A More Efficient Inference Model for Multimodal Emotion Recognition

Liang Jia JIALIANG@SAU.EDU.CN
Jin Tan TANJIN@STU.SAU.EDU.CN
Lijin Qi QILIJIN@STU.SAU.EDU.CN
Mingwen Lin LINMINGWEN@STU.SAU.EDU.CN
School of Electronic Information Engineering, Shenyang Aerospace University, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

With the widespread adoption of the Internet and mobile Internet, an increasing number of individuals are expressing their emotions on short-video platforms. Contemporary multimodal emotion analysis technologies facilitate a more comprehensive recognition and understanding of emotions through the analysis of various data sources including text, facial expressions, audio, hand gestures, among others. Consequently, the significance of sentiment analysis is becoming increasingly pronounced. However, existing research indicates that most emotion analysis techniques are not sufficiently rapid and efficient in light of the exponential proliferation of short video content. In addition, most sentiment analysis models demonstrate significant differences in the contribution of each modality, with text and visual modalities often exerting a greater influence than audio modes. Furthermore, in the pursuit of heightened accuracy, certain models are designed to be exceedingly complex, while others prioritize swift reasoning at the expense of accuracy. This paper proposes a more efficient multimodal sentiment analysis model, presenting three distinct advantages. Firstly, residual-free connectivity modules capable of extracting 3-D attentional weights are proposed to process visual modal features, maintaining accuracy while improving inference efficiency. Secondly, adoption of multi-scale hierarchical context aggregation (aggregation followed by interaction) for audio modality to capture coarse- and fine-grained audio contextual information through multilevel aggregation, thereby enriching audio modality features and minimizing disparities between modalities' contributions. Finally, attainment of a superior balance between accuracy and speed, thereby enhancing adaptability to the fast-paced short video environment and meeting the burgeoning demand for video content processing.

Keywords: emotion analysis, multimodal, efficient, inference.

1. Introduction

In daily life, people can express their emotions not only through text but also through facial expressions and vocal tone. Therefore, modern multimodal emotion analysis techniques can enhance the identification and understanding of emotions by analyzing various information sources such as text, facial expressions, audio, hand posture and so on. Integrating data from different modalities can improve the accuracy and richness of emotion analysis. However, existing multimodal emotion recognition techniques face several challenges. Firstly, the proliferation of short video content on social media platforms has created a need for more efficient processing methods. While the existing visual processing methods can effectively extract the emotions of the characters in the video, they often lack the necessary speed for processing the large volume of short videos. Secondly, the contribution of each modality in most of the existing emotion analysis models varies significantly. For instance, traditional audio processing methods may fail to capture subtle emotional changes in speech, leading to a lack of richness and expressiveness in the extracted audio features compared to other modalities.Finally, most of the existing emotion analysis models typically fall into two categories: those with high accuracy but complex structures and a large number of parameters, resulting in slower processing speeds, and those with fast inference speeds but simpler architectures and lower accuracy. Addressing these challenges requires developing more efficient and balanced multimodal emotion analysis models that can handle the growing volume of short video content while maintaining high accuracy.

In this paper, our goal is to design a more efficient sentiment analysis model to solve the above problems. We aim to better adapt to the fast-paced short video environment and meet the increasing demand for video content processing, and the composition of the model is shown in Fig. 1. Our approach effectively leverages textual, auditory, and visual cues for robust emotion analysis. Firstly, we leverage the state-of-the-art Albert model to process textual data. Concurrently, the visual modality adopts a fully connected model without residuals that can extract 3-D attention weights, first using the reserving and merging operation to remove the residual connections common in traditional networks, and then extracting better visual features without introducing any additional parameters, aiming to reduce the amount of computation and increase prediction speed. Additionally, our approach incorporates multi-scale hierarchical context aggregation networks for analyzing audio modalities. The network extracts contextual information across different temporal scales, thereby enhancing the contribution of audio modalities in multi-modal emotional analysis tasks. The next step involves encoding the audio modal sequence and visual modal sequence using Transformer models, followed by multi-modal weight fusion through a feed-forward neural network to obtain the prediction results. At the same time, in order to achieve a comprehensive connection between the input video and the emotion analysis process, we integrated data preprocessing into a multimodal sentiment analysis model.

Compared to existing multi-modal emotion analysis models, our proposed model exhibits the capability to enhance processing speed while maintaining high accuracy, thereby better accommodating the rapid-paced environment inherent in short video content processing, and meeting the escalating demand for efficient video content analysis.

