PDF: EFFICIENT SPEECH EMOTION RECOGNITION USING MULTI-SCALE CNN AND ATTENTION

Kedar Kore
Roll no.- 210503

1 Introduction

This project implements a multi-modal emotion recognition system using both audio and text data from the IEMOCAP dataset. The model uses deep learning techniques such as MSCNN, attention mechanisms, SWEM pooling, and statistical pooling units, X-vectors, to predict emotions across four classes: *ang*, *hap*, *sad*, and *neu*.

2 Dataset Description

The IEMOCAP dataset consists of multimodal interactions, including audio and text transcriptions, annotated with emotions. For this project, we focus on the following emotions: anger (ang), happiness (hap), sadness (sad), and neutral (neu). Audio features were extracted using Mel-frequency cepstral coefficients (MFCCs), and text inputs were processed using GloVe embeddings.

3 Implementation Details

The model architecture is divided into two branches: an audio branch and a text branch.

3.1 Audio Branch

The audio features are obtained by MFCC. The MSCNN module is used to learn multi-scale features from audio data. Statistical pooling is used, and x-vectors are extracted to create high-level representations. The audio features are passed through completely connected layers.

3.2 Text Branch

The text input is analysed with GloVe embeddings. SWEM pooling and MSCNN are used to identify both sequential and non-sequential patterns. The attention layer generates context-aware features based on audio inputs.

3.3 Model Fusion

Audio and text features are concatenated and routed through dense layers with L2 regularization, including dropout and batch normalisation. The final result is a softmax layer that predicts the emotion class.

4 Training and Evaluation

The model is trained using categorical cross-entropy loss and the Adam optimiser. Class weights are used to address class imbalance. ReduceLROnPlateau and early stopping callbacks are used to optimise training.

4.1 Results

The model achieved the following results on the test set:

- Weighted Accuracy (WA): 78.57%
- Unweighted Accuracy (UA): 73.30%

5 Confusion Matrix

Figure 1 shows the confusion matrix for the test set, indicating the performance of the model across the emotion classes.

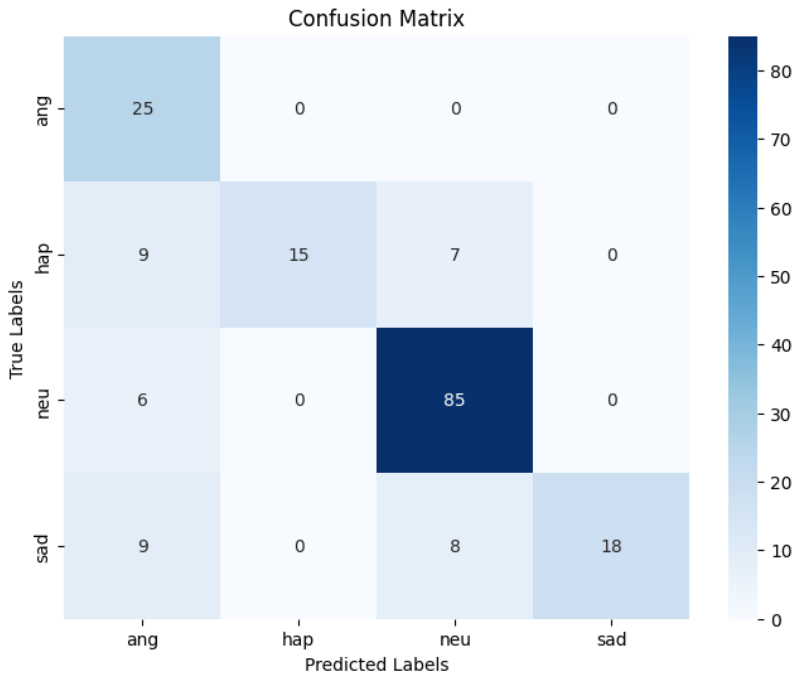

Figure 1: Confusion Matrix for Emotion Classification

6 Conclusion

This model demonstrates the effectiveness of merging audio and text features in emotion recognition. The model performed well on the IEMOCAP dataset due to its usage of MSCNN, statistical pooling, and attention processes, which helped catch subtle patterns in the data.