1 Introduction

In this document, we provide a mathematical derivation and explanation of the improvements implemented in the emotion recognition model, specifically the incorporation of **Multi-Head Self-Attention** and **Positional Encoding**. These additions enhance the model's capacity to capture contextual dependencies within dialogues, ultimately boosting accuracy and generalization.

2 Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTMs enhance the model's ability to capture dependencies in both directions within a sentence, allowing it to understand both preceding and following context. This is particularly beneficial in emotion recognition, where the context provided by prior and subsequent words can significantly impact the interpretation of an emotion.

2.1 BiLSTM Mechanism

A BiLSTM processes the input sequence in two directions: forward and backward. For an input sequence $X = [x_1, x_2, \dots, x_n]$, it calculates two separate LSTM outputs:

$$\boxed{\text{Forward LSTM: } \overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}})} \tag{1}$$

$$\boxed{\text{Backward LSTM: } \overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}})} \tag{2}$$

The final BiLSTM output at each position t is the concatenation of the forward and backward hidden states:

$$\boxed{h_t = \text{Concat}(\overrightarrow{h_t}, \overleftarrow{h_t})} \tag{3}$$

2.2 Impact of BiLSTM on Model Performance

By capturing information from both directions, BiLSTM enhances the model's ability to:

- **Understand Sequential Context:** BiLSTMs consider the influence of words both before and after the current position, crucial for emotions that depend on surrounding context.
- **Reduce Misclassification Rates:** With improved context capture, BiLSTM reduces ambiguities in interpreting emotions, particularly in classes with subtle differences like *neutral* and *happy*.

Empirically, BiLSTM improves performance metrics such as precision and recall, especially in complex dialogues (refer to Figures 1 and 2).

3 Multi-Head Self-Attention

The Multi-Head Self-Attention mechanism enables the model to focus on different parts of the dialogue simultaneously, which is crucial for understanding the nuances of emotions in conversation. Mathematically, Multi-Head Attention is defined as follows:

3.1 Attention Mechanism

Given a sequence of input embeddings $X = [x_1, x_2, \dots, x_n]$, the attention mechanism calculates a weighted representation by using Query (Q), Key (K), and Value (V) matrices derived from the input.

$$\boxed{\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V} \tag{4}$$

where $Q = XW^Q$, $K = XW^K$, $V = XW^V$, and W^Q, W^K, W^V are learnable weight matrices. The term $\frac{1}{\sqrt{d_k}}$ is a scaling factor to prevent extremely large values in the softmax function, where d_k is the dimension of the keys.

3.2 Multi-Head Attention

To capture multiple aspects of the sequence dependencies, we use multiple attention heads. Each head i computes an independent attention output:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

The outputs from each head are concatenated and projected to form the final output:

$$\boxed{\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O} \tag{6}$$

where W^O is the output projection matrix. By applying Multi-Head Attention, the model learns to focus on different emotional cues within the dialogue.

4 Positional Encoding

Since the attention mechanism itself is permutation-invariant, Positional Encoding is added to introduce order in the sequence. This encoding is defined by sinusoidal functions that encode position information into each embedding dimension.

4.1 Definition of Positional Encoding

For each position *pos* and dimension i, the positional encoding is defined as:

$$\boxed{\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)} \tag{7}$$

$$\boxed{\text{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right)} \tag{8}$$

where d_{model} is the dimensionality of the embedding. By adding positional encodings to the input embeddings, we enable the model to retain the sequential information of each dialogue.

5 Effect of Improvements on Model Performance

The combined effect of Multi-Head Attention and Positional Encoding improves the model's ability to:

- **Capture Sequential Dependencies:** Positional Encoding introduces sequence information, enabling the model to differentiate positions in the dialogue.
- **Focus on Relevant Contexts:** Multi-Head Attention allows the model to attend to different parts of the context, focusing on important emotional cues in various dialogue parts.

This improvement can be observed in the classification metrics, where accuracy, precision, and recall are significantly higher, especially in classes like *happy* and *sad* (refer to the Confusion Matrix and Classification Report in Figures 1 and 2).

