Enhancing Video Based Emotion Recognition with Multi-Head Attention and Modality Dropout

Xu Li

{li.xu2@northeastern.edu}

Khoury College of Computer Science, Northeastern University, Boston, MA, 02115, United States

Abstract. Multimodal emotion recognition has become a critical component in enhancing human-computer interaction systems due to its capacity to integrate multiple modalities. In this paper, a novel cross-modal fusion model CFNSR-MSAFNet was proposed with Multi-Head Attention mechanism and modality drop out to improve the accuracy of emotion recognition. The Multi-Head Attention mechanism allows the model to learn and observe multiple aspects from both audio and video input, capturing complex interactions between these two modalities. Additionally, modality dropout is introduced during training, forcing the model to learn representations to handle the missing or noisy data. The proposed model achieved 78.33% of accuracy on the RAVDESS dataset. Our results demonstrate the effectiveness of MHA and modality dropout in improving the performance of multimodal emotion recognition systems by enhancing cross-modal alignment and generalization.

Keywords: Multimodal Model, Emotion Recognition, Modality Dropout.

1 Introduction

A recent trend in AI technology is the growing application of services that draw upon emotions to improve human-computer interaction systems. Emotion recognition has become one of the most popular topics due to its ability to enabling machines to understand and interpret emotions from human [1]. Its ability to recognize emotions has made it applicable in various areas. The ability to help doctors monitor the mental health of patients makes it valuable in healthcare [2], it also can improve the efficiency of online teaching when used in the education [3], and it stood out for capacity to improve the user experience by better handling responses based on the user's emotional state in customer service [4]. As the demand for more stable and accurate systems grows, integrating multimodal data such as audio and video to improve model's performance is one of the most popular trends in current research field.

Previous study on emotion recognition mainly focused either analysing speech or facial expressions independently. Although these methods have advancements in certain area, their performance is relatively poor due to lacking the ability to capture the whole picture of human emotions that are often expressed through multiple modalities [5]. Recent research has focused on the multimodal approaches that combine both audio and visual data to better represent the complexity of emotional expression. However, several challenges remain in effectively integrating these modalities [6-8].

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354608

Zadeh et al. [6] proposed a Memory Fusion Network (MFN) that can independently process the temporal dynamics of each modality, but it cannot effectively interact across modalities and is significantly affected by noise or missing data in a particular data stream. Pham et al. [7] introduced the Deep Canonical Correlation Analysis (DCCA) model to use shared representation and enhanced fusion method to aligns audio and video modalities. However, the model failed to process the weakly correlated or missing modalities. Poria et al. [8] presented a Multi-level Multiple Attention Model, which applied attention at multiple levels to improves interpretability, but its single-modal modality structure restricts its ability to fully exploit cross-modal interactions due to.

To address these limitations, we propose CFNSR-MSAFNet, a cross-modal fusion model that enhancing the SFN-SR architecture proposed by Fu et al. [9] by integrating Multi-Head Attention (MHA) with residual connections, layer normalization, and modality dropout. MHA is known for its ability to handle both audio and video streams at the same time thus improve cross-modal feature integration and ensure high-accurate training [10]. The residual connections preserve crucial information flow, while layer normalization helps stabilize the gradients [11], ensuring more effective learning. Additionally, we introduce modality dropout during training, where either the audio or video stream will be randomly dropped, forcing the model to learn more stable representations. These innovations help the model generalize better when one modality is noisy or ambiguous, enhancing performance on real-world data compared to traditional LSTM-based models.

Our contributions in this paper are as follows:

1. We introduced a Multi-Head Attention with residual connections and layer normalization to more effectively aligns and integrates audio and video data for emotion recognition.
2. We propose the use of modality dropout to improve the stability of the model, enabling it to handle noisy or confused data more effectively.
3. We demonstrate the accuracy of the proposed model on the RAVDESS dataset, achieving a state-of-the-art accuracy of 78.33%, outperforming existing approaches.

The remainder of this paper is organized as follows: In Section 2, we introduce the dataset used for this study and describe the data pre-processing steps, followed by an overview of the proposed CFNSR-MSAFNet model, detailing its key components, including Multi-Head Attention and modality dropout. Section 3 presents the experimental results and provides a comparison with existing models. Section 4 discusses the implications of the results, potential limitations, and some possible areas for improvement. Finally, in Section 5, we conclude the paper and suggest future research directions.

