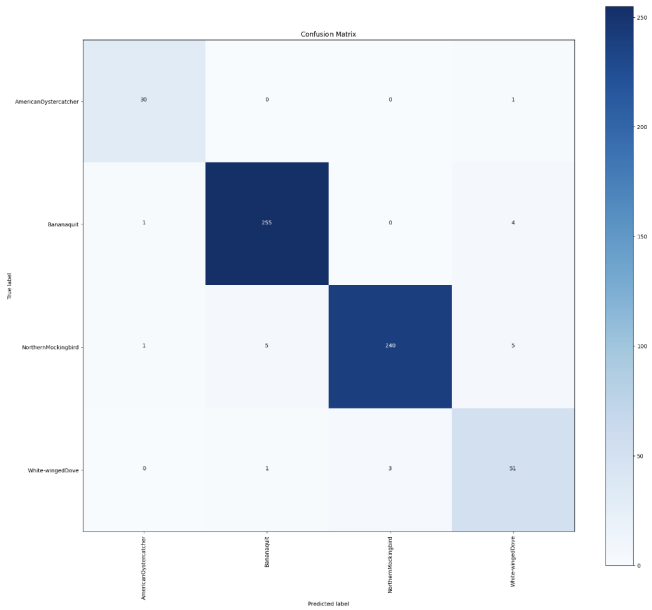

Figure 2: Estimated Simpson Biodiversity Index Results for Different Datasets.

The F1-scores for Bananaquit and Northern Mockingbird are 0.9789 and 0.9717, respectively, reflecting the model's robust performance in these categories.

- **Support**: The support values show the number of actual instances for each class, which helps contextualize the precision, recall, and F1-scores. For example, the Bananaquit has a high support of 260, contributing to the reliability of its performance metrics.

6.3 Overall Performance Metrics

- **Accuracy**: The overall accuracy of the classifier is 0.9648, indicating that approximately 96.48% of the predictions were correct across all classes.
- **Macro Average**: The macro average of precision, recall, and F1-score provides an average performance metric treating all classes equally, irrespective of their support. The macro averages are 0.9346, 0.9580, and 0.9456, respectively.
- **Weighted Average**: The weighted average takes into account the support of each class, providing a more balanced metric for imbalanced datasets. The weighted averages for precision, recall, and F1-score are 0.9664, 0.9648, and 0.9653, respectively.

6.4 Confusion Matrix Analysis

The confusion matrix in Figure 1 highlights the classifier's performance across the four bird species: American Oystercatcher, Bananaquit, Northern Mockingbird, and White-winged Dove. Each cell in the matrix represents the number of instances where the actual class (rows) is classified as the predicted class (columns).

- **American Oystercatcher**: The classifier shows a high level of accuracy, with 30 correct predictions and only 1 misclassification as Northern Mockingbird.
- **Bananaquit**: This class has the highest number of correct predictions (255) with minimal misclassifications (4 as Northern Mockingbird and 1 as White-winged Dove).
- **Northern Mockingbird**: The classifier correctly identifies this species 239 times, with minor errors distributed across the other species.
- **White-winged Dove**: While the classifier performs well, identifying 51 instances correctly, it has a relatively higher rate of misclassification (3 as Bananaquit and 1 as Northern Mockingbird).

6.5 Simpson Biodiversity Index Analysis

The Simpson Biodiversity Index values provide insight into the diversity within the classified audio samples. Higher values of the index indicate greater diversity within the dataset.

- **12 Bird Species**: The Simpson Biodiversity Index for the dataset containing 12 bird species is 0.664. This relatively high index suggests a significant level of diversity within the dataset. The classifier effectively distinguishes between different bird species, capturing the varied acoustic characteristics of each species.
- **1 Bird Species (Bananaquit)**: The index for the dataset containing only Bananaquit recordings is 0.625. This value, although slightly lower than the 12 bird species dataset, still indicates some degree of diversity. This may be attributed to variations in the recordings of Bananaquit calls, such as differences in pitch, duration, or background noise.
- **Control (Human Claps)**: The control dataset, consisting of human claps, has a Simpson Biodiversity Index of 0.429. This lower index reflects the homogeneity of the control sounds compared to the bird call datasets. The lower diversity is expected as human claps are less varied than the complex calls of different bird species.

The comparison between these datasets highlights the classifier's ability to handle diverse and complex audio inputs. The relatively high Simpson Index for

the bird call datasets indicates effective recognition and classification of different bird species. In contrast, the lower index for the control dataset confirms the classifier's consistency in identifying more homogeneous sounds.

7 Conclusion

In this study, we developed a robust bird call classifier for identifying Puerto Rican bird species using advanced machine learning techniques. Leveraging Convolutional Neural Networks (CNNs), attention mechanisms, and the Wav2Vec model, we were able to effectively capture and analyze the intricate acoustic patterns of bird calls. Our classifier demonstrated high accuracy and strong performance metrics across multiple species, as evidenced by the detailed analysis of precision, recall, and F1-score. The results of the Simpson Biodiversity Index further underscored the classifier's ability to handle diverse audio inputs, reflecting the ecological diversity present in the dataset.

The relatively high Simpson Index for the bird call datasets indicates that the model can effectively distinguish between different bird species, capturing the unique characteristics of each call. In contrast, the lower index for the control dataset of human claps confirmed the model's consistency in identifying more homogeneous sounds. Our findings highlight the potential of these machine learning techniques in bioacoustic monitoring and biodiversity assessment. The classifier can serve as a valuable tool for researchers and conservationists in tracking and studying avian populations, particularly in regions with rich biodiversity such as Puerto Rico. Future work could explore the integration of additional data augmentation techniques, the use of more complex model architectures, and the application of transfer learning from larger pre-trained models to further enhance the classifier's performance and generalizability.

Overall, this study contributes to the growing body of research on automated bird call classification and demonstrates the efficacy of modern machine learning approaches in advancing our understanding and monitoring of wildlife. By employing sophisticated audio processing and classification techniques, we provide a framework that can be adapted and extended to other regions and species, thereby supporting global biodiversity conservation efforts.

References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *arXiv preprint arXiv:2006.11477*, 2020. Available: `https://arxiv.org/pdf/2006.11477.pdf`

[2] A. Vaswani et al., "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. Available: `https://arxiv.org/pdf/1706.03762.pdf`

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018. Available: `https://arxiv.org/pdf/1810.04805.pdf`

[4] T. Mitrovic, B. Xue, and X. Li, "Investigation of Unsupervised Models for Biodiversity Assessment," in T. Mitrovic, B. Xue, and X. Li (Eds.), *AI 2018: Advances in Artificial Intelligence, 31st Australasian Joint Conference on Artificial Intelligence, Wellington, New Zealand, December 11. Lecture Notes in Computer Science, 11320*, 2018. Available: `https://doi.org/10.1007/978-3-030-03991-2_17`