We summarize our contribution as follows:

- Residual-free connected networks capable of extracting 3-D attentional weights are proposed to extract visual modal features to improve prediction rate while maintaining accuracy.
- Successfully ported the Focus Modulation Transformer to audio modality, which is able to perform multilevel aggregation to capture fine-grained and coarse-grained audio context information and increase the richness of audio modality information.
- After experiments on two datasets, we find that our proposed model achieves a better balance of speed and accuracy.

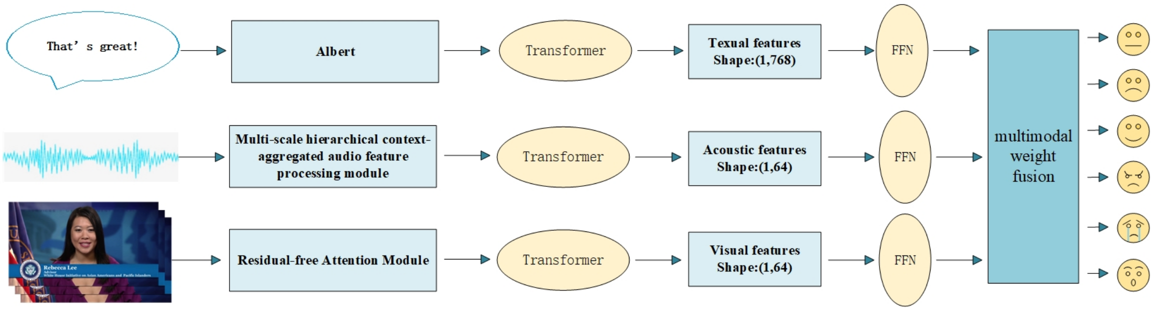
That' s great!
Albert
Transformer
Texual features Shape:(1,768)
FFN
Multi-scale hierarchical context-aggregated audio feature processing module
Transformer
Acoustic features Shape:(1,64)
FFN
Residual-free Attention Module
Transformer
Visual features Shape:(1,64)
FFN
multimodal weight fusion

Figure 1: The data flow of the proposed.

2. Related Work

Multimodal emotion analysis has emerged as a rapidly evolving and extensively researched field within computer vision in recent years. With the proliferation of deep learning, which has become a dominant force in computing, research on emotion recognition increasingly integrates knowledge from deep learning. For instance, Zadeh et al. (2018b) proposed a multi-attention recurrent network that discovers dynamic dependency relationships between different modalities through multiple attention modules, and stores the matrix representation of the dependency relationships in a mixed storage unit of LSTM. Wang et al. (2023) proposes a multimodal encoding-decoding translation network with Transformer, which uses a modality-enhanced cross-attention module to transform unnatural language features into natural language features and improve their quality. In addition, a dynamic filtering mechanism filters out error messages generated in cross-modal interactions. However, the more complex the structure of the deep network, the less computationally efficient the resulting inference is, the greater the storage memory requirement, and the less suitable it may be for real-time analysis.

Li and Tayir (2021) introduced an additional multimodal attention mechanism at the decoder side of the Transformer model to align different text and image features. Peng et al. (2022) proposed a hierarchical fusion CMCN model that utilises an image-text correlation generator to reduce errors due to the presence of erroneous correlations between images and text. However, the above studies only investigated textual modalities and visual modalities, thus ignoring audio modalities.

The evolution of multimodal emotion recognition tasks has gradually permeated various aspects of daily life, exemplified by its integration into diverse domains such as the medical field(Pan et al. (2018); Huang et al. (2021)), distance education(Wang et al. (2018); sen Wang and Wu (2011)), and transportation sector(Liu et al. (2019); Boril et al. (2010)). In the medical realm, practitioners leverage a fusion of psychological expertise and emotion recognition techniques to enhance the treatment efficacy for specific psychiatric patients. This amalgamation facilitates a more holistic understanding of patients' emotional states, thereby guiding personalized interventions and fostering improved therapeutic outcomes. Within the realm of distance education, educators leverage portable devices equipped with emotion recognition capabilities to monitor students' emotional well-being during online classes. By gauging students' emotional states in real-time, instructors can tailor the pacing and content delivery of lessons, fostering a conducive learning environment that promotes efficient knowledge absorption and student comfort. Furthermore, in the transportation sector, the deployment of emotion recognition technology enables continuous monitoring of the mental states of transportation personnel. This proactive approach not only enhances operational safety but also mitigates the occurrence of traffic accidents by promptly addressing potential emotional distress or fatigue among transportation workers. These applications underscore the transformative potential of multimodal emotion recognition technologies in enhancing human experiences and augmenting operational efficiencies across diverse domains, thereby heralding a new era of intelligent and empathetic technological integration into everyday life.