2 Method

2.1 Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a well-known dataset for the recognition of emotions in speech. It consists of 1440 files with audio from 24 professional actors (12 men and 12 women) who express eight different emotions through speeches: calm, happy, sad, angry, fearful, surprised, disgusted, and neutral [12]. The

distribution of these emotions is shown by figure 2. One of RAVDESS's greatest strengths is its professionally produced audio, which reduces background noise and guarantees distinct emotional indications. This clarity is essential for training models in emotion perception especially for the multimodal model. Our work aims to improve the efficacy of emotion recognition systems using RAVDESS, hence advancing the field of affective computing.

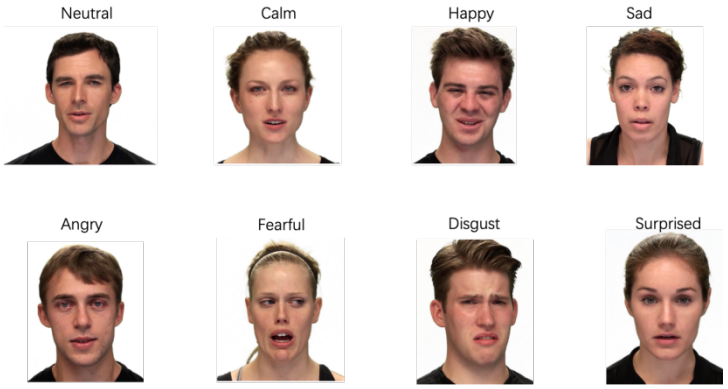

Fig. 1. Representative examples of the dataset [12].

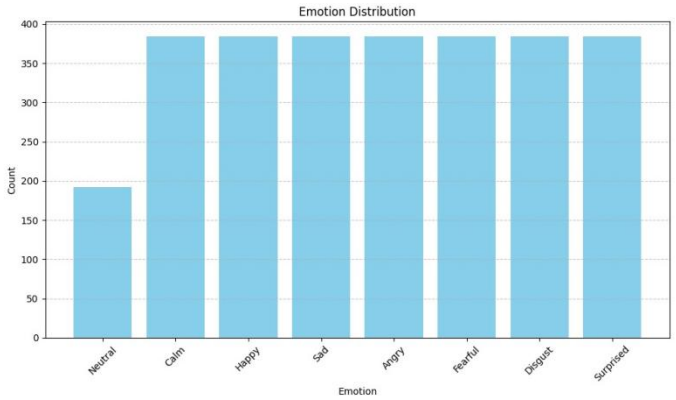

Fig. 2. The distribution of the dataset [12].

2.2 Data Pre-processing

We processed the audio and video data in the RAVDESS separately to ensure consistent input formatting and effective feature extraction. Mel-frequency cepstral coefficients (MFCCs) were used to capturing the perceptual properties from the audio that are crucial for emotion recognition [13], then standardization was applied to normalize features. For the video part, Frames were extracted at 30 frames per second and resized to 224x224 pixels, additional

augmentation through cropping and flipping ensure the high quality of the data. Next, by using the temporal alignment, corresponding segments of audio and video can be processed together. Finally, all features were normalized to zero mean and unit variance, and mini-batch processing was employed with a batch size of 16, supporting accurate training and generalization.

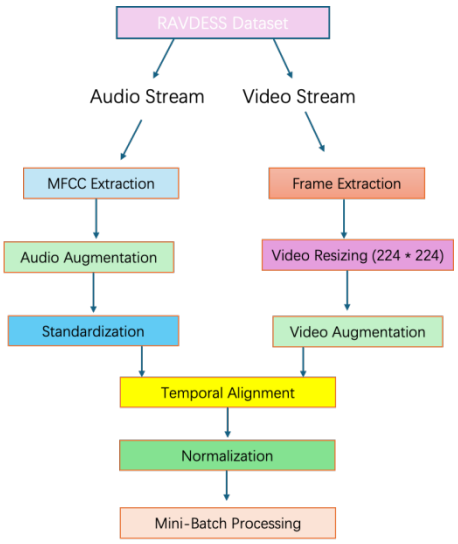

Fig. 3. Architecture of Data Pre-processing (Figure Credits: Original).

