Deep Learning Final Project

Adhithya Kiran | Xiao Qi | Akshath

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 "It's eleven o'clock", as well as longer paragraphs, such as the story of "The boy who cried wolf".

We are using the RAVDESS emotional dataset to classify and understand which model out of CNN,RNN and LSTM perform better in the dataset. We are aware that numerous papers available till date rely on CNN with a pretrained model giving an accuracy of 80-85% and the highest is 92% with Tim-Net. As most of the models are CNN models, we are examining how the RNN and LSTM works.

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

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