Enhancing Emotion Recognition Accuracy with CFRSN-LSTM

Yidan Zhang[1,*,†] , Yangyue Zheng[2,†]

{yidanzhang.claire@outlook.com[1], yangyue.zheng@outlook.com[2]}

Taiyuan University of Technology, College of Software, Taiyuan, China[1]
University of Sydney, Faculty of Engineering, Sydney, Australia[2]
*corresponding author
[†]co-first authors

Abstract. Multimodal emotion recognition is a challenging problem partly because the loss of original semantic information during intra- and inter-modal interactions. In this paper we propose to a novel cross-modal fusion network based on self-attention and residual structure (CFNSR-LSTM) for multimodal emotion recognition. We first perform representation learning for audio and video modalities to obtain the spatio-temporal structural features of video frame sequences and the MFCC features of audio sequences by efficient ResNeXt and a simple and effective one-dimensional CNN method, respectively. We then feed the features of the audio and video modalities into the cross-modal blocks separately. The audio features are processed using LSTM to capture temporal characteristics, and the self-attention mechanism is employed for intra-modal feature selection, which enables efficient and adaptive interaction between the selected audio features and the video modality. The residual structure ensures the integrity of the original structural features of the video modality. Finally, we conduct experiments on the RAVDESS dataset to verify the effectiveness of the proposed method. The experimental results show that the proposed CFNSR-LSTM model, with a parameter count of 26.40M, achieves the best performance with an accuracy of 76.25% and outperforms other models.

Keywords: multimodal emotion recognition, convolutional feature residual network, LSTM (Long Short-Term Memory), self-attention.

1 Introduction

Emotions are the basic bond of daily communication between people, and emotion recognition can be applied in various fields, such as the smart car driving safety in life, which recognizes the correspondent negative emotions of the driver to make timely adjustments to help avoid accidents [1], and it is also possible to use models based on facial emotion recognition for predictions used for specific conditions (Parkinson's disease) [2], [3].

In late decades of research emotion recognition models have gradually transformed from unimodal to multimodal, and the application scenarios have become more, but how to efficiently fuse between different modalities is still the focus of research.

Traditional fusion methods are mostly linear paradigms, such as feature splicing and multi-system fusion, which are difficult to capture the complex associations between audio and video.

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354597

Moreover, existing studies tend to ignore the complementary learning of static and dynamic features in automatic speech emotion recognition, limiting the performance [4].

In contrast, our research develops a fully end-to-end model that connects the two stages and performs joint optimization, avoiding the drawbacks of traditional two-stage pipelines. In addition, full end-to-end training is achieved by reconstructing the dataset. To reduce the computational overhead, a sparse cross-modal attention mechanism is introduced for feature extraction, which maintains the performance while reducing the computational effort by about half.

Our research is to develop the CFNSR-LSTM cross-modal fusion network, which integrates audio and video data to improve the accuracy of emotion recognition with less data. Specifically, the model extracts Mel Frequency Cepstrum Coefficients (MFCC) [5] from audio using a 1D CNN and spatiotemporal features from video using ResNeXt[6]. Then, an LSTM processes these features and augments them with a self-attentive [7] mechanism for efficient and adaptive interactions between modalities.

In this research, our main objectives are as follows:

– By proposing CFNSR-LSTM cross-modal fusion network to comprehensively capture emotional signals and overcome the limitations of unimodal data.

– Develop a full end-to-end model that is more efficient in order to reduce information loss and the fusion of features.

– Introducing a cross-modal attention mechanism reduces the model overhead and improves the efficiency of the model, making it suitable for application in resource-constrained environments.

– Combining LSTM and self-attention mechanisms to analyse the feature changes over time series of audio and video modalities to further improve the accuracy and adaptability of emotion recognition.

2 Related Work

In the past few decades, the field of emotion recognition has changed significantly, in the early days the main focus was on unimodal emotion recognition methods, and the development of deep learning was not as well established. Fatemeh Noroozi et al. in 2017 proposed to use random forests combined with features of speech signals to recognize the emotional state , and the results achieved an average 66.28% correct recognition rate, which is decisive for signal processing to recognize voice emotions[8]. With the development of deep learning, Jianfeng Zhao et al. in 2018 designed a Network which has two branches (1D CNN branch and 2D CNN branch) for learning high-level features and achieve 89% correct rate. However, it was found that CNN has a memory function that affects the performance of the target network [9].