2.3 Model

The CFNSR-MSAFNet model is a modified based on the original CFNSR model [12]. The CFNSR -MSAFNet model processes multimodal data consisting of video frames and audio (MFCC) features. The video stream is passed through several convolutional layers and max-pooling layers, followed by ResNeXt blocks for feature extraction, while the audio stream is processed through a series of convolutional layers and pooling operations. Both streams are fused in the Cross-modal Blocks, where interactions between the audio and video modalities are captured and aligned, and the resulting fused representation is passed through a dense layer for final emotion prediction. Key modifications include adding residual connections to maintain information flow and prevent gradient degradation, Multi-Head Attention (MHA) to focus on important features from both streams, and layer normalization to stabilize training. What's more, including the modality dropout during the training process can enhance the stability and generalization of model by let it handle noisy or missing data.

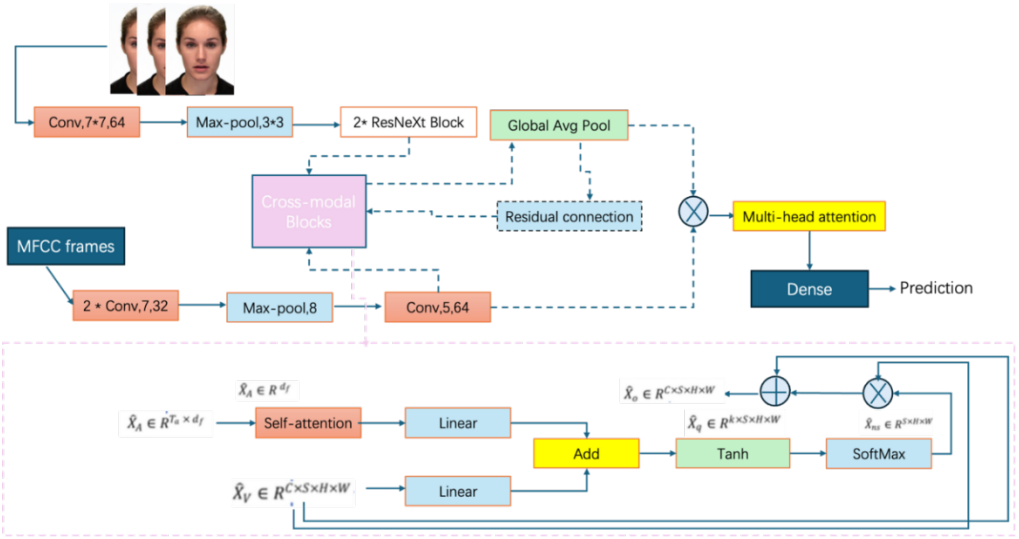

Fig. 4. Architecture diagram of CNN (Figure Credits: Original).

2.4 Multi-head attention

By allowing the model to simultaneously focus on multiple aspects of the input data, Multi-Head Attention can enhance enhancing model's ability to process complex relationships for both within and between modalities. This structure was first introduced in the Transformer architecture by Vaswani et al. [10], its effectiveness at handling long-range dependencies and the ability to integrate complementary information from multiple modalities making it an ideal choice for improving recognition accuracy. In the tasks of multimodal emotion recognition, MHA plays a crucial role in matching and fusing features from both modalities. In addition, it can also facilitate a deeper interaction between the streams by focusing on key elements such as speech patterns and facial expressions. Each head processes different subspaces of the data, ensuring the model captures a wide range of temporal and spatial dependencies.

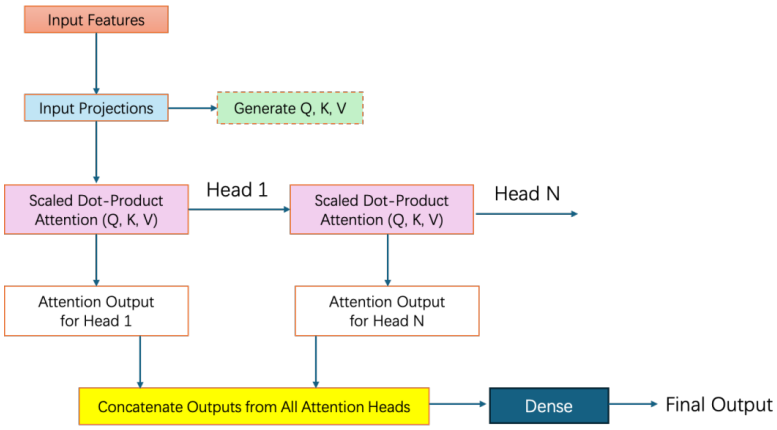

Fig. 5. Architecture diagram of MHA (Figure Credits: Original).

