Kaggle Competition Submission

Freesound General Purpose Audio Tagging Challenge

"You're challenged to build a general-purpose automatic audio tagging system using a dataset of audio files covering a wide range of real-world environments. Sounds in the dataset include things like musical instruments, human sounds, domestic sounds, and animals from Freesound's library, annotated using a vocabulary of more than 40 labels from Google's AudioSet ontology. To succeed in this competition your systems will need to be able to recognize an increased number of sound events of very diverse nature, and to leverage subsets of training data featuring annotations of varying reliability"

Data Preparation

Raw audio files contain time-domain information only, and so there exists a semantic gap between our perceptual experiences of sound, which are based on frequency, and the data to be found in a .wav file. In order to extract relevant information from the training set of audio files, these had to be first converted into the frequency domain. This is achieved by performing a Fast Fourier Transform on the audio signal, which generates phase and amplitude information for each individual frequency contained within the signal. Using this information, a spectrogram can be created, which shows amplitude as a function of frequency over the course of the entire signal. As the signals were short in duration, it was not necessary to window them, as would be the case in a Short-Time Fourier Transform in order to determine how the frequency content of the signal is changing over time.

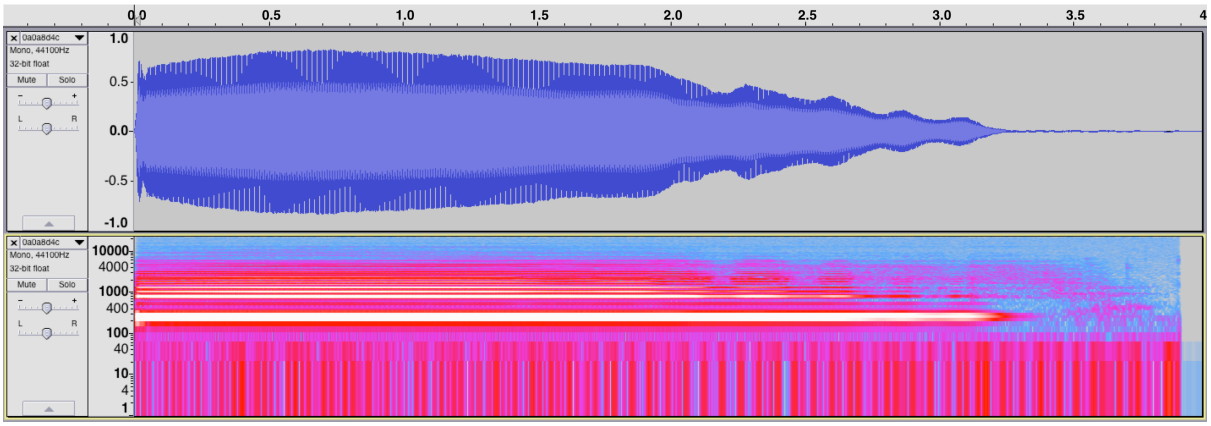

Figure 1: Top - waveform time-domain representation of an audio signal from the test set. Bottom - spectrogram frequency-domain representation of same

Although a traditional logarithmic scaling of the frequency spectrum is an effective representation of how the sound may be perceived, using the *mel* scale may result in a more perceptually salient representation. The mel scale is closely related to the logarithmic scale but divides the frequency spectrum in terms of how pitches are perceived, with increased distances for intervals at higher frequencies and decreased distances for intervals at lower frequencies. Thus, taking the power spectrum of the signal with reference to the mel scale results in a more perceptually accurate representation of a sound. This is done by deriving the Mel Frequency Cepstral Coefficients of the signal.

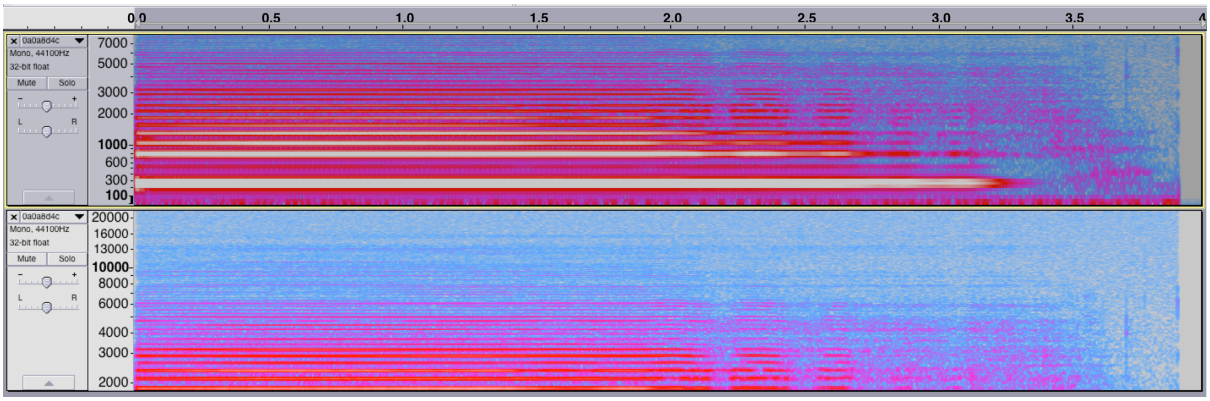

Figure 2: Top – mel frequency spectrogram of an audio signal from the test set. Bottom - linear frequency spectrogram

Using the librosa audio analysis library in python, 2 descriptors of MFCC values (mean and standard deviation) were generated over 20 windows for each audio signal. A data frame containing 40 MFCC values for each audio signal was imported into R. This data frame was then split into a train and test set (60/40) which were used to refine the Random Forest model.

Tuning the Random Forest model

Within R, the caret package was used to create and evaluate the Random Forest model. Random forest models contain hyperparameters, which must be set as inputs to the predictive model. The most important of these hyperparameters is the *mtry* value, which represents the amount of variables to be considered at each split point. In the *ranger* random forest algorithm in the caret package, three random *mtry* values are generated by default, and the resulting accuracy for each *mtry* value can be viewed in order to determine a new range by the user. The *mtry* value with the highest accuracy determines the final model. The *tunelength* hyperparameter allows the user to determine how many *mtry* values should be generated. It was found through trial and error that the model on this particular dataset had maximum accuracy with an mtry value of 21. 5-fold cross validation was

Evaluation

In the caret package, evaluation metrics are available within the *train* function.

Kaggle Results

The score achieved in the competition by the model was 0.684. This score seemed sufficient for the limited nature of Random Forests in comparison with some of the neural network systems and advanced signal processing techniques contained in some of the kernals submitted by other competitors.

Future Work

Although MFCCs are sufficient for determining salient features of an audio signal, the inclusion more spectral features, such as spectral centroid values and zero-crossing rate may yield better classification results. In addition, more time to fine tune the parameters of the Random Forest algorithm may have resulted in the algorithm having a better fit to that particular type of input. The competition evaluation metric was Mean Average Precision @ 3 (MAP3), meaning that the average of three best predictions were taken, before the average of the entire data set was taken, however only one prediction was made for each instance. With more time, more predictions could have been submitted in the required format.