With the increase of demand, the unimodal emotion recognition system could not satisfy the needs of development, so Egils Avots et al. started to build a multimodal model for recognizing human emotions based on audio-visual information with different corpora samples for training, but it was found that resulting in poorer fusion results than individual results, the accuracy for

the testset was only 27.1% [10]. While the data for training and testing are more closely aligned to real-world situations, model fusion is not well worked out.

In order to achieve better results in the extraction of visual features based on fusion model, Liam Schoneveld et al. have proposed the use of the latest advances in deep learning to develop a method by using self -distillation for CNN model facial feature extractor training results with fine-tuning VGGish backbone for audio feature model training results in model-level fusion, the results on Google FEC dataset accuracy has a significant improvement [11], but the audio environment still affects the recognition, which needs to be combined with better algorithms in audio feature recognition.

To solve the existential problem of better fusion between facial expression recognition models and audio recognition models we propose the development of a cross-modal fusion network based on self attention and residual structure (CFNSR-LSTM) for multimodal emotion recognition by selecting features from their respective modalities to obtain emotionally relevant regions through attention-guided factorial bilinear pooling in deep fusion [12], and completing the whole process in neural networks. Meanwhile, the proposed hierarchical network based on the fusion of static and dynamic features can effectively integrate static and dynamic features in speech emotion recognition [13], and shows superiority on the balanced dataset through well-designed modules and attention mechanisms.

3 Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a well-known dataset for therecognition of emotions in speech [14]. It consists of 7356 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. The distribution of these emotions is shown by figure 1. 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 through the use of RAVDESS, hence advancing the field of affective computing.

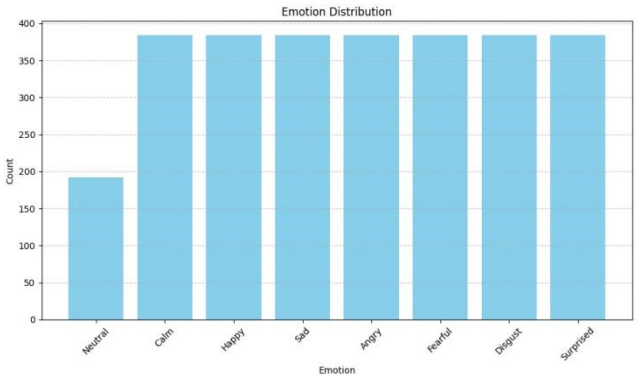

Fig. 1. Emotion Distribution

4 Methdology

4.1 Research Design

This study aims to improve the accuracy of emotion recognition through audiovisual deep learning, for which a CFNSR-LSTM cross-modal fusion network is proposed. In this paper, this approach is chosen mainly because emotion recognition has a key role in several important fields; in healthcare, it is crucial to accurately identify diseases related to emotional characteristics; in human-computer interaction, good emotion recognition can enhance the interaction experience. By fusing audio and video information, this paper expects to capture emotion signals more comprehensively and improve recognition accuracy.

The objects of this study are speech and facial expressions in the RAVDESS dataset. In this paper, the audio and video modalities are processed separately, with the audio extracting the MFCC features by a one-dimensional CNN method, and the video obtaining the spatio-temporal structure using ResNeXt. Then these features are fed into the cross-modal module, and LSTM is used to obtain the audio features on the time series, and the audio is allowed to select the internal modal features through the self-attention mechanism to realize the efficient adaptive interaction with the video modality.

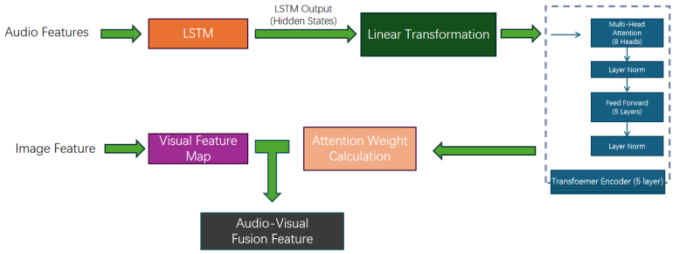

Fig. 2. LSTM-Attention

LSTM-Attention effectively captures long-term dependencies in sequential data.

4.2 Data Collection

For data collection, the RAVDESS dataset was used as the source for this study. This dataset contains rich speech and facial expression data, which provides a reliable basis for multimodal emotion recognition. The specific data elements collected include audio modality and video modality. For audio, a 1D CNN method is used to obtain its MFCC features, which can effectively reflect the spectral characteristics of audio. For video, ResNeXt is utilized to extract the spatiotemporal structure of the video frame sequences, so as to capture the changes of facial expressions over time. This dataset has the distinct advantage of containing 1,440 short voice-over video clips performed by 24 actors (12 male and 12 female). These actors performed after being explicitly informed of the emotions to be expressed, ensuring the relevance and authenticity of the data. The dataset is of high quality in terms of both video and audio recordings, providing reliable base data for the study. Among the eight emotion categories covered in the dataset were neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. This rich categorization of emotions enables us to conduct multimodal emotion recognition studies to comprehensively explore the performance characteristics of different emotions in audio and video modalities.

