DEEP LEARNING TECHNIQUES - CS6005 PROJECT

Multimodal Speech Emotion Recognition System

Gautham Kuman G	2020103521
Praveen Kuman R	2020103036
Nirubama A	2020103030

Need for the System

Recognising emotions is essential for customer service, mental health evaluation, human-computer interface, and other fields.Emotion recognition from textual and audio sources improves user comprehension and experience. The multimodal strategy uses many data sources to increase resilience and accuracy. One popular approach for identifying emotions via voice signals is speech emotion recognition. Through a multimodal approach, the system seeks to increase accuracy. When compared to unimodal approaches, the multimodal system obtains a remarkable accuracy rate of 98%. For text emotion identification, it makes use of the BERT model, while for audio emotion recognition, it makes use of the AlexNet model. This system provides an in-depth understanding of emotions by merging textual and auditory information.

Objective

- Accurately identify and categorise a broad spectrum of human emotions.
- Use a multimodal strategy that combines textual and audio data to capture complex emotional expressions.
- Increase emotion recognition accuracy beyond what can be achieved with unimodal methods.
- Employ sophisticated models to improve performance, such as AlexNet for audio emotion detection and BERT for text emotion identification.

Dataset Description

IEMOCAP (Interactive Emotional Dyadic Motion Capture)

The IEMOCAP dataset (Interactive Emotional Dyadic Motion Capture Database) is a multimodal emotion recognition dataset that contains audiovisual data, including video, speech, motion capture of face, and text transcriptions.

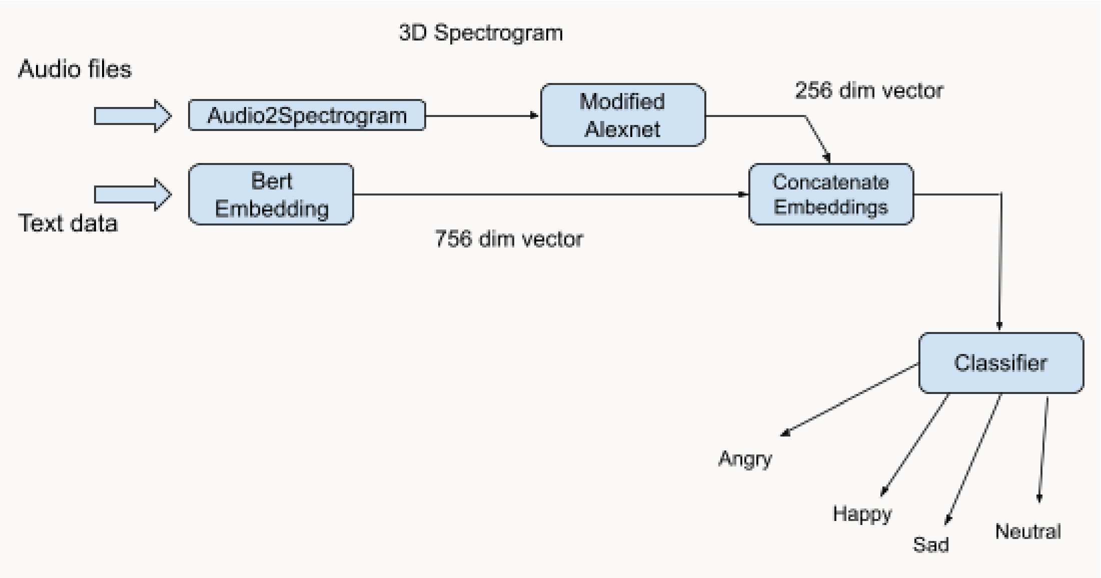
It is a valuable resource for developing and evaluating new machine learning algorithms for emotion recognition, and it has been used in a wide range of studies, including:

- Automatic emotion recognition from speech and video
- Modeling the relationship between emotions and facial expressions
- Understanding how emotions are expressed in dyadic interactions
- Developing new methods for human-computer interaction that take into account emotions

The IEMOCAP dataset is available for free download to academic researchers. It can be downloaded from the IEMOCAP website: https://sail.usc.edu/iemocap/

The dataset is split into five sessions, each with five pairs of speakers. Each session contains approximately 12 hours of audiovisual data. The dataset contains a total of 12,000 utterances, each labeled with one of the nine categorical emotions and the three dimensional labels.

