CONVOLUTIONAL RECURRENT NEURAL NETWORK BASED BIRD AUDIO DETECTION

Technical Report

Rajdeep Mukherjee[1]*, Dipyaman Banerjee*[2]*, Kuntal Dey*[2]*, Niloy Ganguly*[1]

[1] Indian Institute of Technology, Kharagpur, India, {rajdeep1989.iitkgp, ganguly.niloy}@gmail.com
[2] IBM Research - India, New Delhi, India, {dipyaban, kuntadey}@in.ibm.com

ABSTRACT

We propose a Convolutional Recurrent Neural Network (CRNN) based approach, implemented as a Convolutional Neural Network (CNN) followed by a Recurrent Neural Network (RNN), for the task of detecting the presence of birds in audio recordings. As part of the IEEE DCASE 2018 Challenge, we were provided with three separate development datasets containing recordings from three very different bird sound monitoring projects. We performed a stratified 3-way cross-validation mechanism for training our model by considering two datasets for training and the remaining one for validation in each fold in order to generalize our model well when exposed to data from unseen conditions. We obtained an Area Under Curve (AUC) measure of **88.7%** on the leaderboard test set. We compare our results with the CNN version of our model which achieves an AUC measure of **87.74%** on the same test set.

Index Terms— Bird Audio Detection, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, freefield1010, warblrb10k, BirdVox-DCASE-20k datasets

1. INTRODUCTION

Detecting bird sounds in audio is an important task for automatic wildlife monitoring. The audio modality is well-suited to bird monitoring because many birds are much more clearly detectable by sound than by vision or other indicators. The details of the first DCASE Bird Audio Detection Challenge can be found in [1]. In DCASE Bird Audio Detection Challenge 2018 we are given three datasets for training and validation of our machine learning models. The primary goal here is to create a model that can detect occurrence of bird sound in a given dataset. It is a binary decision problem and does not involve detecting multiple occurrence or detecting the start and end times for such occurrences. The main challenge here is to create a highly generalized model which can adapt well to previously unseen datasets where the data is collected from environments other than the training dataset or uses different methods for bird sound collection. To that end, the evaluation datasets provided in the challenge are widely different from that of the training dataset [1]. In addition, any model trained on the given training set also requires to be sufficiently robust to tolerate any error in ground truth labeling. As such, the three training datasets have varying label accuracy which also poses a significant challenge.

In this report, we describe our deep neural network based approach which we used for detecting bird sound in a given sound file. We first used a convolutional neural network and then extend it to a convolutional-recurrent neural network by adding a recurrent block on top of the convolutional block. We finally use a sigmoid layer to detect the occurrence or absence of bird sound. While the convolutional layers help us in detecting useful features which are invariant to spectral changes, the recurrent layers help us in detecting any temporal feature which may be present in the in the sound file. We use a 128-band log-mel spectrogram as our input to the model. To improve generalization, we performed a stratified 3-way cross-validation mechanism for training our models by considering two datasets for training and the remaining one for validation in each fold in order to generalize our model well when exposed to data from unseen conditions. The details of the feature extraction method and the neural network architecture is given in Section 2.

The initial results have been highly encouraging. With CNN architecture, our model achieves approximately 87.74% accuracy over a subset of the evaluation dataset on the leader-board submission portal provided in the DCASE challenge website. The accuracy measure is chosen to be Area Under Curve (AUC) in accordance with the challenge specification. The accuracy further improves to 88.7% when we add the recurrent block. The details of the results across different versions of our model architecture is given in Section 3.

2. OUR APPROACH

2.1. Overview and Feature Extraction

Our proposition is based on a deep neural network learning methodology. The input to the training phase of the system comprises of Log-scaled Mel Spectrogram representation of the audio signals. For each audio file under analysis, each of which is of 10 second duration, we randomly consider 8.91 seconds of data and extract a 128-band Log Mel Spectrogram, that covers the audible frequency range of audio signals, namely 0-22, 050 Hz, with 23.22 ms windows ($1,024$ samples at 44.1 KHz), and a hop size of 512 samples to allow for 50% overlap among consecutive frames. The obtained segments (128 rows/bands * 768 columns/frames) were provided together with their deltas (computed with default librosa settings) as a 2-channel input to the network. The features are subsequently passed over to the deep neural network for training and testing, in the respective phases.

