Machine Learning Engineer Nanodegree

Capstone Proposal

Sam Hiatt
Aug 9, 2019

Background

Many social animals communicate using vocalizations that can give clues to their species as well as to their intent. Studies of animal vocalizations have been hindered by the cost required for manual expert analysis of audio signals. Machine learning offers tools than can automate classification of these audible signals, opening up countless opportunities for sound-aware computer applications and could help accelerate studies of these animals. For example, imagine a computer trained to recognize the call of a specific species of bird. This model could be used to trigger a camera recording, or automatically tag a live audio stream containing avian calls with the species of the bird that made it, producing a time-series record of the presence of this species.

Sonograms (spectrograms based on sound frequencies) are commonly used for visual representation of audio information and have long been used for interpreting recordings of animal vocalizations. Bird Song Research: The Past 100 Years describes how a device called the Sona-Graph™, developed by Kay Electric in 1948, began to be used by ornithologists in the early 1950's and greatly accelerated avian bioacoustical research.

By applying machine learning techniques for image classification on these sonograms, automated classification of audio clips is possible. The project DeepSqueak at the University of Washington in Seattle takes this approach for classifying recordings of ultrasonic vocalizations of rodents. Their publication in Nature, DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations, uses this classifier to prove correlations between specific behaviors and types of vocalizations.

Problem Statement

This project will build a classifier for labeling audio recordings of avian vocalizations. The classifier will take a digital audio recording as input and will return a single label representing the most prevalent recognized species in the recording.

Datasets and Inputs

Xeno Canto is an online community and database of crowd-sourced recordings of birds from around the world with Creative Commons licensing, indexed by species and labeled by call type. This project draws on inspiration from The British Birdsong Dataset, which includes 264 recordings from 88 species commonly heard in the United Kingdom. In order to procure a dataset with sufficient samples for a robust analysis, the Xeno-Canto Avian Vocalizations CA/NV, USA dataset was compiled from a selection of recordings obtained by scraping Xeno Canto. It contains 2730 samples from 91 different species (30 recordings of varying length for each species) each recorded in California or Nevada, USA.

Solution Statement

Drawing on insights from the project DeepSqueak, this effort will apply a CNN architecture trained on sonograms generated from the audio files in the Xeno-Canto Avian Vocalizations CA/NV, USA dataset. It is expected that the temporal invariance introduced through convolution and max pooling will achieve improved results compared to models that do not preserve the temporal relationship between input features.

Benchmark Model

A Gaussian Naive Bayes classifier will be trained on flattened sonograms and used as a benchmark model for the classifier.

Evaluation Metrics

sklearn.metrics.accuracy_score will be used to evaluate the overall accuracy of the classification model. It is simply defined as the number of correctly labeled samples divided by the total number of samples. During model selection and hyperparameter tuning 5-fold cross-validation will be used. Final test accuracy will be evaluated against a subset kept aside from the beginning to be used as a test set.

Project Design

First, analyze the chosen dataset verify the distributions of represented classes and report the length of each audio file. Create sonograms and analyze statistics. Determine the main frequency ranges. Determine the mean amplitude of each frequency band.

Optionally, follow the approach taken in Edoardo Ferrante's Kaggle kernel to generate sonograms and filter out the quietest frames in each sample to remove any silence.

Create a stratified 5-fold cross-validation training / test split with 1/3rd of the dataset reserved for testing.

Create a data generator for creating fixed-length clips from a random window in each sample.

Use these data generators to train a Naive Bayes classifier. Report test accuracy and use as a benchmark result.

Now try to improve on the benchmark by implementing a CNN-based classifier. Try several different configurations of CNN architectures. Try settings that only apply convolution and max pooling along the temporal dimension.

Produce a confusion matrix to see which classes are most difficult to distinguish. Are commonly confused species closely related? Or is presence of multiple species in in the recordings confusing the classifier?

Evaluate the test accuracy of the best-performing model and compare to the benchmark results.

References

- Baker, Myron C. (2001). Bird Song Research: The Past 100 years (PDF). Bird Behavior. 14: 3–50.
- Kevin R. Coffey, Russell G. Marx & John F. Neumaier (2019) DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations Neuropsychopharmacology vol 44, pp 859–868.
- Xeno Canto, founded by Bob Planqué and Willem-Pier Vellinga.
- Edoardo Ferrante. Extract features with Librosa, predict with NB (2019) Kaggle kernel. Version 10.