Project Report

Qiaochen Sun

Nowadays, human activities such as wildfires and hunting have become the largest factor that would have serious negative effects on biodiversity. In order to deeply understand how anthropogenic activities deeply affect wildlife populations, field biologists utilize automated image classification driven by neural networks to get relevant biodiversity information from the images. However, for some small animals such as insects or birds, the camera could not work very well because of the small size of these animals. It is extremely hard for cameras to capture the movement and activities of small animals. To effectively solve this problem, passive acoustic monitoring (PAM) has become one of the most popular methods. We could utilize sounds we collect from PAM to train certain machine learning models which could tell us the fluctuation of biodiversity of all these small animals. The goal of the whole program is to test the biodiversity of these small animals (most of them are birds). However, the whole program could be divided into plenty of small parts. I and Jinsong will pay attention to the intermediate step of the program.

The Neurips 2021 paper provides us with an overview of the intermediate steps we need to cope with in the project. What is the Neurips paper about? And how does the project in this paper relate to the huge program we discuss in the first paragraph? Like I illustrate in the first paragraph, we need to collect plenty of audio recordings, and then group members would let many

volunteers generate species-level ground truth labels on a subset of audio recordings for the field and some people in the program will utilize these data to train, test and validate machine learning models to understand biodiversity situation. However, the problem is that there are lots of audio data we collected from South America that do not have bird sounds we are interested in at all. Even though volunteers spend a lot of time labeling these “useless” data, we could not utilize these blank data to train our model. Therefore, we need to come up with some solutions to generate subsets of audio recordings that have higher probability of vocalization of interest. The solutions could help us reduce down the amount of time and resources required to achieve enough training data for species-level classifiers. Thus, by deploying the solutions, we could save volunteers’ time and let them label more useful audio data in a given time period.

In the paper, four methods have been proposed and tried to solve the problem. The four methods are baseline stratified random sampling, sampling with knowledge of diurnal bird 48 vocalization trends, using a Microfaune, neural network model designed for audio event detection, and the combination of diurnal bird vocalization trend and Microfaune. In order to test the different methods, we constructed four separate datasets of 240 audio clips and applied four methods into these four datasets. We also need to label our audio data as presence or absence and high activities or low activities in order to test the accuracy of different methods and find the best one. We found out that the

combination of diurnal bird vocalization trend and Microfaune, a CNN-RNN machine learning model, could provide us with the highest presence rate of 95%. The presence rate means that 95 percent of audio data in the subsets we collect from raw audio data have bird vocalization that we are interested in. With 95% probability that audio will have the vocalization of interest, volunteers could work more efficiently and our coworkers could use less time and resources required to achieve enough training data for species-level classifiers. The figure1 illustrates the result of four methods.

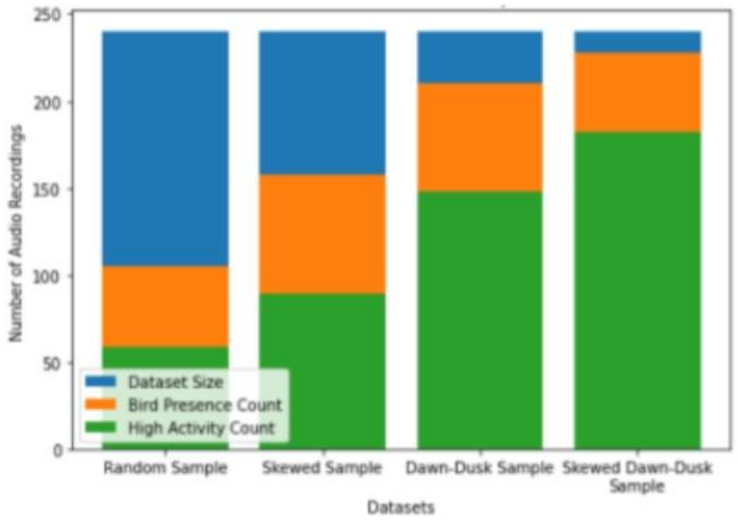

Figure1

The y-axis is the number of audio recordings and the x-axis is datasets.

Now, we would like to illustrate the Microfaune model and diurnal bird vocalization trends. Microfaune model is a hybrid CNN-RNN model built for

event detection on low resource datasets and this model is a binary classifier. The input data of this model should be .wav audio datasets (will be transformed into Mel spectrograms in the model) and predict the presence or absence of the bird in the audio. In the model, we have a local score and global score. Local score is the score that predicts the probability of bird presence in each position of the Mel spectrogram and global score is the maximum probability of bird presence, which means that global score is the maximum local score. In this project, we first choose all the subsets of the audio dataset with a global score that is higher than 50% after putting them into Microfaune. After that we need to use diurnal bird vocalization trends to select another subsets from the subsets we selected by using the microfaune model. Diurnal bird vocalization trends mean that birds are more active in dusk and dawn, which means we have a higher chance to record audio with the vocalization of interest. By selecting the overlap subsets of them, we get that the probability of the audios with the vocalization of interest is around 95%, which is much higher than 44 percent and 44% is just the probability if we randomly select the audios. We could see the overlap of the subsets in figure2.

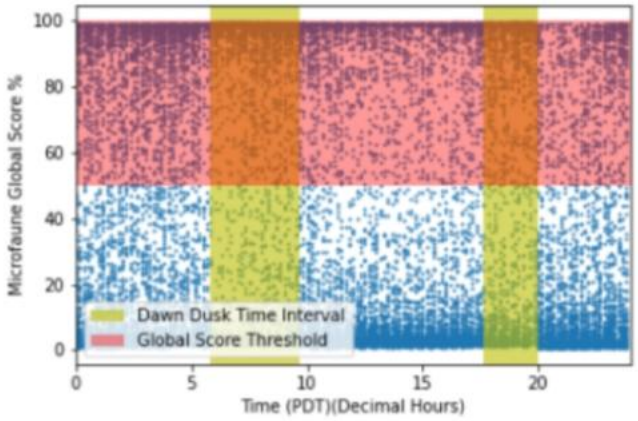

Figure 2

From figure1, we can see that the y-axis is the global score and the x-axis is Time hours. The time lengths contained in yellow columns are dask and dawn.

In this quarter, we have already understood the logic of our project and how to use microfaune mode. We wrote some codes of microfaune to get started and get prepared for next quarter. We randomly selected four different wav data (birds sound) on xeno-canto website and predicted the global score using microfaune.

In [143…

```
global_score = []
for i in x:
    glob, loc = detector.predict_on_wav(i)
    global_score.append(glob)
global_score
```

Out[143…

```
[array([0.8777194], dtype=float32),
 array([0.99038446], dtype=float32),
 array([0.9671296], dtype=float32),
 array([0.99264914], dtype=float32)]
```

From the result, we could see that most of audio's score are more than 0.95, which is pretty good. Because we choose these wave files from Xeno-canto and audio on this website have already been processed by some methods. Therefore, it's very normal to have higher global score. This is just the start code for our project, in next quarter, we will continue work on microfaune model and have a better understand of this model and our project.

In the next quarter, we will continue working on the microfaune model and the solutions described in the paper. This is a good starting point for our whole project and I believe we could finish this project very well next quarter!