Jamie McQuire
CSC8635
J.Mcquire2@newcastle.ac.uk

Deep Learning for Automatic Speech Recognition: Structured Abstract and Key Images

2020/01/24

1 Abstract

Automatic speech recognition is an important area of research in natural language processing that focuses on the classification of human voice signals with computational algorithms. The recognition of human voice commands offers potential benefits in developing future devices for smart homes and the assistance of people with visual impairments who might struggle to operate digital devices not tailored to them. This project explores the use of deep learning techniques for the classification of a vocabulary of human voice commands, providing a comparison between the performances of different types of deep learning models for audio recognition. The audio files used in this report are taken from Google's speech commands dataset, and the deep learning models are trained to classify a subset of 6 words and the full vocabulary specified by the Kaggle competition. The voice data was transformed into log-spectrograms allowing for the classification of the audio signals to be done using computer vision algorithms. The performances of the models were evaluated based on the correct classification of each command, and if the models could correctly identify when an audio signal contained silence. We found that the ResNet model achieved the highest accuracy for both sets of words, with the model achieving an accuracy of 84.6% on the competition's testing set. This work expanded on existing papers that treated the audio recognition problem as an image recognition problem by using both deep residual learning and recurrent models for image classification, comparing the results to the basic convolutional neural network model. The use of other well established deep learning models could improve the performance of the audio recognition model and should be investigated in future research.

2 Key Images

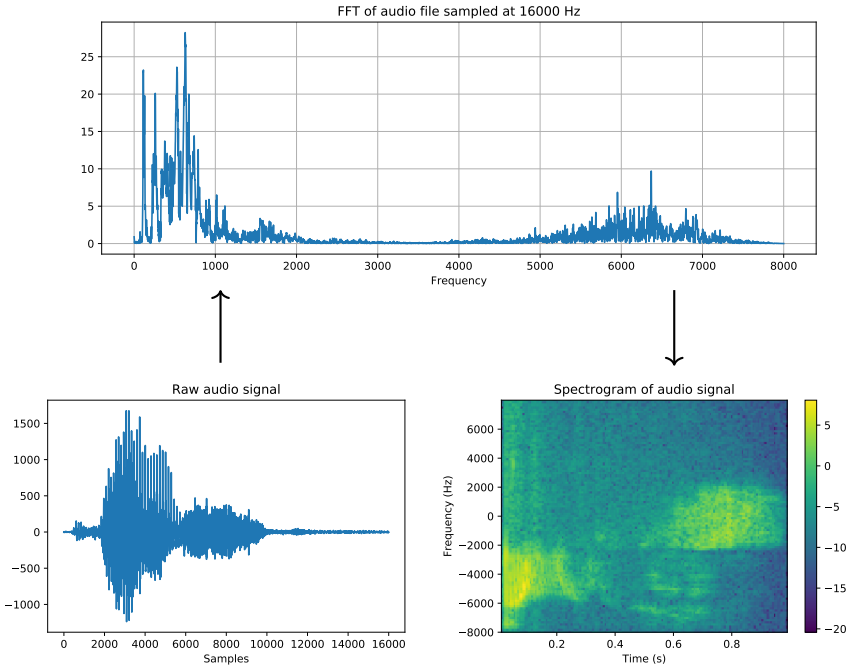
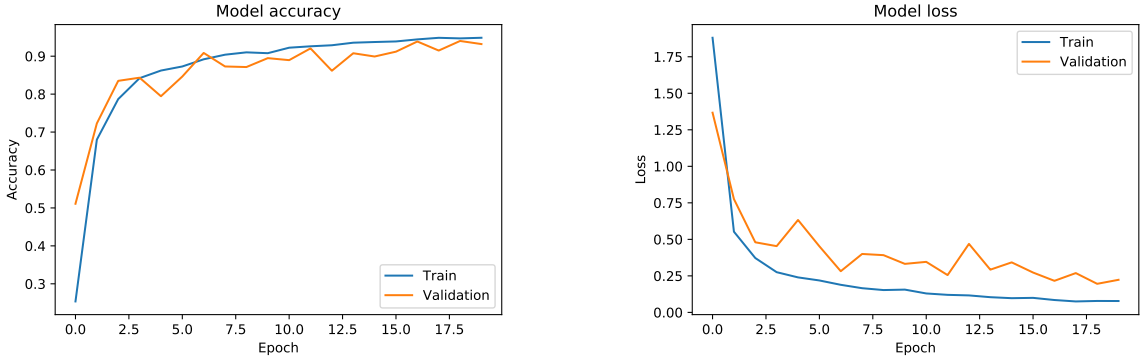


Figure 1: The data transformation process mapping the raw audio signal to a spectrogram.



(a) Accuracy of the training and validation sets.

(b) Loss of the training and validation sets.

Figure 2: Plot of the accuracy and the loss for the ResNet model trained on the full competition dataset.