Are You Seeing What I'm Hearing?

Convolutional Neural Network Model's Classification of Spectrogram Images

Sammy Gagou, Joseph Ross, Nadine Hughes, Julia Guzzo, Aneesh Sallarm

Abstract

This paper explores the results of the study to see how effectively a convolutional neural network (CNN) model can be trained to analyze a spectrogram image of an audio recording and classify the species of animal that made the sound. The dataset was collected from a public GitHub repository and consists of 875 animal sounds and is split into 10 different types (cat, dog, bird, cow, lion, sheep, frog, chicken, donkey, and monkey).

Introduction & Motivation

Spectrogram classification is useful for a variety of reasons, such as, but not limited to, biodiversity monitoring, animal behavior studies, taxonomic research, and conservation efforts. For example, with more and more species being placed on protected and endangered lists, scientists need a way to track the location and population levels of these species accurately. While we can record audio in an attempt to perform this research, it is not possible to invest the manpower to analyze enough audio to produce statistically meaningful results. This is where the possibility of AI and model training comes in. We decided to research a proof of concept that a model could be effectively trained to classify Spectrogram images and classify what animal produced it. Different models and techniques will have different levels, so being able to tune/use models for different use cases improves efficiency in pursuing/supporting these efforts.

Previous classification models for acoustic data have traditionally converted audio waveforms to images of spectrograms, which has proven to be successful for longer end-to-end audio scene classification. (Eghbal-Zadeh et al. 2016) Spectrograms are typically chosen for

these longer audio data because they can condense long-term and high-frequency audio data into a much smaller matrix of long term spectral averages (LTSA). However, for audio lengths on the order of 100,000 samples or less (~5 seconds at ~20kHz audio) a highly detailed spectrogram will contain roughly the same number of data points as the waveform itself.

Additionally, some bioacoustics research has begun to use spectral probability density graphs to identify the presence or absence of animals in a soundscape. By using microphones with onboard processing capabilities, high-frequency audio data can be effectively compressed even further into a matrix of size 9xN, where N is the Nyquist frequency (sampling rate divided by 2) of the audio recording. (Martin et al. 2021) Current research in AI audio classification has yet to incorporate this representation in the machine learning process, which may prove useful for data analysis on large-scale bioacoustic sampling efforts like SanctSound (NOAA).

Our goal in this experiment is to qualify the success of traditional machine learning models in learning the classification boundaries between different animal sounds using various audio representations.

Experimental Design

The first step in this experiment was to define the scope of the research by identifying the classification goal. For this, we chose to classify based on the available dataset of ten animal species. From there, we were able to begin looking for an appropriate dataset and we found one on GitHub that consisted of 875 distinct audio files of ten different animal sounds.

For the development of this model, we first had to do audio preprocessing to ensure that all of the audio samples had the same peak audio level and sampling rate. By normalizing the peak audio level, we can avoid the model underfitting the classification pattern based solely on volume. Normalizing the sampling rate of the audio allows all of the spectrograms to have the

same upper frequency bound, allowing all of the images being processed to have the same dimensions. Completing these two preprocessing steps ensures that all of the audio files have roughly the same fundamental acoustic properties, and eliminates many possible extraneous variables.

After we have the audio files standardized and converted into spectrograms using a short-time fourier transformation (STFT), we can choose an AI model. We have chosen to use convolution neural networks (CNNs) for image processing in this experiment because they are optimal for learning contour-based features of an image without the need for additional feature engineering, similarly to the function of human vision-recognition (LeCun et al 2015).

Then, we are going to split our data into training, validation, and test sets in order to train our model. At this point, we will evaluate our model on its ability to classify which of the species in the dataset the testing data belongs to (error percentage?).