Deep learning classify the Bird songs with STFT.

Mr.Phiphat Chomchit, 630631028
Data Science Consortium Faculty of Engineering Chiang Mai University

November 18, 2020

Abstract

This report need to study about Sound Classification. The data that we want to study is Bird Song Dataset[1]. We have audio files of 50 species (Birds species in Europe). We try to use Deep learning to classify soud each bird species that we tranform it with STFT(Short time fourier tranform). And we want to make a dashbord to show STFT graphs for each species.

Keywords:
Neural network, Deep learning, Sound classification, Short time fourier tranform

1 Introduction

Audio(Sound) is one of the main sensory information we receive to perceive our environment. Almost every action or an event in our surroundings has its unique sound. Audio has 3 main attributes which help us in distinguish between two sounds.

Amplitude — Loudness of the sound

Frequency — The pitch of the sound

Timbre — Quality of the sound or the identity of the sound

However, the audio event recognition systematically (preferably using a computer program or an algorithm) is very challenging. This is mainly because of,

The noisiness of recorded sound clips — transducer noise and background noise.

An event can be occurring at various loudness levels and various time durations.

Having a limited number of examples to feed into an algorithm.

The preprocessed audio files themselves cannot be used to classify as sound events. We have to extract feature from the audio clips to make the classification process more efficient and accurate. Let's extract the absolute values of Short-Time Fourier Transform (STFT) from each audio clip. To calculate STFT, Fast Fourier transform window size(n_fft) is used as 512. According to the equation n_stft = n_fft/2 + 1, 257 frequency bins(n_stft) are calculated over a window size of 512. The window is moved by a hop length of 256 to have a better overlapping of the windows in calculating the STFT.[2]

2 Research methodology

2.1 Data

Meta data, there are 2150 rows and 61 Features.
For example

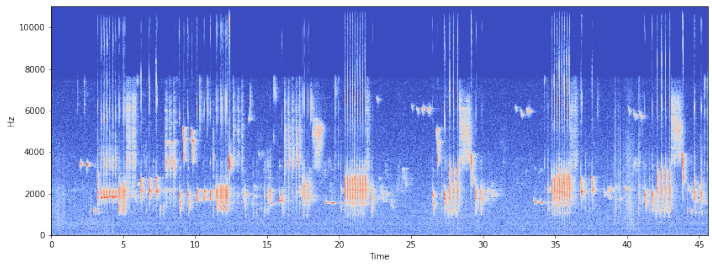
- Recording_ID (Audio file name),
- Genus,
- Specific_epithet,
- Subspecies,
- Other_species (Sound of other bird species in sound clip),
- Length (Length of sound clip),
- Vocalization_type (Behvaiour of bird)
- Species. (There are 50 species. Each species has 43 file)

2.2 Data classication

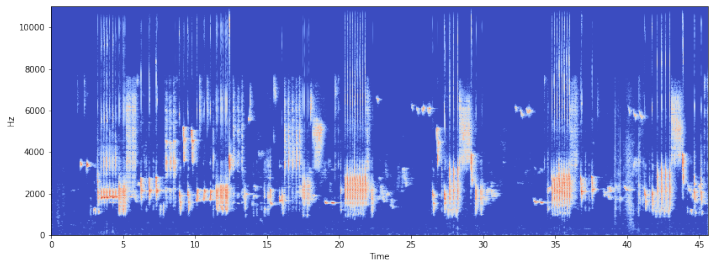
2.2.1 Data Classication Techniques

1.) Change "Length" (type str) to time in second.
2.) Drop row that have "Length" more than 135 second.
3.) Drop row that have other species in sound clip.
4.) Slect the sound clip with "Recording_ID".
5.) Clean sound with Noise reduction and Trimming.

Before Clean noise.



After Clean noise.