System Design Diagram



Layer Architecture

<u>Speech emotion recognition</u>

AlexNet Architecture

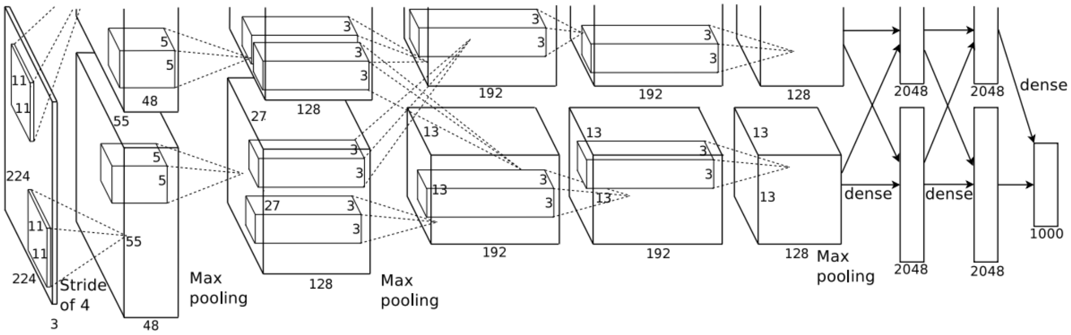
AlexNet is a convolutional neural network (CNN) architecture that was designed for image recognition tasks. However, it can also be used for speech emotion recognition (SER) by converting speech signals to spectrograms and then feeding the spectrograms into the network.

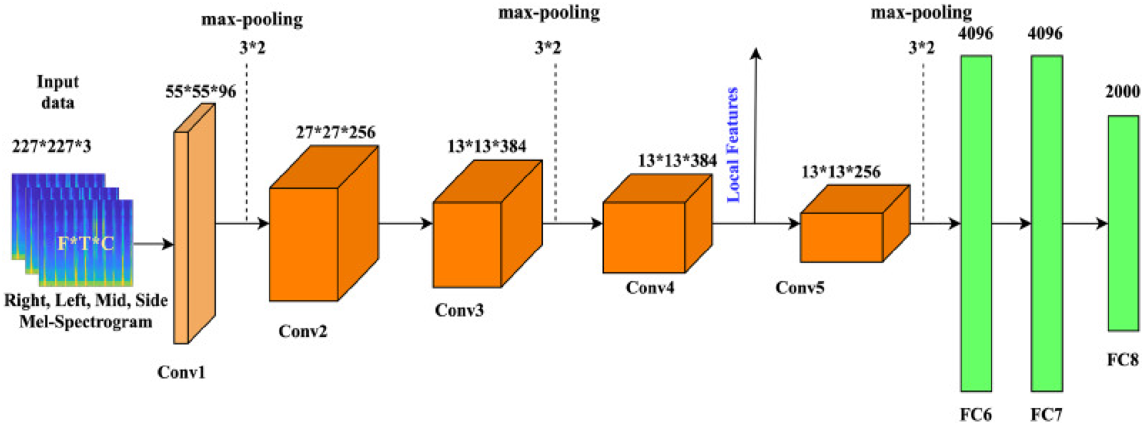
To use AlexNet for SER, you would typically follow these steps:

1. Convert the speech signals to spectrograms.

2. Extract features from the spectrograms using AlexNet.

   Train a classifier on the extracted features to predict the emotion of the speake