Scaled Dot-Product Attention. The Scaled Dot-Product Attention is the fundamental part of the Multi-Head Attention. It computes attention scores by applying the SoftMax function to the dot product between the query (Q) and key (K) matrices scaling by $\frac{1}{\sqrt{d_k}}$ to obtain a probability distribution. This distribution represents how much attention each key should receive and the values (V) are weighted accordingly. The formula is given by equation (1)

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d_k}})\mathrm{V} \quad (1)$$

SoftMax function ensures the attention scores sum up to 1 and emphasizes the relevant parts of the input based on the similarity between queries and keys. The values will be scaled by the attention scores before summing up as an output, allowing the model to focus on important features in the input.

Multi-Head Attention Structure. Multi-Head Attention extends the single attention mechanism by using multiple attention heads. Each head operates on its own set of projections for Q, K, and V, and focuses on different parts of the input sequence, learning diverse representations across different subspaces. The key benefit of having multiple heads is that it allows the model to attend to various positions and interactions in parallel, providing a more comprehensive understanding of the input [10]. For each attention head, the following steps are performed:
Linear projections of the input features are computed to generate the queries Q, keys K, and values V, where W_Q, W_K, and W_V are learnable weight matrices. The formula is given by equation (2)

$$\mathrm{Q} = \mathrm{X}W_Q\ ,\quad \mathrm{K} = \mathrm{X}W_K\ ,\quad \mathrm{V} = \mathrm{X}W_V \quad (2)$$

Scaled dot-product attention is applied to each head independently, as described earlier. The formula is given by equation (3)

$$Attention_i(Q_i, K_i, V_i) = SoftMax(\frac{Q_i {K_i}^T}{\sqrt{d_k}})V_i \quad (3)$$

The outputs of the attention heads are concatenated and projected through a linear transformation, where W_O is a learnable output projection matrix, and h is the number of attention heads. The formula is given by equation (4)

$$\mathrm{MultiHead\ (Q, K, V)} = \mathrm{Concat}\ (head_1, \ldots., head_h)\ W_O \quad (4)$$

Finally, after the operations above, each attention head will output the weighted sum of the value vectors (V),

2.5 Residual Connections and Layer Normalization

In order to ensure the stability during the training process, residual connections are introduced after the MHA module to make model capable for retaining information from previous layers. The output of the attention mechanism is added in to the input X, forming a residual connection. The formula is given by equation (5)

$$\mathrm{Output} = LayerNorm(\mathrm{X} + \mathrm{MultiHead(Q\ , K\ , V)}) \quad (5)$$

By allowing gradients to flow more easily, this residual connection can helps prevent vanishing gradients and facilitates the training of deep models [14]. What's more, layer normalization is applied to ensure that the distribution of activations remains stable throughout the network after the residual connection.

2.6 Modality Dropout

Modality Dropout is a regularization technique designed to enhance the generalization capability of multimodal models. In our model, modality dropout is applied with a predefined probability during training. For each batch, either the audio or video modality is randomly set to zero, while the remaining modality is processed normally. For each training step, modality dropout can be expressed as the formula given by equation (6).

$$x'_{audio}\begin{cases} x_{audio} & with\ probability\ 1-p_{drop} \\ 0 & with\ probability\ p_{drop} \end{cases}$$
$$x'_{video}\begin{cases} x_{video} & with\ probability\ 1-p_{drop} \\ 0 & with\ probability\ p_{drop} \end{cases} \quad (6)$$

The model will then process the remain modality as the formula given by equation (7)

$$x_{fused} = f(x'_{audio}, x'_{video}) \quad (7)$$

This removal of modalities encourages the model to learn useful, independent feature representations from both streams, which leads to better overall fusion during intersection when both modalities are present.

