← Go Back to Deep Learning

Course Content

Session Problem Statement - Audio MNIST Digit Recognition

Context

In the past decades, significant advances have been achieved in the area of audio recognition and a lot of research is going on globally to recognize audio data or speech using Deep Learning. The most common use case in this field is converting audio to spectrograms and vice versa.

Audio in its raw form is usually a wave and to capture that using a data structure, we need to have a huge array of amplitudes even for a very short audio clip. Although it depends on the sampling rate of the sound wave, this structured data conversion for any audio wave is very voluminous even for low sampling rates. So it becomes a problem to store and computationally very expensive to do even simple calculations on such data.

One of the best economical alternatives to this is using spectrograms. Spectrograms are created by doing Fourier or Short Time Fourier Transforms on sound waves. There are various kinds of spectrograms but the ones we will be using are called MFCC spectrograms. To put it in simple terms, a spectrogram is a way to visually encapsulate audio data. It is a graph on a 2-D plane where the X-axis represents time and the Y-axis represents Mel Coefficients. But since it is continuous on a 2-D plane, we can treat this as an image.

Objective

The objective here is to build an Artificial Neural Network that can look at Mel or MFCC spectrograms of audio files and classify them into 10 classes. The audio files are recordings of different speakers uttering a particular digit and the corresponding class to be predicted is the digit itself.

Dataset

The dataset we will use is the Audio MNIST dataset, which has audio files (having .wav extension) stored in ten different folders. Each folder represents the digit or the class to be predicted by the Artificial Neural Network.

Previous

Next