4.3 Data Analysis

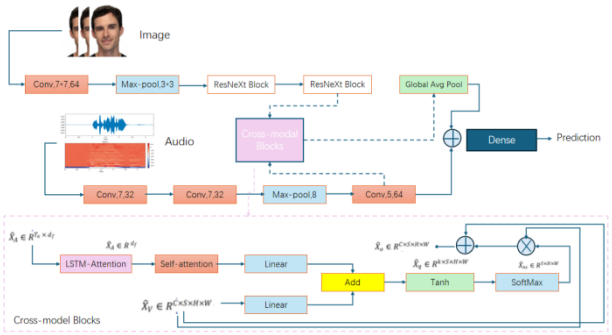

Fig. 3. Model design

In this paper, the one-dimensional CNN method is chosen to extract the MFCC features of audio because MFCC has a wide range of applications and good results in the field of audio processing, which can effectively capture the key information of audio signals. The use of ResNeXt to process video data is due to its strong ability in extracting image features, which can well capture the spatio-temporal information in video. For the use of LSTM and self-attention mechanism, the importance of time series and modal interaction in emotion recognition is taken into account, which can better integrate the information of different modalities and improve the effect of emotion classification.

4.4 Model Evaluation

In terms of model evaluation, this paper focuses on the recognition effect and accuracy of the CFNSR-LSTM model by examining the recognition effect and accuracy of the CFNSR-LSTM

model. Accuracy is one of the key indicators of model performance, which is calculated by comparing the model's prediction results on emotions with the actual emotion labels. In addition, recognition effectiveness takes into account the model's performance on different sentiment categories, its ability to distinguish complex sentiment states, and its stability under different data sizes.

In order to ensure the objectivity and reliability of the evaluation, this paper divides the dataset into training, validation, and testing sets, and comprehensively evaluates the performance of the model by its performance on different datasets. It is also compared with other existing emotion recognition models to highlight the advantages and innovations of the CFNSR-LSTM model.

The experimental results show that the CFNSR-LSTM model exhibits better recognition results with 76.25% accuracy with less data. This result shows the powerful ability of the model in dealing with limited data and has high practical value. The cross-modal fusion design in this paper can make full use of audio and video information to make up for the lack of single-modal data. On the other hand, the use of the selfattention mechanism and LSTM enables the model to better capture the emotion changes and inter-modal interactions in the time series, thus improving the accuracy of emotion recognition.

4.5 Ethical Considerations

In this study, the RAVDESS dataset was used for the training and evaluation of emotion recognition models. The primary ethical consideration for the use of data is to ensure the legality and legitimacy of the source of the data. The RAVDESS dataset should have been collected and published through legal means, thus ensuring that the data underlying the study is ethical.

5 Result

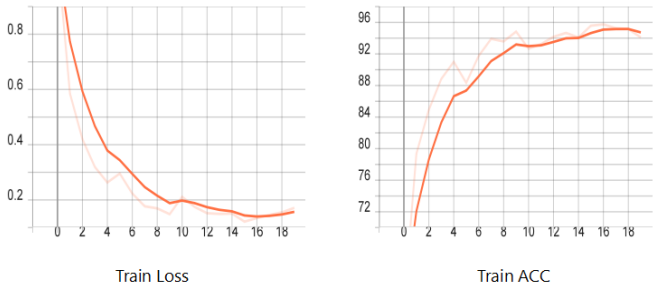

Fig. 4. Train Performance

After training through the building of the model, we analyse the results by visualising the loss value and the accuracy rate of the training set as shown in the line graph in Fig. 4. We can observe that the left graph shows that the loss decreases with the increase of epoch, which shows that the model is constantly learning, and combined with the right graph we can see that the accuracy improves with the epoch and becomes stable, which shows that the classification ability of the model in the training set is constantly improving.