2.2. The Deep Learning Architecture

For deep learning, we design and compare two different model architectures. The architecture details for the Convolutional Neural Network (CNN) based approach is presented in Table 1. We extend this model with two additional Gated Recurrent Unit (GRU) based Recurrent Neural Network (RNN) layers to come up with the design of the Convolutional Recurrent Neural Network (CRNN) architecture. The output of the RNN is passed over a fully connected layer,

wherein the final class label is generated. The architecture details are provided in Table 2.

Input: $128 \times 768 \times 2$
5×5 Convolution 2D (32) + ReLu 5×5 Convolution 2D (64) + ReLu Batch Normalization 4×4 MaxPooling 2D (Strides 4×4) + Drop-Out (0.3)
5×5 Convolution 2D (64) + ReLu Batch Normalization 5×5 Convolution 2D (64) + ReLu Batch Normalization 5×5 MaxPooling 2D (Strides 5×5) + Drop-Out (0.3)
3×3 Convolution 2D (64) + ReLu 3×5 MaxPooling 2D (Strides 3×5) + Drop-Out (0.5) ReShape (7, 64)
Dense(1) + L2 Weight_Regularizer (0.001) + Drop-Out (0.5)
Activation (Sigmoid)

Table 1: CNN Model Architecture. ReLu ← Rectified Linear Unit.

Input: $128 \times 768 \times 2$
5×5 Convolution 2D (32) + ReLu 5×5 Convolution 2D (64) + ReLu Batch Normalization 4×4 MaxPooling 2D (Strides 4×4) + Drop-Out (0.3)
5×5 Convolution 2D (64) + ReLu Batch Normalization 5×5 Convolution 2D (64) + ReLu Batch Normalization 5×5 MaxPooling 2D (Strides 5×5) + Drop-Out (0.3)
3×3 Convolution 2D (64) + ReLu 3×5 MaxPooling 2D (Strides 3×5) + Drop-Out (0.5) ReShape (7, 64)
Gated Recurrent Unit (200)
Gated Recurrent Unit (200)
Dense(1) + L2 Weight_Regularizer (0.001) + Drop-Out (0.5)
Activation (Sigmoid)

Table 2: CRNN Model Architecture. ReLu ← Rectified Linear Unit.

3. EXPERIMENTS

We perform training on the features extracted from the audio files as explained in Section 2. For stratified 3-way cross validation, we train our model for 50 epochs for each of the folds. The obtained AUC score on the Development datasets as reported in Table 3 is the Harmonic Mean of the AUC scores of the three folds.

4. CONCLUSION

In this paper we present a deep neural network based model to detect occurrence of bird sound in a given sound file as part of the Bird Audio Detection Challenge in DCASE 2018. We provide two variations of our model architecture. We first use a convolution neural network and then extend it to add a recurrent layer on top. Both the models achieve good accuracy on both development datasets and a subset of the evaluation dataset provided in the leaderboard portal. The C-RNN architecture however slightly outperforms the CNN architecture on these datasets. We would like to continue to work on this problem and use data augmentation and transfer learning techniques to further improve accuracy. Any additional improvement on training and validation accuracy will be reported in a DCASE workshop paper which we intend to submit in the near future.

Method	**AUC (Development Dataset)**	**Leaderboard AUC Score**	**Model Summary**
Our_CNN	84.22%	87.74%	Five conv layers with Batch Normalization, followed by a single dense layer.
Our_CRNN	85.52%	88.7%	Five conv layers with Batch Normalization followed by two GRU layers and one dense layer

Table 3: Performance comparison of various methods

5. REFERENCES

[1] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, 2018.