Resnet - Speech emotion recognition

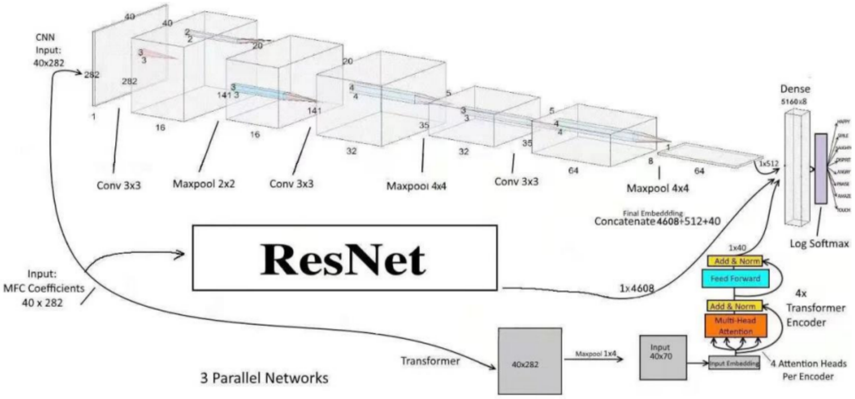
ResNet, or residual neural network, is a type of convolutional neural network (CNN) that is well-suited for speech emotion recognition. CNNs are a type of machine learning model that is particularly good at learning spatial patterns in data. Speech signals can be represented as spectrograms, which are visual representations of the frequency and intensity of the sound over time.

ResNets are able to learn deep representations of spectrograms by using a technique called residual learning. Residual learning allows the network to learn the differences between consecutive layers, rather than having to learn the entire representation from scratch. This makes ResNets more efficient and easier to train than traditional CNNs. Steps for using ResNet-

1. Convert the speech signals to spectrograms.

2. Extract features from the spectrograms using ResNet.

3. Train a classifier on the extracted features to predict the emotion of the speaker.

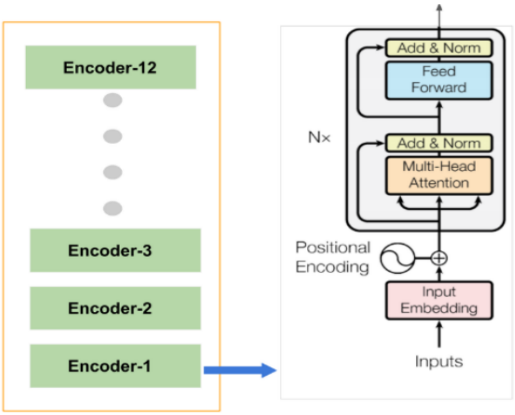


Bert - text based emotion recognition

BERT (Bidirectional Encoder Representations from Transformers) is a powerful natural language processing model used for understanding contextual relationships between words in text.

For text-based emotion recognition, BERT's encoder, with its attention mechanism, is typically employed. It's trained using a masked language model approach and can capture nuances in language, making it a valuable tool for recognizing emotions in text data.

BERT's architecture includes 12 layers with multi-head self-attention mechanisms, and it uses input embeddings combining token, segment, and position information to understand the emotional context of text.



Working of model with dataset

Working of AlexNet with IEMOCAP dataset

The IEMOCAP dataset contains audiovisual data, including video, speech, motion capture of face, and text transcriptions. For speech emotion recognition, only the speech data is needed.

Extract spectrograms from the speech data: Spectrograms are visual representations of speech signals that show the frequency and intensity of the sound over time.

Preprocessing spectrograms: This may involve resizing the spectrograms to a consistent size, normalizing the pixel values, and adding data augmentation to improve the robustness of the model.

Fine-tune the AlexNet model on the prepared spectrograms: The AlexNet model has been pre-trained on a large dataset of images, so it is important to fine-tune it on the IEMOCAP dataset to learn the specific features of speech signals that are relevant to emotion recognition.

Once the model has been fine-tuned, it can be evaluated on the IEMOCAP test set to see how well it performs on unseen data.

