Deep Learning Group 1 Final Project

PROJECT PROPOSAL

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset is a public dataset of emotional speech and song recordings. It was developed by the Ryerson University's Multimedia Laboratory in Toronto, Canada.

The dataset contains a total of 24 professional actors (12 male and 12 female) who have each recorded two repetitions of 48 different scripts in 8 different emotions (calm, happy, sad, angry, fearful, surprise, disgust, and neutral). The scripts include short statements, such as "Kids are talking by the door", and "The dogs are sitting by the door". The sentences are recorded in different tones from all the actors.

We are using which model out of CNN, RNN and LSTM can perform better in the dataset. Historically, for audio sets especially speech data, numerous papers available till date rely on convolutional or CNN models with majority giving an accuracy of 80-85% and some that are provide a 92% accuracy, for eg. Tim-Net Model. As most of these models are CNN models, the goal of the project is to do a comparative analysis of how the RNN and LSTM models could perform against the CNN models.

So, we will be comparing CNN, RNN, LSTM with each other and will be comparing how their architecture differs and works for the audio speech recognition using RAVDESS dataset.

We will also be investigating why a certain model might outperform the other for the dataset.