3 Results

3.1 Experiment Setup

In this study, we used the RAVDESS dataset to perform multimodal emotion recognition, leveraging both audio and video data. The proposed CFNSR - MSAFNet model integrates audio and video streams using Multi-Head Attention (MHA) and modality dropout to enhance feature fusion and robustness. The audio stream was processed using a Transformer-based encoder, while the video stream was passed through a CNN. The model was trained on the L4 GPU provided by Colab, with a batch size of 16, using the Adam optimizer and a learning rate of 0.001. Categorical cross-entropy was employed as the loss function, with both standard and modality dropout applied to prevent overfitting. The model was trained for 45 epochs, and accuracy was used as the primary evaluation metric.

3.2 Performance Analysis

The Training and Test Accuracy/Loss Comparison Over 45 Epochs shown by Figure 6 demonstrates that the model effectively learned from the training data, achieving a near-perfect accuracy of 98-99% by epoch 7. However, there is a clear issue with overfitting, as seen from the significant gap between training and test accuracy. The test accuracy fluctuated throughout training, ranging between 65% and 78%, with the highest test accuracy of 78.33% occurring at epoch 18. The fluctuation of test accuracy and the increasing test loss in later epochs indicate that the model struggles to generalize well to unseen data, even as it fits the training data almost perfectly.

This overfitting problem be revealed in the loss curves. While the training loss continually decreased, almost reaching zero, the test loss showed a volatile pattern. By the final epoch, the test loss had increased to 2.5261, indicating that the model was memorizing the training data rather than learning features that generalize to the test set. These results suggest that despite the strong fit to the training data, the model's generalization capability could benefit from other techniques to reduce overfitting, such as early stopping, stronger regularization, or data augmentation.

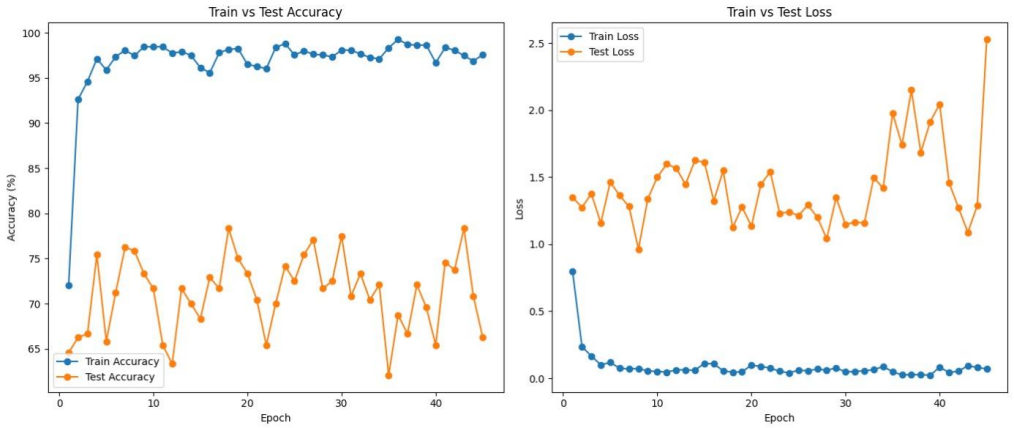

Fig. 6. Training and Test Accuracy/Loss Comparison Over 45 Epochs (Figure Credits: Original).

3.3 Confusion Matrix

The confusion matrix shown by Figure 7 offers a detailed insight for model's performance across various emotion categories, with predictions shown on the x-axis and true labels on the y-axis. The correspondence between emotions and index is shown in the Table 1. This model shows its best performance in predicting "Happy" (index 2) and "Surprised" (index 7), as indicated by the darkest blue diagonal cells at (2,2) and (7,7). These darker cells indicate a high number of correct predictions, suggesting that the model effectively captures the distinguishing features of these emotions from both audio and video modalities.

However, this model also struggles with certain categories, particularly with distinguishing "Neutral" (index 0) and "Sad" (index 3). The confusion between these two emotions is highlighted by the lighter shading along cells (0,3) and (3,0), indicating frequent misclassification. This difficulty might caused by shared subtle expressions in facial and vocal cues between "Neutral" and "Sad" [15], making it challenging for the model to differentiate them accurately.

