WildNet: Wildlife Audio Classification Using CNN-Attention-Transformer Architecture

Abstract

Wildlife audio classification is a critical tool for ecological monitoring and biodiversity conservation. In this paper, we present a novel multi-label classification model for recognizing species-specific vocalizations, focusing on birds, mammals, and amphibians. Our approach utilizes a hybrid architecture combining convolutional neural networks (CNNs), channel attention mechanisms, and transformer encoders. This model is optimized for the BirdCLEF 2025 dataset, which comprises tens of thousands of annotated soundscape recordings. We further employ focal loss and stratified multi-label splitting to handle class imbalance and improve generalization.

1. Introduction

Automated audio classification of wildlife species has become increasingly important with the advent of passive acoustic monitoring systems. These systems generate large volumes of data, demanding robust and scalable machine learning solutions. Our work builds upon recent advances in deep learning, proposing a modular pipeline that processes raw audio into spectrogram representations and leverages attention-enhanced deep models for multi-label prediction.

2. Dataset Description

We use the BirdCLEF 2025 dataset provided by the LifeCLEF challenge. It includes:

- **Audio Clips**: Over 80,000 `.ogg` and `.wav` files sampled at 32kHz or 44.1kHz.
- **Labels**: Multi-label annotations including primary and secondary species IDs, often with background overlaps.
- **Meta-data**: Species scientific names, common names, geolocation, and recording timestamps.

Each clip may contain overlapping calls from multiple species, presenting a realistic and challenging scenario for classification.

3. Data Preprocessing

3.1 Audio Loading and Resampling

All audio is resampled to a uniform rate of 32kHz for consistency. We ensure mono channel formatting for standardized spectrogram generation.

3.2 Spectrogram Generation

We convert audio into Mel spectrograms using Librosa, with the following parameters:

- Sampling rate: 32000
- Number of Mel bands: 128 (fixed height)
- Duration: Unfixed; we do not truncate or pad the waveform
- Dynamic width: Resized to 206 pixels using TensorFlow bilinear interpolation

Spectrograms are resized to a fixed shape of **128 x 206** using bilinear interpolation, without applying zero-padding or waveform truncation. This preserves the original audio content and avoids introducing artificial silence or clipping. The interpolation ensures all inputs match the model's dimensional requirements while retaining the spectral-temporal structure.

Additionally, we expand the spectrogram dimensions to add a single channel, making the input shape suitable for CNNs: **(num_samples, 128, 206, 1)**.

Why Mel Spectrograms? Mel spectrograms mimic the human auditory system's sensitivity to frequency, making them effective for capturing bioacoustic signals. They also reduce model complexity by eliminating irrelevant frequency information.

3.3 Normalization

- Spectrogram values are converted to decibels (log scale)
- Min-max normalization scales the values to the [0, 1] range

3.4 Label Encoding

Labels are one-hot encoded for multi-label classification. Primary and secondary species labels are both included. Secondary labels are parsed from string representations where available.

3.5 Stratified Multi-Label Split

We use `MultilabelStratifiedKFold` from scikit-multilearn to ensure balanced label representation across training and validation folds.

4. Model Architecture

Our architecture combines spatial feature extraction, channel-wise attention, and temporal context modeling. The model is implemented using TensorFlow and trained on a Tesla P100 GPU, enabling efficient handling of large-scale audio data and complex model computations.

4.1 CNN Backbone with Channel Attention

We start with a 3-stage 2D CNN to extract local spatial features from spectrograms. Each stage includes:

- Convolutional Layer (Conv2D)
- Batch Normalization
- ReLU Activation
- Max Pooling
- Spatial Dropout

Channel Attention Module (CAM) is inserted after each block to enhance discriminative channel features. It computes:

- Global Average Pooling (GAP)
- Dense layers with ReLU and sigmoid activations to generate channel-wise weights

This boosts the model's sensitivity to relevant frequency patterns.

4.2 Transformer Encoder Block

The output from the CNN layers is reshaped into a temporal sequence before entering the transformer. Learned positional embeddings are added to encode temporal order. Each transformer block contains:

- Multi-head Self-Attention (8 heads, 64-dimensional keys)
- Feed-forward network (2-layer MLP with GELU activations)
- Dropout and residual connections
- Layer Normalization

We use two such transformer layers, enabling the model to capture long-range dependencies in the spectrogram.

4.3 Global Pooling Classifier Head

After transformer encoding, we apply Global Average Pooling across the sequence dimension, followed by:

- Dense Layer with ReLU activation (512 units)
- Dropout (0.5)
- Final Dense Layer with Sigmoid activation (206 units for multi-label classification)

This design eliminates the need for token-wise decoding and allows the model to summarize all time steps into a fixed-length vector.

5. Loss Function

We use **Focal Loss**, defined as:

$$FL(pt) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

Where:

- p_t: predicted probability
- α_t: class balance factor
- γ : focusing parameter

This loss down-weights well-classified examples, focusing the learning on hard and minority class samples.

6. Evaluation Metrics

We evaluate performance using:

- Area Under ROC Curve (AUC)
- PR-AUC (Precision-Recall AUC)
- Precision@Recall
- Recall@Precision

These are suited for imbalanced multi-label problems.

7. Results and Discussion

The BirdCLEF 2025 dataset posed a significant challenge due to extreme class imbalance across 206 species:

- The most represented class had **990** samples, while the least represented class had only **2** samples.
- Despite this imbalance, **WildNet** achieved impressive results:
    - **88%** score on both the training and test sets.
    - Evaluated using a **macro-averaged ROC-AUC score variant** that **ignores classes without any true positive labels**, making it especially reliable in highly imbalanced settings.

Key observations:

- Channel attention significantly improved the results.
- Focal loss proved more robust compared to class weighting.
- No data augmentation techniques were used.

8. Future Work

- Integrate audio event localization
- Use wavelet or harmonic-percussive transforms
- Deploy lightweight version for real-time edge inference
- Generalize the architecture across various spectrogram classification tasks

9. Conclusion

We present a high-performing, modular pipeline for multi-label wildlife audio classification. Our WildNet model begins with initial convolutional blocks to extract spatial features, followed by transformer layers for temporal modeling. Spectrograms are generated without truncating or padding the waveform; instead, they are uniformly resized using bilinear interpolation. This design choice helps preserve the richness of the acoustic signal while ensuring compatibility with fixed-size model input.

Instead of traditional token-wise decoders, we utilize a global pooling classifier head that simplifies the prediction pipeline. This architecture demonstrates state-of-the-art performance on complex natural audio scenes and holds promise for scalable biodiversity monitoring systems.

References

- Vaswani et al., "Attention is All You Need", NeurIPS 2017.
- Lin et al., "Focal Loss for Dense Object Detection", ICCV 2017.
- BirdCLEF 2025 Challenge, LifeCLEF.
- Tokozume et al., "Learning from Between-class Examples for Deep Sound Recognition", arXiv:1711.10282.
- Kong et al., "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition", IEEE/ACM Trans. ASLP, 2020.
- Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021.
- Hershey et al., "CNN Architectures for Large-Scale Audio Classification", ICASSP 2017.
- Huang et al., "AST: Audio Spectrogram Transformer", arXiv:2104.01778.