3. Proposed Methods

There are three modalities included in the model, which are text, acoustic, and visual. For textual modality, textual modality features are extracted using Albert Lan et al. (2019). For visual modality, residual-free connected networks capable of extracting 3-D attentional weights were used to obtain frame information, and then Basic Transformer Vaswani et al. (2017) was used to obtain visual sequence information. For acoustic modality, audio features are extracted for each spectrum of audio modality using focal modulation, and then the audio sequence information is obtained using Basic Transformer. Following the extraction of modality-specific features, our model employs a weighted fusion mechanism to integrate predictions from each modality and outputs the final emotion category predictions.

3.1. Residual-Free Attention Module(RFAM)

To tackle the issues of sluggish processing speeds and inefficient inference observed in current multimodal sentiment analysis models, we propose a Residual-Free Attention Module to extract visual modal features, and the general framework of this module is shown in Fig. 2. RFAM consists of a Residual-Free Block (RFB) and a 3-D attention Block (TDAB). Influenced by RMNet Meng et al. (2021), the RFB removes residual connections between nonlinear layers by reserving the input feature mappings and merging them with the output feature mappings. This method converts a network with residual connections into a rectilinear network, which can greatly improve the processing speed of the model. The TDAB is different from existing channel or spatial attention modules in that it is lightweight in that it can derive 3D attention weights for the feature maps without additional parameters. In summary, RFAM is able to process visual features at a faster rate while maintaining accuracy in visual modal feature extraction, which is consistent with our original intention. Next, we will introduce each module in detail.

3.1.1. Residual-Free Block (RFB)

Reserving: the number of channels in the input image frame is 3. In Conv1, insert a convolutional kernel initialized by the Dirac filter with the same number of channels. The Dirac filter is capable of constant mapping for each channel separately, setting the weight of the desired channel to 1 and the weight of other channels to 0. The input features are preserved by this method (is preserved to remain as after convolution), as shown on the left side of Fig. 3.

To preserve the input features, for the BN layer, it is necessary to adjust the weight ω and bias b in the BN layer so that the BN layer behaves like an identity function. Assuming that the running mean and running variance of the feature map are μ and σ^2 respectively, we set $\omega = \sqrt{\sigma^2 + \varepsilon}$ and the bias is $b = \mu$, where $\varepsilon = 10^{-5}$.

For ReLU, if the value of each input is non-negative, ReLU can be used directly; if there are negative values in the input, PReLU can be used instead of ReLU. for the input features, the alpha parameter of PReLU is set to 1 to keep the linear mapping. For convolved features, set the alpha parameter of PReLU to 0, which is equivalent to ReLU.

Merging: as shown on the right side of Fig. 3, this operation incorporate the weighted convolutional kernel initialized by the Dirac filter, which was utilized in the reserved operation of Conv1, into Conv2. This integration serves as a substitute for the residual connection typically present in Conv2. By implementing this approach, we effectively transform the network with residual connections into a linear network. Consequently, this conversion significantly enhances the processing speed and prediction efficiency of the model, as it eliminates the need for complex residual connections. This streamlined architecture facilitates smoother and more efficient model inference.