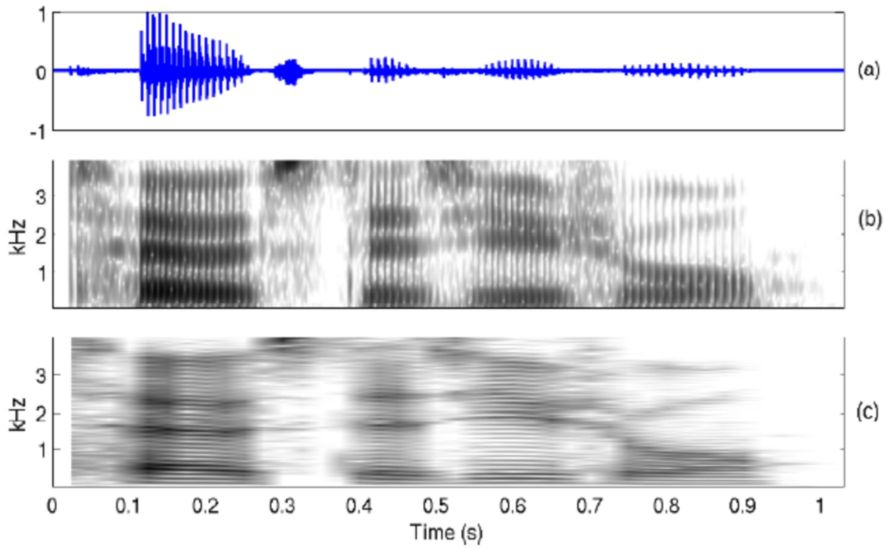
Working of Bert with IEMOCAP

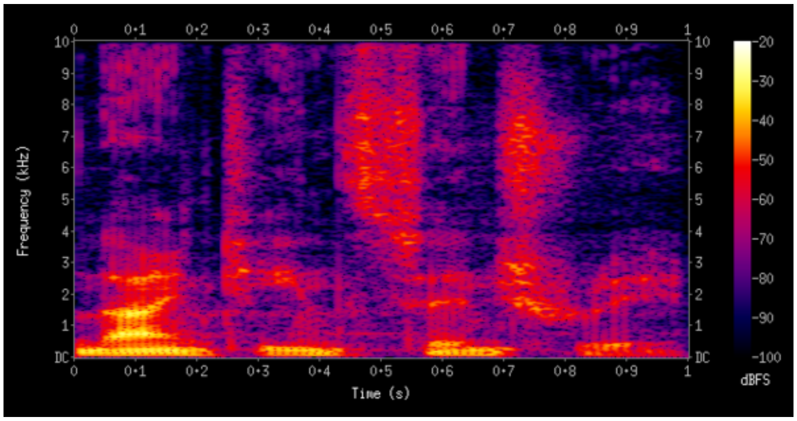
The IEMOCAP dataset is preprocessed by removing punctuation and stop words, converting the text to lowercase, and stemming or lemmatizing the words.

It is fine-tuned on the preprocessed IEMOCAP dataset. This involves updating the model's parameters to learn the specific features of the IEMOCAP dataset that are relevant to text-based emotion recognition.

The fine-tuned BERT model is evaluated on the IEMOCAP test set. This involves feeding the test set sequences into the model and getting the predicted emotions. The predicted emotions are then compared to the actual emotions to calculate the accuracy of the model.

Spectrogram Images





Working of Multimodal data with IEMOCAP

Multimodal emotion recognition (MMER) is the task of recognizing emotions from multiple modalities, such as audio, video, and text. For the proposed work, we combine the Bert(Text based) and Alexnet (Speech based) model that will run with IEMOCAP dataset.

To use AlexNet for speech-based emotion recognition and BERT for text-based emotion recognition on the IEMOCAP dataset, you would typically follow these steps:

1. Preprocess the data: Involves converting the speech signals to spectrograms, resizing the video frames (if available), and cleaning the text data.

2. Extract features from each modality.

   ○ For speech, use AlexNet to extract features from the spectrograms.

   ○ For text, use BERT to extract features from the text data.

3. Fuse the features from each modality: Done by concatenating the features from each modality or by using a more sophisticated fusion technique, such as a multimodal neural network.

4. Train a classifier on the fused features to predict the emotions of the speakers.

RESNET34 will be tried in addition to AlexNet (for audio based) with the same multi model configuration.

Tools and software requirements

- Kaggle
- Pytorch
- Python
- Tensorflow
- Colab