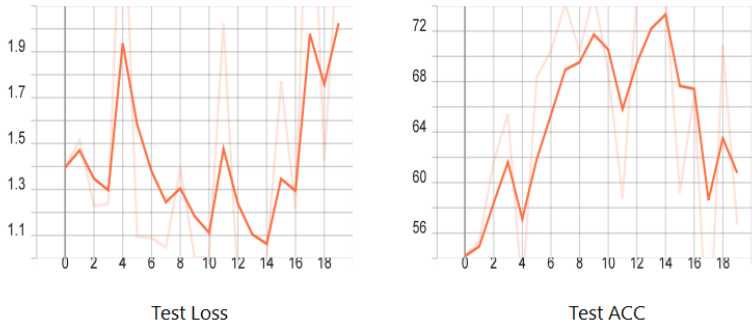

Fig. 5. Test Performance

At the same time, we also visualise the loss and accuracy of the test set, we can observe on the left side of Fig. 5 that the fluctuation of the loss on the test set is relatively large and the end of the test has an upward trend, for the loss on the training set probably overfitting has occurred, which performs well on the training set but is unstable on untrained data, and on the right side of Fig. 5, it shows that the accuracy of the test set have the fluctuating trend, which proves that the model has a weak generalisation ability on the test set. In the future we considered using regularisation methods or adding more training data to improve the generalisation ability of the model.

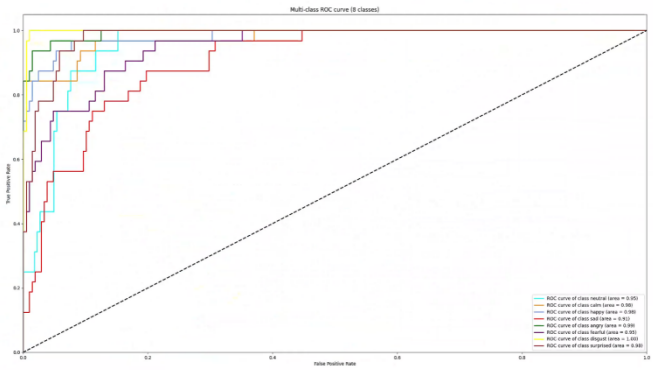

Fig. 6. ROC curve

We represent the performance of the model on each emotion classification by the multiclassification ROC curves in Fig. 6, where each curve represents a emotion class, and we can see that most of the ROC curves have good performance with stick close to the top left corner, in comparison ”disgust” performs better with an AUC closer to 1 (yellow line), and ”sad” performs worse with a lower value of the AUC (red line) indicating that model performs worse in ”sad” class.

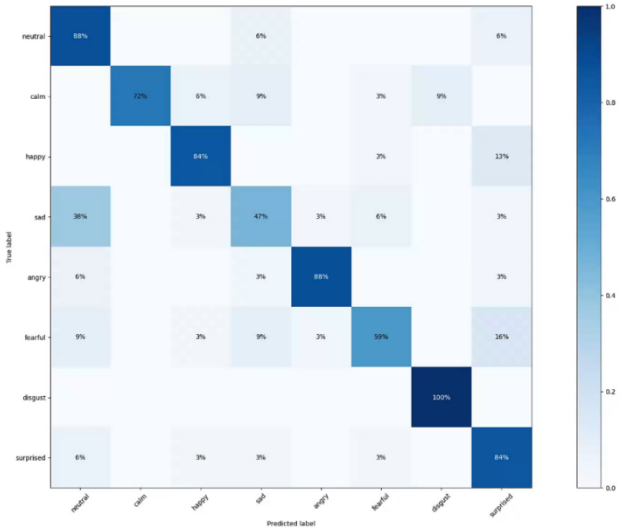

Fig. 7. Mixed matrix

We can combine Fig. 6 and Fig. 7 to show the results of different emotion classification more obviously, we can see that the emotion of "sad" is easily confused with the emotion of "neutral", we also studied the training video and found that it is more similar among these two emotions, which makes the model perform poorly in recognising the emotion of "sad", only have the 47% accuracy.

6 Discussion

Table 1. Comparisons

Model	Fusion Stage	Accuracy	Params
3D RexNeXt50 (Vid.)	-	62.99	25.88M
1D CNN (Aud.)	-	56.53	**0.03M**
Averaging	Late	68.82	25.92M
Multiplicative β=0.3	Late	70.35	25.92M
Multiplication	Late	70.56	25.92M
Concat + FC	Early	71.04	26.87M
MCBP	Early	71.32	51.03M
MMTM	Model	73.12	31.97M
MSAF	Model	74.86	25.94M
ERANNs	Model	74.80	-
CFN-SR	Model	75.76	26.30M
CFNSR-LSTM(ours)	Model	**76.25**	26.50M

In this section we compare our proposed CFNSR-LSTM model with various models that already exist for emotion recognition tasks using video and audio data:

1) Accuracy: In terms of accuracy our CFNSR-LSTM model stands out with the highest accuracy of 76.25%, the model fusion we used is better than the other models with early fusion and late fusion, we can see through Table 1 that model fusion is higher than the other fusion methods, even though the same complex model level fusion is used, such as ERANNs (74.8%) [15] and CFNSR (75.76%) [16], but still lower than our model.

