Name: Divyansh Anand

Student ID: 22240217

Question-Answers

1) How are the CNN and BiLSTM models appropriate choices for emotion detection? What is the complexity of the employed models?

Answer: The models CNN and BiLSTM have several strengths to handle speech data to recognize the emotions from the speech. CNN are effective at extracting local patters and features from the data. This is useful in analyzing audio data to pick-up important audio characteristics and features such as pitch, tone, and rhythm which are important to distinguish different emotional states from the speech data. CNN can also learn from the spatial features in data which means they have the capability to learn complex patterns by stacking convolutional layers. This is helpful for processing audio data as different frequencies have data related to different emotion states. LSTM models are designed to work with sequences of various lengths which makes it effective to analyze data audio files. By using the Bidirectional LSTM, data is processed in both forward and backward direction which takes the entire context of the audio input and makes it better at recognizing emotions regarding surroundings and patterns of voice in the audio input. BiLSTM networks are also capable to learn long term dependencies which is important in analyzing speech data considering emotions may change during the speech. This is crucial for emotion recognition given the fact that state of emotion is dynamic over time and may change slowly over dialog or speech.

The hybrid model used in this research has combined two branches 1D CNN branch and BiLSTM-Transformer branch which parallelly processes the feature file. The first branch 1D CNN is used to capture local patterns in the input audio by using various filters and kernel sizes. It has different layers for different functionality, max pooling layers helps in reducing dimensions of the feature maps which reduces computational complexity and prevents overfitting during processing. The BatchNormalization layers are used after every Conv1D layer in order to maintain and enhance the training process. Another important layer which is dropout layers is used to prevent overfitting. Flatten layer is used to covert the 3D output of a convolution layer into one-dimensional vector.

The computational complexity of each Conv1D layer is, O(n x k x m) where n represents the number of input samples, k represents size of the kernel and m represents the size of the filters. Since max pooling layer perform simple reduction, they have minimal complexity. Whereas, BatchNormalization layers linearly add complexity of the model depending on the size of the input data.

In the case of second branch, which is combination of BiLSTM and Transformer models. BiLSTM Layer incorporates forward and backward context in the sequence data to the model and simplify dependency thus improving understanding of context. Transformer block includes Multi-Head Self-Attention layer to capture the pattern between sequences of inputs and the Feed-Forward Neural Network layers to enhance the features required by the model. The computational cost of LSTM layer is O(n x d^2) where n is the length of the input sequence, d is the number of LSTM units, as it is bidirectional, actual complexity of the BilSTM will be O(2 x n x d^2). Since, Transformer block is complex as compared to CNN and BiLSTM, it has computation complexity of O(n^2 x d) where n is the sequence length and d is the dimensionality of the embedding. This complexity is multiplied by the number of heads in the multi-head attention and depth of transformer layer.

2) It's good to see testing with two datasets. What was the motivation for choosing CREMA-D over Emo-DB?

Answer: The motivation for choosing CREMA-D over Emo-DB dataset for speech emotion recognition in this research was mainly due to the diversity and size of CREMA-D. CREMA-D has high volume of data with 7,442 audio clips of 91 actors of different ages, races, and ethnicities. All such variety offers a wider and more detailed representation of emotions in speech, which is important for the emotion detection analysis. Furthermore, the multi-modal nature of CREMA-D, which includes both face and voice emotions, is beneficial as compared to other databases. This enables understanding of emotion from a broader perspective and may also enhance the efficiency of emotion recognition from speech. In contrast, Emo-DB dataset has 576 recordings from 10 actors which does not offer that much diversity of the speakers and emotional expressions. Additionally, Emo-DB has very less amount of data which is not enough to test the deep neural network model's efficiency. Hybrid model's usually get overfit when the dataset size is very small. By using CREMA-D, we can enhance the capability of the model to be generalized across different speakers and demographic, which makes it a better option for developing speech emotion recognition systems.

3) Considering the preprocessing steps, how effective is the approach by McFee et al. (2020) for removing unwanted portions of speech, followed by noise injection, time stretching, and shifting, in the context of CNN and BiLSTM models?

Answer: In the preprocessing phase, effects.trim function from librosa was used to remove non-speech segments from the audio. Removing non-speech segments from the audio data is important as it introduces noise and impact the performance of the emotion detection systems. By removing non-speech segments model can focus on the important parts of the audio data and can learn emotion related information effectively.

CNN models are great at extracting local patters and features from the data which is important in analyzing audio data to understand important audio characteristics and features such as pitch, frequency, tone, and rhythm. These voice characteristics are important to distinguish different emotional states from the speech data. Using noise injection techniques audio samples are created by injecting random noise to the original audio clips. CNN models by analyzing noise injected audio instances become more robust in handling real-time scenarios where background noise might be present in the audio clips. Time stretching technique changes the speed of audio without changing the pitch of the audio. CNN model can benefit from time stretching technique as it allows the model to be trained on various speaking rates. Some people speak quickly when excited or slow when sad, by training the model on time-stretched data helps CNN to learn and recognize emotion related information irrespective how fast or slow speaker speaks. This generalization of the speaking rate of the model makes it more robust and able to provide the right emotions across the various speakers and environments.

