Title of Project: When Computers Learn Frog Calls
Project Category: Convolutional Neural Networks for Frog Call Classification
Team Members: Ami Kano (ak7ra), Kate Meldrum (kmm4ap), and Tyler Valentine (xje4cy)

Motivation

This project aims to apply a deep learning approach to identify frog species by their calls in the form of audio recordings. The previous work that inspired this project used a variety of extracted features to inform a Gaussian mixture model. However, our approach is to convert audio files of individual frog calls to Mel spectrogram images for use in training a convolutional neural network. This is advantageous because it would allow for the future use of the model with minimal preprocessing of audio files, and without extensive variable engineering.

The ability to identify species of frogs by their calls is valuable to researchers who wish to monitor which types of frogs are present in various locations over time. For example, the presence of an endangered species of frog in an area might prompt conservation efforts for that ecosystem. Being able to monitor this through audio files provides a non-invasive way to track the movement of species over time. Moreover, using machine learning for classification reduces the manpower required to achieve this on a large scale.

Dataset

The data[1] is organized into 10 folders. Each folder is named after a frog species and contains 5~10 audio files of the call of that species. The length of the audio file can range from 19 seconds to more than 10 minutes. Most audio files contain silence/noise in addition to the frog calls. An audio file could also contain the voice of the person who is recording. It is necessary to cut out human speech and silence out of the audio files before processing them for our model.

We intend to translate the audio files into images, or Mel spectrograms, to build a neural network image classification model. Initial attempts to convert audio files to Mel spectrograms have been successful. From the spectrograms, it was observed that frog calls from individual frogs of different species have clearly differentiable shapes. For example, the calls of the frog species *E. petersi* have shapes comparable to a sideway apostrophe, whereas the calls of the species *B. lanciformis* have relatively more vertical shapes.

Related Work

The creators of the dataset - Terneux, Nicolalde, Nicolalde, and Merino-Viteri - produced a paper called *Presence-absence estimation in audio recordings of tropical frog communities*.[2] This paper shows a Gaussian Mixture Model trained to predict the species of a frog call. Although the specific kind of model is different from what we intend to make, this paper is quite relevant to our work as it uses the same data. We also intend to loosely mimic the segmentation of audio files described in the paper.

In addition, there are many past works like this project, for example, classifying bird calls with neural networks has been attempted countless times, with a notable example being *Bird sound recognition using a convolutional neural network*[3] by Incze, et al. The authors of this paper convert audio data into spectrograms and create a convolutional neural network. The specific CNN described in this paper is a fine-tuned version of a pre-trained MobileNet model, but other papers have attempted to design their own CNN. Past work like this will be helpful to inform our efforts.

Technical Plan

The audio files in our data set consist of a series of numbers, each indicating the amplitude of the sound wave for each timestep. Because this format only considers the amplitude of the sound wave, it is difficult to differentiate between sounds that have similar intensities. As a result, we have found that most deep learning models do not take raw audio as an input. It is more common for researchers to use Fourier Transforms to decompose the sound waves and obtain the amplitude of each frequency in the signal. This information is represented in an image called a spectrogram. We have chosen to use the Mel spectrogram, which plots the frequencies using the Mel scale vs. time and uses the Decibel scale to indicate colors.

Our technical plan involves segmenting our audio files based on whether the amplitudes are above a defined threshold. This will provide a systematic methodology for determining at which times in the audio a frog can be heard calling. Next, we will pre-process the segmented audio files so that the data all have the same sampling rate. This standardization will ensure that all the arrays have the same dimensions. We will then use the Librosa python library to convert the audio segments into Mel spectrograms and save the images. Following these methods will provide a set of images containing the spectrogram representations of frog calls from each species that can be used to train a convolutional neural network.

We have not yet decided on the architecture of our convolutional neural network, but we plan to determine a reasonable layer structure and loss function schema through trials informed by evaluations as well as additional research on prior models that have had success in spectrographic classification.

Evaluation Plan

We will evaluate our model through accuracy, precision, and recall metrics. Accuracy will give us a good overview of the success of our model, whereas the precision and recall metrics for each species will allow us to identify whether that species is being overrepresented or underrepresented by the model. Additionally, we will look at cross entropy loss, which will be a useful addition to looking at accuracy, especially if we proceed with using our full dataset, which is imbalanced. We also will test the model on different types of audio files than used for training, such as clips with more background noise, unsegmented audio files with multiple calls, and clips with multiple calls from multiple species to test the extent to which this method could be useful.

Citations

1. https://data.mendeley.com/datasets/5j852hzfjs/1
2. Terneux, Andrés Estrella, Damián Nicolalde, Daniel Nicolalde, and Andrés Merino-Viteri. "Presence-absence estimation in audio recordings of tropical frog communities." *arXiv preprint arXiv:1901.02495* (2019).
3. Incze, Agnes, Henrietta-Bernadett Jancsó, Zoltán Szilágyi, Attila Farkas, and Csaba Sulyok. "Bird sound recognition using a convolutional neural network." In *2018 IEEE 16th international symposium on intelligent systems and informatics (SISY)*, pp. 000295-000300. IEEE, 2018.