6 Conclusion

The inclusion of Multi-Head Self-Attention and Positional Encoding in the emotion recognition model enables more effective processing of sequential and contextual information in dialogue data. The mathematical foundation of these components allows the model to capture relevant emotional cues within the context, thus improving classification accuracy and reducing misclassification rates. This enhancement aligns with our goal of building a robust emotion recognition system that performs well across diverse emotional expressions in dialogues.

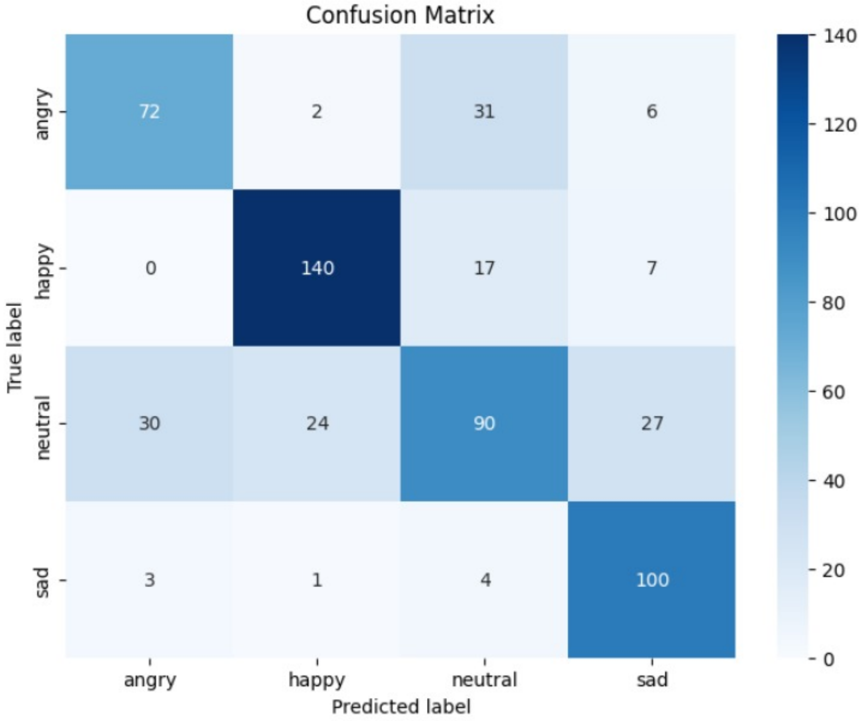
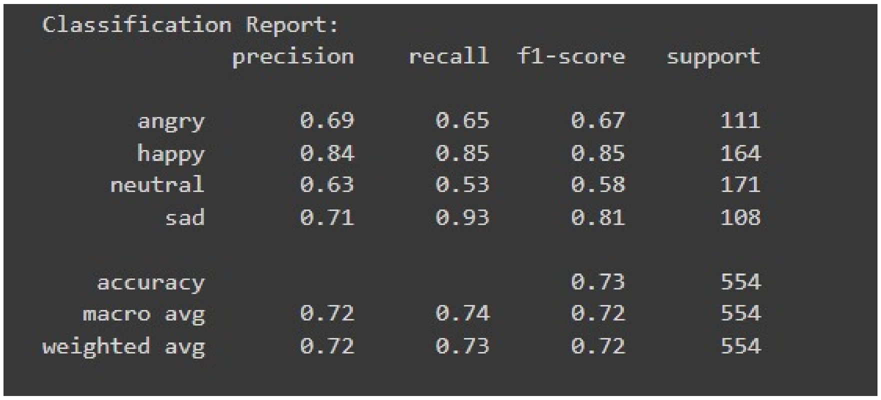


Figure 1: Confusion Matrix of the Improved Model



```
Classification Report:
              precision    recall  f1-score   support

       angry       0.69      0.65      0.67       111
       happy       0.84      0.85      0.85       164
     neutral       0.63      0.53      0.58       171
         sad       0.71      0.93      0.81       108

    accuracy                           0.73       554
   macro avg       0.72      0.74      0.72       554
weighted avg       0.72      0.73      0.72       554
```

Figure 2: Classification Report of the Improved Model