PDF: EFFICIENT SPEECH EMOTION RECOGNITION USING MULTI-SCALE CNN, SPU, ATTENTION AND MONTE CARLO DROPOUT

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 employs deep learning techniques such as Multi-Scale Convolutional Neural Networks (MSCNN), attention mechanisms, SWEM pooling, statistical pooling units (SPU), Attention mechanism and Monte Carlo Dropout 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

3.1 Audio Branch

The audio features were obtained through MFCCs. The MSCNN module was used to extract multi-scale features from audio data, combined with statistical pooling and x-vector representations. Monte Carlo Dropout was incorporated to improve model robustness and uncertainty estimation.

3.2 Text Branch

Text inputs were analyzed using GloVe embeddings. SWEM pooling and MSCNN modules were utilized to capture sequential and non-sequential patterns after which Monte Carlo Dropout layer was applied. An attention layer was added to generate context-aware text features based on audio inputs.

3.3 Model Fusion

Audio and text features were concatenated and passed through dense layers with L2 regularization, Monte Carlo dropout, and batch normalization. The final output layer used a softmax activation function to predict the emotion class.

4 Training and Evaluation

The model was trained using the categorical cross-entropy loss and the Adam optimizer. Class weights were applied to address class imbalance. ReduceLROnPlateau and early stopping callbacks were used to optimize training. Monte Carlo Dropout was utilized during testing to perform multiple stochastic forward passes for robust predictions.

4.1 Results

The model achieved the following results on the test set:

- **Weighted Accuracy (WA): 86.78%**
- **Unweighted Accuracy (UA): 90.06%**

5 Confusion Matrix

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

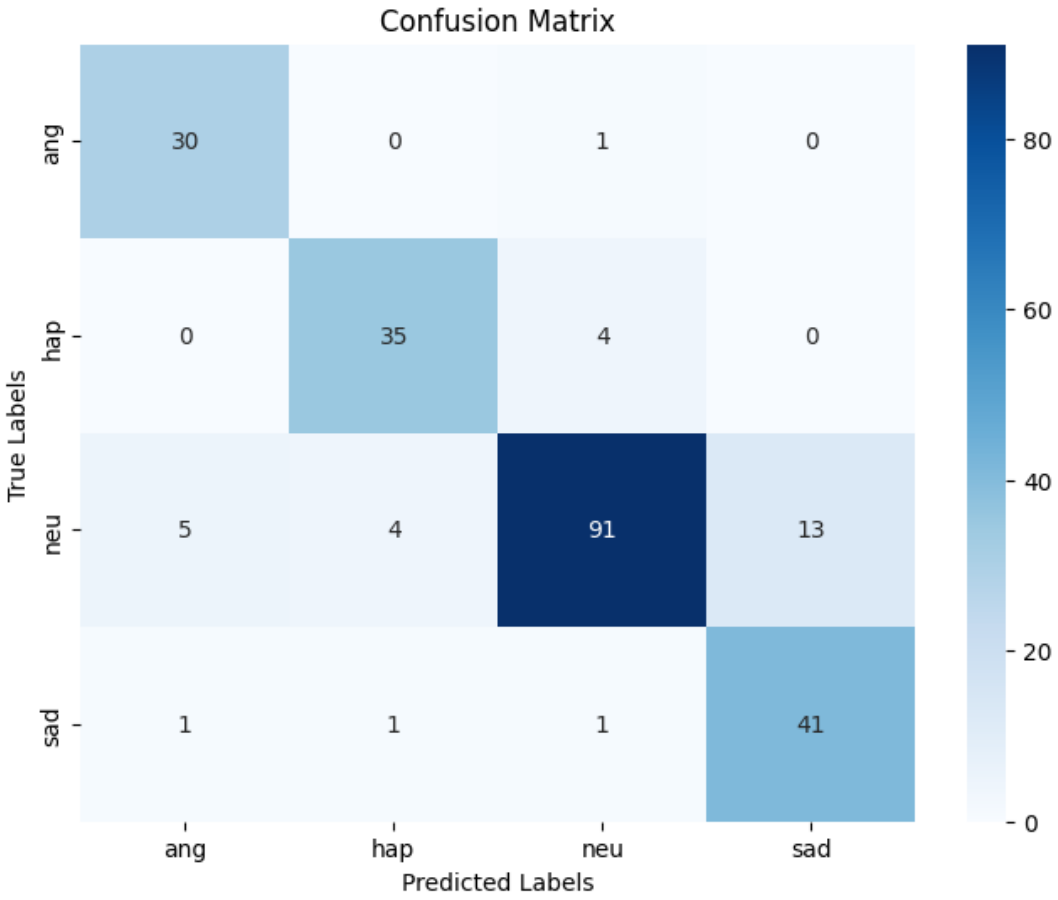

Figure 1: Confusion Matrix for Emotion Classification

6 Conclusion

This project demonstrates the effectiveness of integrating Monte Carlo Dropout into a multi-modal emotion recognition model. The enhancements improved the model's robustness and performance, achieving high weighted and unweighted accuracies on the IEMOCAP dataset.