ECE 114: Speech and Image Processing Design

Class Project

Due: Tuesday, December 8th, 2020 at 11:59pm PT

To be completed in groups of up to 4 students

In this project, we will explore machine learning for speech and image processing. The given task is to perform speech recognition on a subset of the Google Speech Commands Dataset . The data can be accessed here. You will access it through google drive and google collab to avoid having to download the set and setup the environment on your own computer. Download the project python notebook from CCLE and add a copy of it to your google drive. Then open it in google colab to run it. Make sure that the data is searchable through your google drive path. This data subset contains the spoken digits zero to nine. Our goal is to create a neural network pipeline that will be able to classify which digit was spoken given a 1 second recording. In doing so, we will explore two approaches.

1. Feature extraction – we will extract some feature like the DFT, LPCs, or MFCCs from the signal at each frame. Once concatenated across all frames, those features will form a spectrogram-like image. We can then perform image recognition on the spectrogram.
2. End to End model – we will feed the raw audio signal into the neural network without prior feature extraction

The data processing steps and neural networks for these approaches are given to you. However, they are trained on clean speech signals taken in ideal recording conditions. In real life scenarios, we rarely obtain such high-quality audio recordings. We would normally expect there to be some degradations due to outside noise, disadvantageous microphone placement, overlap from multiple speakers, incomplete utterances, and other phenomenon. In this project, we will deal with one common such signal degradation: reverberation. It is common for a microphone to not only pick up the speaker's original utterance but also echoes of that utterance reflected off surrounding surfaces. This can be thought of as attenuated and delayed copies of the signal being added to the original before it reaches the receiver. Reverberation is also often called convolutional noise since this effect can be modeled through convolving the clean signal with the impulse response of a room to generate an echo. In this project, we create reverberated signals from the clean ones and see that the speech recognition system does significantly worse with the reverberated signals. You task is to pre-process the reverberated signals being input into the network so as to improve the performance of the system. Follow the instructions provided in the python notebook. Your group must implement 5 distinct preprocessing methods to attempt to improve the performance of the image classification network and 5 distinct preprocessing methods to attempt to improve the performance of the end to end model. You are not graded by the actual performance of then network but rather by the novelty of your approaches and how much knowledge of speech, image processing, and DSP your approaches show. Submit a report of up to 4 pages (excluding references) justifying your approaches. You may choose to use a conference paper template from a relevant conference like ICASSP or ICML. Be sure to include a clear table of results showing the performance of the network for each approach.

The grading will be as follows:

10 distinct approaches used – 5pts

Justification of approaches – 5pts

Shows knowledge of class concepts – 5pts

Clear and well written report -5pts

Total – 20 points

Include the report and all code in your submission. You do not need to code every method you wish to implement yourself. You may use python and MATLAB functions from trustworthy sources including the Numpy, Scipy, Sklearn, Liborosa, and Tensorflow libraries in Python and the MATLAB Voicebox and DSP toolboxes.