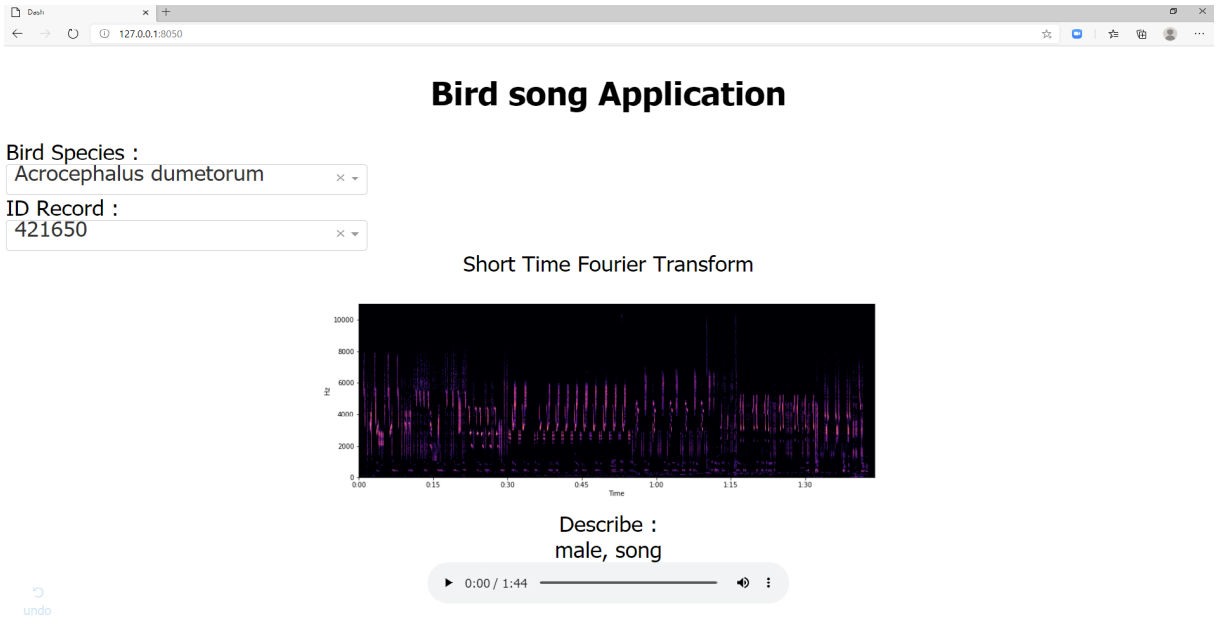


6.) Extract features with STFT and save it in "Recording_ID.npy" file. And save STFT graphs.
7.) Create dummy data for column "Species" ['Fringilla', 'Parus', 'Turdus', 'Sylvia', 'Emberiza'] (we choose only 5 species, It made model has high accuracy).
8.) Split test train data (test size = 0.3, X is Extracted features, y = dummy Species).

9.) Create Deep leraning model (input_shape = 257, output_shape=5).

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 256)               66048
_________________________________________________________________
dense_1 (Dense)              (None, 256)               65792
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_3 (Dense)              (None, 5)                 645
=================================================================
Total params: 165,381
Trainable params: 165,381
Non-trainable params: 0
_________________________________________________________________
```

10.) Train the model.
11.) Make the dashbord show STFT graphs and describe of the sound.



2.2.2 Evaluation

1.) Loss function, The loss function that we use in this studdy is categorical cross entropy.
Let y be actual and $\hat{y}$ be a prediction. We define "categorical cross entropy"

$$L(\hat{y}, y) = -\frac{1}{M}\sum_{i=1}^{M}[y_i \log(\hat{y}_i) + (1-y)\log(1-\hat{y})]$$

2.) Accuracy. When we input data into medel, model will return vector $V_{(1,5)}$.
We define the prediction

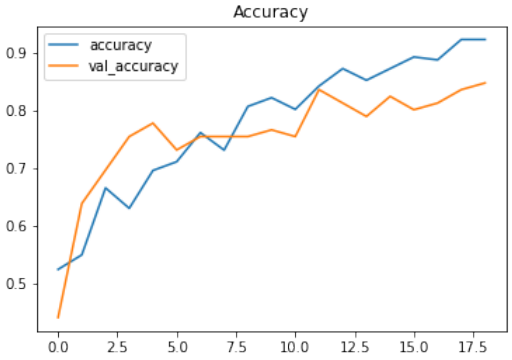
$$Prediction = argmax(V)$$

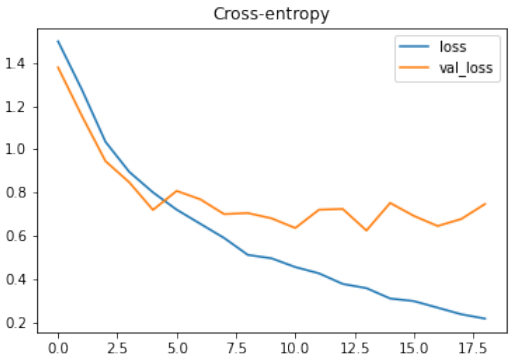
Then find the accuracy from prediction.

$$Prediction\_accuracy = \frac{Correct\_test}{All\_test}$$

2.3 Result

2.3.1 Loss value and Accuracy





Acual = [0, 4, 3, 0, 0, 3, 0, 0, 1, 1, 4, 0, 1, 3, 4, 0, 0, 3, 2, 3, 3, 0, 0, 3, 3, 3, 3, 0, 0, 0, 0, 1, 1, 4, 0, 0, 3, 1, 0, 3, 4, 3, 3, 3, 4, 3, 3, 3, 0, 0, 4, 4, 3, 1, 2, 0, 3, 3, 0, 0, 2, 3, 3, 0, 0, 2, 3, 3, 3, 0, 0, 0, 3, 3, 0, 3, 0, 2, 3, 3, 3, 0, 4, 3, 1, 3]

Predict = [0, 4, 3, 3, 2, 3, 0, 0, 1, 1, 3, 0, 1, 3, 4, 0, 0, 3, 2, 3, 3, 0, 0, 3, 3, 3, 1, 0, 0, 3, 2, 3, 4, 4, 0, 0, 3, 1, 0, 1, 4, 3, 3, 3, 4, 3, 3, 4, 0, 0, 4, 4, 3, 3, 2, 0, 3, 3, 3, 0, 2, 3, 3, 0, 2, 0, 3, 3, 3, 0, 0, 3, 3, 3, 0, 1, 0, 2, 3, 1, 3, 0, 4, 3, 3, 3]

$$Prediction\_accuracy = \frac{68}{86} = 0.8$$

3 Conclusion

Deep learning can classify sound of 5 species of bird with accuracy 80%

References

[1] https://www.kaggle.com/monogenea/birdsongs-from-europe

[2] https://towardsdatascience.com/sound-event-classification-using-machine-learning-8768092beafcet