Bidirectional Long Short-Term Memory (BiLSTM) networks incorporated in the architecture have capability to work with sequences and capture long dependencies in speech. BiLSTM models are particularly valuable for speech emotion detection related tasks because they are capable to learn long term dependencies which is important in analyzing speech data considering emotions may change during the speech, phrases or sentences. In time shifting technique, the start time of the audio is shifted. This technique helps BiLSTM models to adapt small level of misalignment of the data. Time-shifted data help BiLSTM models to learn emotions when the audio may start at any random position and not at the exact starting point which makes the model more robust and useful for real-time scenarios. Time stretching has significant contribution in BiLSTM models because it

helps the model to capture temporal dependencies at various time scales. Speed of speech can be slow or fast depending on the speaker and emotional state of the speaker. This type of training on time-stretched audio makes the BiLSTM models adaptive to emotional patterns irrespective of how long it takes or how short it is. This flexibility allows the model to work with new data, different from the training data with variations in speaking rate.

Through the preprocessing techniques applied by McFee et al. (2020) for non-speech segment removal, noise insertion, time stretching, and time shifting the CNN and BiLSTM models become more robust to handle variations in actual speech data with background noise, variation in speaking rates, and varying starting points.

4) The proposed approach is not compared with any other methods. How well the model performs compared to LSTM and other transformer models is not unclear.

Answer: The main objective of this research was to check the performance of a hybrid model that integrates one-dimensional Convolutional Neural Network, Bidirectional Long Short Term-Memory and Transformer for robust speech emotion detection from features of speech. This proposed approach has been compared with previous research papers proposed by Kim and Lee (2023) which introduced a hybrid model approach by combining 2D CNN with BiLSTM-Transformer parallelly for emotion recognition from Ravdess dataset. [1] Another comparison was done with the study by Ullah et al. (2023) which used CNN with multi-head transformer model for analyzing Ravdess dataset. [2]

However, this approach was not directly compared to individual models which could provide better understanding of the model's effectiveness. Future work for this research could test this hybrid model against individual LSTM models and Transformer models. Only use of LSTM model could highlight the added value of CNN and Transformer models, especially for feature extraction and capturing sequence information. LSTMs are beneficial in capturing Temporal patterns but fail to capture local dependencies as that of a CNN and Long-range dependencies as that of a Transformer. Transformer models are very proficient when it comes to capturing patterns within the sequence to sequence tasks due to its self-attention mechanism. With the help of self-attention mechanism in transformers, the model can apply weights on the input elements and score them to differentiate and extract essential aspects of the input sequence. This capability further improves the modelling of relations with the environment and can be effectively applied to the problems related to sequences, such as natural language processing and speech-based emotion recognition.

Although, this research successfully evaluated the performance of the proposed architecture with other similar hybrid architectures, it lacks to evaluate the performance of each model individually. This can be done in future to test the contribution of each model (CNN, BiLSTM and Transformer) with the overall performance of the hybrid model to get better understanding of the approach used to detect emotions from the speech.

5) What is your research design, Is it qualitative or quantitative?

Answer: This research focuses on collecting the audio data from two different datasets which is Ravdess and Crema-D, processing it and later evaluating it based on the interpretations of the output provided by the proposed hybrid model integrating CNN, BiLSTM and Transformer. Both the datasets Ravdess and Crema-D have quantifiable audio data for each emotion category. Key audio features such as Zero Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCCs), and Root Mean Square Energy (RMSE) are extracted from raw audio files which is quantified. These features provide numerical input for the integrated model which further analyses the data and produces output. The model proposed in this research integrates layers of Convolutional Neural Networks, Bi-directional

Long Short-Term Memory Networks and Transformer and its performance is tested for both the given datasets. In this research, numerical metrics such as accuracy, recall, precision and F1-score are used to evaluate the performance of the proposed hybrid architecture. Overall, based on the quantitative data, feature extraction steps, and performance evaluation techniques of the model, this research is quantitative in nature.

6) What is meant by the neutral sample in your dataset? How did you deal with its imbalance as compared to other samples?

Answer: The neutral audio samples present in Crema-D and Ravdess datasets represents audio files which does not convey any emotion in the input as compared to other audio samples which convey emotions like happy, sad, angry, etc. The purpose of keeping the neutral emotion in an audio dataset is to use it as a baseline to compare performance of the model in detecting various emotion classes.

In the initial phase, in the raw data files the count of neutral emotion category for both the datasets was slightly higher in case of Ravdess (288) and slightly less in Crema-D (1090) datasets as compared to other emotion categories which had almost similar count of 192 for Ravdess and 1272 for Crema-D. To address this imbalance, this research has used data augmentation techniques which diversify and increase the data samples for the model. The data augmentation techniques used in this research include noise injection, time stretching, time shifting, pitch shifting which ensures that the data samples are balanced by creating more neutral samples. These techniques make the model more generalised and robust to perform better across all the emotion categories thereby decreasing the effects of imbalance in the dataset samples.

