Project Proposal Bird sound classifier

Group NaN

This sample solution contains artificial data and insights. It should not be taken as truth, but as a help to set expectations. Some of the plots are missing axes labels and titles because they're fake. It's also better to do it in LaTeX. This is not a gold standard!

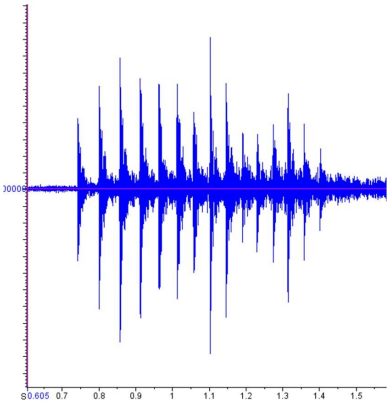
For our custom project we're attempting to build a classifier to recognise birds by their song. Labeled song recordings are available from fakebirds.com/notadataset.

1. Preliminary domain knowledge

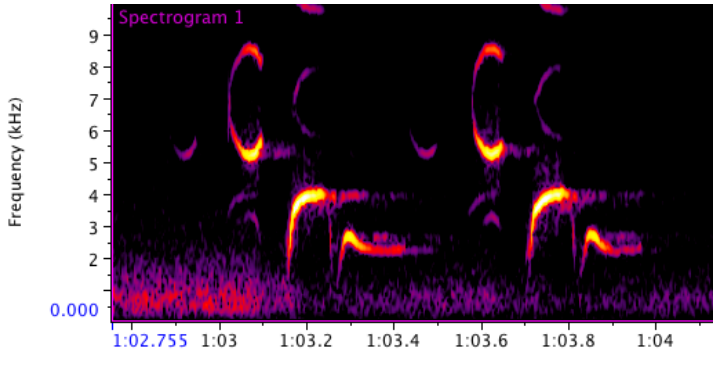
Cornell Lab has run a citizen science project offering BirdNET as a classifier to monitor bird populations (https://birdnet.cornell.edu/). They used a ResNet CNN to classify the bird species based on spectrograms (image representation of frequencies over time). Worth highlighting is that they used 2 spectrograms, one from 0-3000Hz, and one from 500-15000Hz. They then scaled the images to a resolution of 96x511 pixels. BirdNet was trained on more than 6000 species, whereas the dataset from fakebirds.com only considers 5 "Orders" of birds: owls, seagulls (and relatives), woodpeckers (and relatives), pigeons (and relatives) and ducks (and relatives). This higher order perspective means there will be more variance within each class, but also makes sure that all the classes are a really distinct sound.

Since woodpeckers are often heard by their drilling, rather than their song, there's a good chance that fakebirds.com has the drilling sound of woodpeckers, rather than the song.

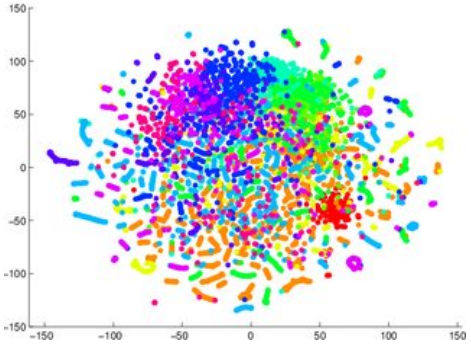
2. Preliminary data exploration



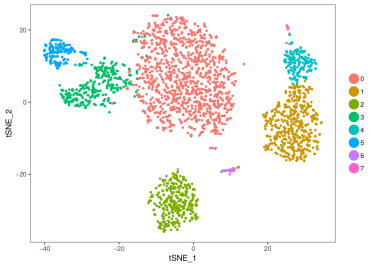
Woodpecker waveform



Seagull spectrogram



t-SNE plot of raw data (color=class)



t-SNE of frequencies

As suggested in 1. there was a possibility that the woodpeckers were actually drilling, rather than singing. We listened to some samples and show a plot that indicate this is the case. As a result, we find that the woodpecker is very identifiable by the waveform (specifically envolope/changes in volume), and that it is not dependent on frequency/pitch. For other birds the spectrogram was a much better representation.
We used t-SNE on the raw data to see if classes were distinct. We found that this did not work, possibly because the raw representation is very high dimensional, and the dimensions aren't specifically meaningful (a delay of 1/40000th of a second means you have an entirely different sample). Instead we plotted t-SNE of frequencies as well (over the entire recording), and found this to be much more separable.
From this we expect the raw data to perform poorly, but the distribution of volume and frequency might work. Possibly even without using spectrograms and CNNs.

3. Proposed preprocessing

We found that for some recordings there was silence before or after the song. We want to crop this out. We will not be filtering out any frequencies, as we could not establish a good range for this. We'll use the fourier transform to extract the amplitude of certain frequencies, and bin that at intervals of 25Hz, up to 25kHz. This will give 1000 features, which we'll reduce with PCA. We do this because the t-SNE of frequencies suggested that we may not need efforts of time that a spectrogram would capture. 25kHz is above the nyquist frequency of recordings that are sampled at 48kHz, so there is no frequencies lost.

4. Proposed Model + baseline

Our baseline model will be 5 small one-vs-rest Multi-Layer-Perceptrons. The "full" model that we'll be using is a small Multi-Layer-Perceptron with 5 outputs. We hope to show that this can benefit from combining things it's learned from different classes. We'll use one hidden layer, and use gridsearch to determine how big the hidden layer should be. Another important hyperparameter that we want to tune is the number of Principal Components. For the Minimum Viable Product we'll also first have a logistic regression.

5. Proposed evaluation

We'll be assessing our models with a classification report and the global accuracy. Data is balanced and we do not favor any particular class. We'll also look at the confusion matrix to see if certain birds appear similar. We expect the woodpecker to be easiest to classify.

6. Model usage

Due to the oversimplified birds, this model is not useful as an app for experienced bird enthusiasts like BirdNET is. Instead, it may be a nice educational tool for children. Alternatively, the model could be put on a Raspberry Pi in the forest. It can do constant recordings (before), and foresters can use this as analytics to assess bird population (after). This may need a step to make sure a sound is a bird sound, and not e.g. a human (before).

7. Risk assessment

Our design assumes that birds are classifyable by frequency, but ignores time. This has the benefit of easily scaling to longer timespans, but it may be that changes in frequency over time are actually essential. We may spot this problem in the MVP stage, and modify the feature extraction method if this proves to be a problem. We do not foresee any societal risks for this project.

8. Individual Learning Outcomes

Alice: I would like to learn how to write unit and integration tests. This was covered in a course before, but I never really worked with it.
Bob: I want to improve my familiarity with audio processing. I'm a musician so I have some applied familiar, but I would like to be closer to the low-level aspects.

9. Member contributions

Alice and Bob got together to go over the questions and did the writeup together. Bob did the typing. Alice make the tSNE plots, Bob the spectrogram & waveform plots. Carol has unfortunately dropped out of the course.