Table 1. Emotions and their corresponding index

Emotions	Index
Neutral	0
Calm	1
Happy	2

Sad	3
Angry	4
Fearful	5
Disgust	6
Surprised	7

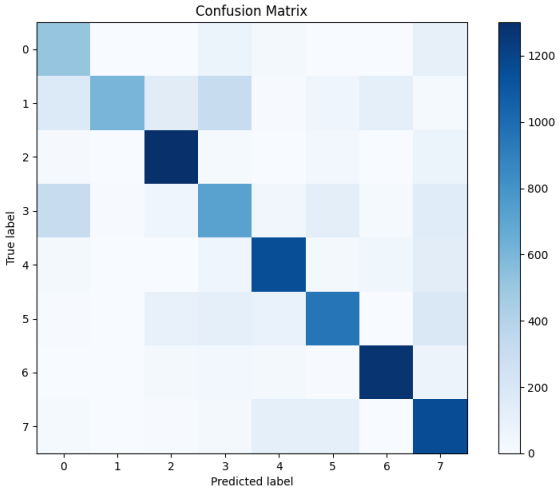

Fig. 7. Confusion Matrix (Figure Credits: Original).

3.4 Performance Comparison

The performance comparison table shown by Table 2 highlights the evaluation results for various methods used in emotion recognition. The CFNSR-MSAFNet model we proposed achieved the highest accuracy of 78.3%, demonstrating significant improvements over other models in the comparison. The CFN-SR model, which also leverages cross-modal fusion, obtained an accuracy of 75.76%, marking it as a strong performer but slightly below than the enhanced CFNSR-MSAFNet.

Among the baseline methods, Multiplicative fusion approaches and MCBP (Modality Confidence-Based Pooling) methods demonstrated comparable results with accuracies around 70-71%. Models such as MSAF and ERANNS, which focus on integrating spatial and temporal features, performed in the 74-75% range. The Averaging model achieved the lowest accuracy at 68.82%, suggesting that simple fusion techniques are less effective compared to more advanced strategies involving attention mechanisms and residual connections.

The results clearly shows that models utilizing advanced fusion techniques, like CFNSR-MSAFNet, has the best performance, highlighting the effectiveness of leveraging both spatial and temporal information across modalities. These advancements address feature redundancy and enhance emotion recognition accuracy.

Table 2. Performance Comparison

Model	Accuracy
Averaging	68.82
Multiplicative	70.35
Multiplication	70.56
Concat + FC	71.04
MCBP	71.32
MMTM	73.12
MSAF	74.86
ERANNS	74.8
CFN-SR	75.76
CFNSR- MSAFNet (Ours)	78.3

4 Discussion

4.1 Analysis of Model Strengths and Weaknesses

The CFNSR-MSAFNet model proposed in this study achieves better result on the emotion recognition tasks and has following advantages. The use of Multi-Head Attention (MHA) to efficiently capture and align features across both audio and video modalities is one of the notable strengths in this model, which been justified by the 78.3% of accuracy in the real-world data. MHA can observe and learn from different aspects for input data, thus helps the model identifies crucial features for emotion recognition such as tones and facial expressions from both modalities. This significantly enhances the cross-modal fusion process, leading to more effective representation of emotional states.

Another strength of the model is the usage of Modality Dropout, which improve the stability by help the model handling the less informative or unreliable modality. This technique will force the model learn from both modalities, thus vanishing over-relying on a certain modality and improving the generalization performance. Additionally, the use of residual connections and layer normalization plays a vital role in stabilizing the training process and preventing vanishing gradients, particularly in deep architectures, which allows for efficient gradient flow through the network layers.

The CFNSR-MSAFNet model designed in this paper also has some shortcomings that need to be further improved. As the Figure 6 shows, overfitting occur during the later epochs. The increasing test loss while the training loss remains relatively low, as well as fluctuating pattern shows in both test accuracy and test loss also supported this opinion. Overfitting can result from the model being too complex, learning noise or minor details from the training set that do not generalize well to the validation or test sets.

Another weakness is model struggles with distinguishing between certain emotions, particularly between neutral and sad emotions. This indicates that while the model performs well in general,

there are specific emotional classes where feature overlap or ambiguity persists, making it harder for the model to make accurate predictions.

4.2 Suggestions For the Future Work

Incorporating Additional Modalities. Currently, only audio and video data are the currently considered in this model. Expanding the model to include additional modalities such as physiological signals (e.g., heart rate or skin conductance) might improve emotion classification accuracy since it will provide richer contextual data to the model. This would address the model's potential insensitivity to subtle details that are not easily captured by visual and auditory cues, thus decrease the frequency of occurrence for the emotion misclassification.

