Deep Learning Project
- Speech Emotion Recognition

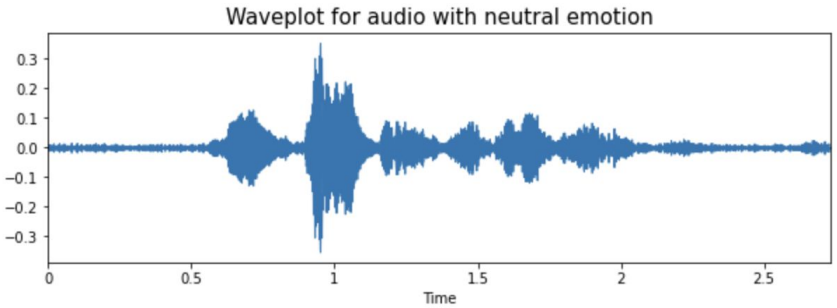
Zhipeng Hong Kaihang Zhao

Motivation

- Understanding emotion behind audio could be widely used to understand and analyze customer or user needs

- Audio data can be parsed into two ways (CNN, RNN) to capture patterns for different classes

- Creating algorithms that processes audio sounds and understands their inherent meanings



Speech Emotion Recognition

- Data Source: Crowd Sourced Emotional Multimodal Actors Dataset

  01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful

  Around 7k audios in total; Each around 3s

  https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en

- Techniques we consider:

  CNN-base, RNN-base, Transformer-base

- Data Preprocessing

  Reference on Kaggle notebook to sample and extract features as a 1D vector with 848 dimensionalities.

CNN

1D-CNN:

- **Used same structure as AlexNet and ResNet. Since we convert audio to array, we need to use 1d convolution layer.**
- **We follow the same structure as AlexNet and ResNet, then we rebuild the model from scratch and turned the hyperparameters.**

Result:

Model	Learning Rate	Weight Decay	Best Test Accuracy
ResNet	1e-4	5e-4	0.9
AlexNet	1e-4	5e-4	0.8

Sequence Model

- **We built three sequence deep learning model, Bidirectional GRU, Bidirectional LSTM and Attention Model.**

- **Because the stack and bidirectional models' training speed is quite slow, we only trained for 25 epochs.**

Model	Learning Rate	Final Test Accuracy
GRU	0.005	0.41
LSTM	0.005	0.53
Attention	0.005	0.63

Well & Not Well

Well:

- 1D-CNN model performs quite well on the audio classification. This would be an example to prove that 1D-CNN can also apply into prediction of sequence data.

Not well:

- Dependent on only one kind of feature extraction techniques; can explore more methods
- Can play around more hyper parameters and train more epochs to get better results

Thanks for listening!