During the training phase, the hybrid model integrating CNN, BiLSTM and Transformer have balanced the emotion classes appropriately for the analysis. While evaluating the performance of the model across all emotion categories, we can see the f1score values for neutral emotion category is 77.36 % for Ravdess and 88.55 % for Crema-D datasets. These high values shows that the model performed well on this emotion class, even if there was an initial imbalance.

7) What is the difference between the adopted two datasets? Could you let me know why you don't combine them?

Ravdess dataset contains recording from 24 different actors which include 12 males and 12 females. This dataset included both audio and audio-video files, for this research only audio files were considered. This dataset includes audio files for seven emotion categories namely neutral, surprise, happy, angry, disgust, fear and sad. Whereas, Crema-D dataset include 7442 original audio clips from 91 actors which include 48 males and 43 females. It includes audio data of six emotion categories sad, angry, happy, fear, disgust and neutral. One major difference is that Ravdess dataset has an additional emotion category “Surprise”. Another difference, the count of emotion instances is 1440 for Ravdess and 7452 for Crema-D dataset which represents Crema-D has higher volume of data.

This research did combine the Ravdess and Crema-D datasets but faced few problems. First problem was after combining there was major difference in emotion categories count for Surprise emotion compared to other emotions, since “Surprise” emotion was just present in Ravdess and not in Crema dataset. To handle this, additional piece of code was required to eliminate “Surprise” emotion and balance the combined dataset. Another problem, the time complexity and computation resource requirement were very high as the entire code took more than 7 hours to complete as compared to individual datasets which was around 2 and half hours for Ravdess and 4 hours for

Crema-D dataset. Another problem was unexpected results for “sad” emotion. In the combined confusion matrix given below in figure 1, the number of correctly predicted "sad" instances is 138 which is significantly lower than the other emotional classes. This suggests, performance of the model is poor in classifying “sad” emotions when trained on the merged dataset which is abnormal to what we observed in the confusion matrix of Ravdess and Crema-D which is given in figure 2 and figure 3.

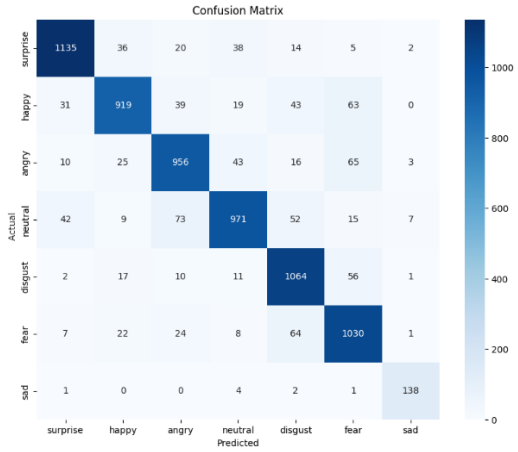

Figure 1: Confusion matrix of combined dataset (Ravdess & Crema-D)

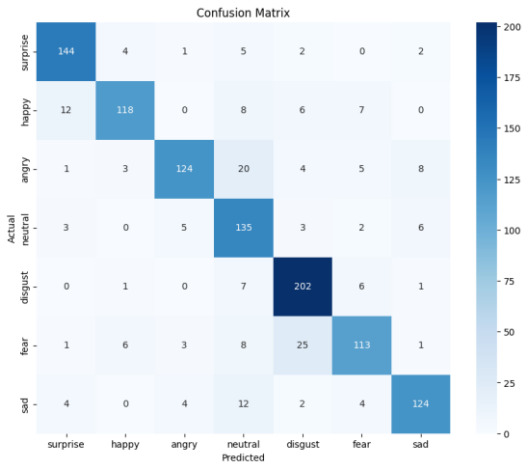

Figure 2: Confusion matrix of Ravdess dataset

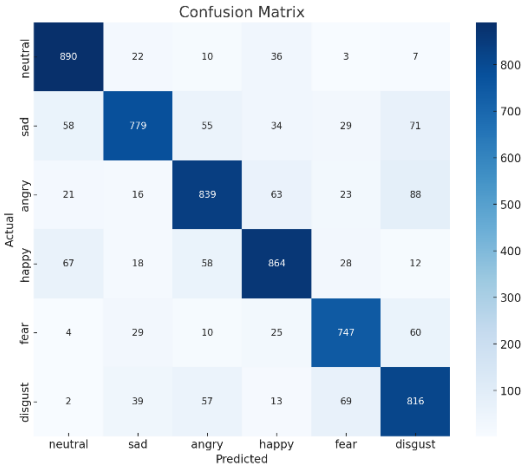

Figure3: Confusion matrix of Crema-D dataset

References:

[1] Kim, S. and Lee, S.-P. (2023). A bilstm–transformer and 2d cnn architecture for emotion recognition from speech, Electronics 12(19). URL: https://www.mdpi.com/2079-9292/12/19/4034

[2] Ullah, R., Asif, M., Shah, W. A., Anjam, F., Ullah, I., Khurshaid, T., Wuttisittikulkij, L., Shah, S., Ali, S. M. and Alibakhshikenari, M. (2023). Speech emotion recognition using convolution neural networks and multi-head convolutional transformer, Sensors 23(13): 6212.