Improvement of Modality Interaction. While Multi-Head Attention (MHA) improves the interaction between the audio and video modalities. Relationship between different modalities could be further improved by incorporating co-Attention networks, graph-based interaction models or other advanced cross-modal fusion mechanism [16]. These fusion mechanisms could help model address cases where the model struggles with subtle emotional cues that require more nuanced multimodal processing from different ways. For Instance, Co-attention mechanisms can allow modalities to attend to each other during feature extraction to fuse relevant information and Graph-based models' framework can help models capture complex and indirect relationships between modalities.

Addressing Overfitting and Generalization. Despite the introduction of Modality Dropout during training, overfitting is still an essential problem in current model. Future work may involve integrating more advanced regularization techniques such as variational dropout or label smoothing [17]. These techniques can help models generalize better on different emotion data sets by reduce models' confidence towards their predictions or treating dropout as a learnable parameter rather than a fixed one, thus mitigating overfitting especially when encountering new or unseen emotional expressions.

5 Conclusion

In conclusion, the CFNSR-MSAFNet model proposed in this study showcase a great performance in multimodal emotion recognition by leveraging Multi-Head Attention (MHA) and Modality Dropout. These techniques eventually lead to a more accurate and steady emotion recognition system by enabling the model to effectively fuse features from both audio and video modalities, which is justified by the 78.33% of accuracy in real world dataset. The residual connections further enhance the model's performance by facilitating gradient flow during training and contributing to the stability and efficiency of deep model optimization.

However, there are still areas for further improvement. The fluctuating pattern throughout the training test and loss indicates that overfitting occurred during the process of training. Misclassifications between neutral and sad state that model often confused nuanced emotional expressions. These limitations suggest CFNSR-MSAFNet can be improved by focusing on enhancing the model's generalization capabilities through advanced regularization techniques like label smoothing. Future research could base on this work by exploring more sophisticated fusion mechanisms and different datasets to capture different kinds of emotions. The promising

results of this study lay a solid foundation for further development in the domain of multimodal emotion recognition.

References

[1] Salah, M., Alaa, A., Adel, A., Amr, A., Saad, S., Gamal, A., & Aly, S. (2023, November). Emotion Recognition: Enhancing Human-Computer Interaction. In 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 542-548). IEEE.

[2] Ayata, D., Yaslan, Y., & Kamasak, M. E. (2020). Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering*, *40*, 149-157.

[3] Barron-Estrada, M. L., Zatarain-Cabada, R., & Bustillos, R. O. (2019). Emotion Recognition for Education using Sentiment Analysis. *Res. Comput. Sci.*, *148*(5), 71-80.

[4] Gupta, S., Saxena, S., Verma, V., & Nishad, D. K. (2024, May). Facial Emotion Recognition for Virtual Customer Service Agents. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)* (pp. 321-326). IEEE.

[5] Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. P. (2018, April). Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[6] Ahmed, N., Al Aghbari, Z., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, *17*, 200171.

[7] Pham, H., Liang, P. P., Manzini, T., Morency, L. P., & Póczos, B. (2019, July). Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*(Vol. 33, No. 01, pp. 6892-6899).

[8] Poria, S., Cambria, E., Hazarika, D., Mazumder, N., Zadeh, A., & Morency, L. P. (2017, November). Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*(pp. 1033-1038). IEEE.

[9] Fu, Z., Liu, F., Wang, H., Qi, J., Fu, X., Zhou, A., & Li, Z. (2021). A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. arXiv preprint arXiv:2111.02172.

[10] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[11] Ba, J. L. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

[12] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

[13] Debnath, S., & Roy, P. (2021). Audio-visual automatic speech recognition using PZM, MFCC and statistical analysis.

[14] Borawar, L., & Kaur, R. (2023, March). ResNet: Solving vanishing gradient in deep networks. In *Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022* (pp. 235-247). Singapore: Springer Nature Singapore.

[15] Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS one*, *12*(5), e0177239.

[16] Zhao, Z. W., Liu, W., & Lu, B. L. (2021, May). Multimodal emotion recognition using a modified dense co-attention symmetric network. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 73-76). IEEE.
[17] Zhang, C. B., Jiang, P. T., Hou, Q., Wei, Y., Han, Q., Li, Z., & Cheng, M. M. (2021). Delving deep into label smoothing. *IEEE Transactions on Image Processing*, *30*, 5984-5996.