2) Number of parameters: In terms of the number of parameters (26.50M) we also keep a low level, through the Table 1 we can see that the most efficient 1D CNN uses only 0.03M of data but the accuracy rate is only 56.53%, which shows that there is still a trade-off between the model and the performance. We achieve a better accuracy rate with a certain amount of data.

The comparison shows that our CFNSR-LSTM model finds the best balance between model complexity and performance.

7 Future Work

In future work, several areas can be further explored to improve the performance of our:

– Overfitting problem: In the future we are going to apply to complex emotional states, overfitting is still a concern question, we will try more regularisation techniques to choose one more fit our model to enhance generalization in the future.

– Low accuracy of specific emotions:According to the resuls, we found ’sad’ and ’calm’ are more difficult to classify accurately beacause of similarity. In the future, we can train on a larger dataset to complement the emotion categories that did not perform so well to improve the accuracy of each emotion classification.

By solve above problem to make our next generation of emotion recognition models achieve even greater accuracy and stability, making them better to implement in real-world scenarios.

8 Conclusion

In this research, we propose a new CFNSR-LSTM model that combines multiple advanced features in order to solve several key challenges in audio-visual emotion recognition, in particular the imbalance of processing modalities and the missing of information during the suboptimal fusion of audio and video features. We use the LSTM-attention allows for targeted analysis of the input data without the need for extensive global context, increasing the focus on important features. In addition to use items level model fusion technique to combined the strengths of multiple models to better handle changes in the input data so that the model achieves a higher level of accuracy. The test results demonstrate the superiority of the CFNSR-LSTM model, achieving an accuracy of 76.25% with small dataset, which outperforms several baseline model. At the same time, we still need to integrate more datasets to improve the accuracy of emotion recognition of all classes, and try more regularisation techniques to improve the generalisation ability of the model.

Acknowledgment

Yangyue Zheng and Yidan Zhang contributed equally to this work and should be considered co-first authors.

References

[1] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," ACM Computing Surveys (CSUR), vol. 53, no. 3, pp. 1–30, 2020.

[2] S. Argaud, M. V rin, P. Sauleau, and D. Grandjean, "Facial emotion recognition in parkinson's disease: a review and new hypotheses," Movement disorders, vol. 33, no. 4, pp. 554–567, 2018.

[3] [3] M. Lee and W. Zhang, "Challenges and advances in multimodal emotion recognition," in Proceedings of the 2024 International Conference on Affective Computing, 2024, pp. 95–105.

[4] Q. Cao, M. Hou, B. Chen, Z. Zhang, and G. Lu, "Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6334–6338.

[5] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recog- nition using mfcc," in International conference on computer graphics, simulation and modeling, vol. 9, 2012.

[6] B. V. Kumar, R. Jayavarshini, N. Sakthivel, A. Karthiga, R. Narmadha, and M. Saranya, "Evaluation of deep architectures for facial emotion recognition," in International Conference on Computer Vision and Image Processing. Springer, 2021, pp. 550–560.

[7] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in 2018 IEEE Conference on multimedia information processing and retrieval (MIPR). IEEE, 2018, pp. 196–201.

[8] G. Anbarjafari, F. Noroozi, T. Sapinski, and D. Kaminska, "Vocal-based emotion recognition using random forests and decision tree," 2017.

[9] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep cnn," IET Signal Processing, vol. 12, no. 6, pp. 713–721, 2018.

[10] E. Avots, T. Sapin(´)ski, M. Bachmann, and D. Kamin(´)ska, "Audiovisual emotion recognition in wild," Machine Vision and Applications, vol. 30, no. 5, pp. 975–985, 2019.

[11] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," Pattern Recognition Letters, vol. 146, pp. 1–7, 2021.

[12] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Ko ir, "Audio-visual emotion fusion (avef): A deep efficient weighted approach," Information Fusion, vol. 46, pp. 184–192, 2019.

[13] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," IEEE access, vol. 9, pp. 7943–7951, 2021.

[14] S. R. Livingstone and F. A. Russo, "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, vol. 13, no. 5, p. e0196391, 2018.

[15] S. Verbitskiy, V. Berikov, and V. Vyshegorodtsev, "Eranns: Efficient residual audio neural networks for audio pattern recognition," Pattern Recognition Letters, vol. 161, pp. 38–44, 2022.

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