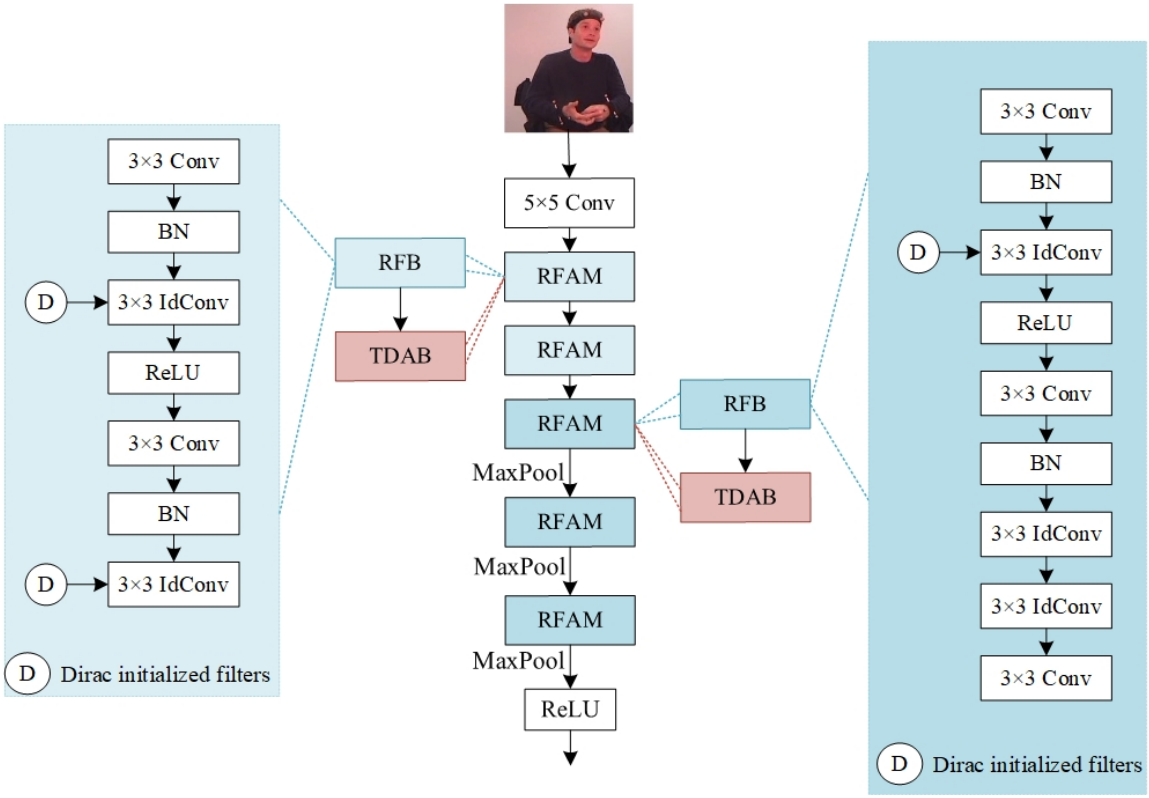

Figure 2: Overall framework diagram of RFAM.

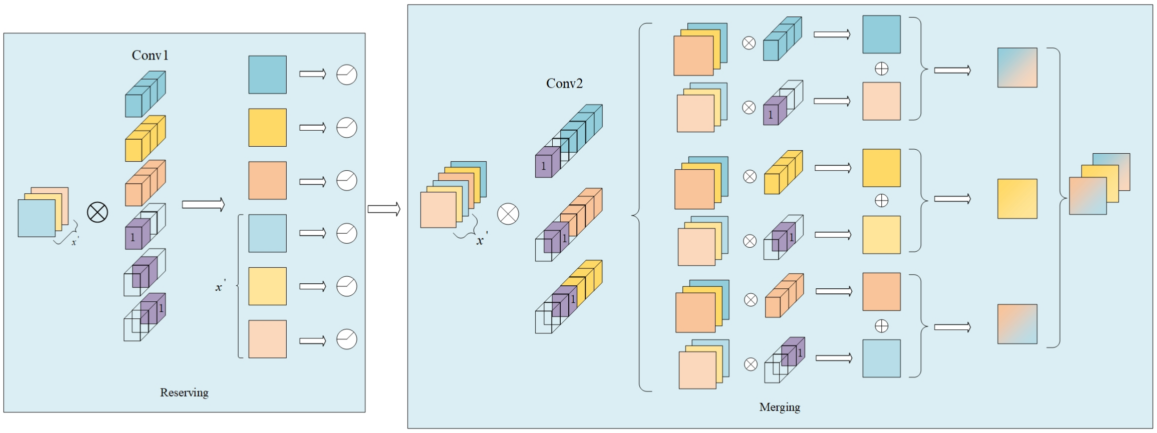

Figure 3: Framework diagram of the RFB.

3.1.2. 3-D Attention Block(TDAB)

Existing attention modules, such as spatial attention and channel attention, which can only refine features along channel or spatial dimensions, produce 1-D or 2-D weights that either treat all neurons in a channel equally or all neurons in a spatial location equally, which limits their flexibility to learn attention weights that vary across channels and across space. Secondly, these two attentional mechanisms correspond exactly to feature-based and spatial-based attention in the human brain Carrasco (2011). In the human brain, however, these two mechanisms work in tandem to jointly facilitate information selection during visual processing. In visual neuroscience, the most informative neurons are usually those that exhibit a different firing pattern than the surrounding neurons. In addition, an active neuron may also inhibit the activity of surrounding neurons, a phenomenon known as spatial inhibition Webb et al. (2005). In other words, neurons that exhibit significant spatial inhibition effects should be given higher priority (i.e., importance) in visual processing. The simplest way to find these neurons is to measure the linear separability between a target neuron and other neurons. Based on these neuroscientific findings, the minimum energy function of a target neuron t is defined as Yang et al. (2021):

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{\left(t - \hat{\mu}\right)^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

Here,$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$,$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}\left(x_i - \mu\right)^2$,$\lambda = 0.0001$. The above equation shows that the lower the energy e_t^*, the greater the difference between neuron t and peripheral neurons and the more important it is for visual processing. Therefore, the importance of each neuron can be obtained by $1/e_t^*$.

Suppose the input feature map is $\mathbf{V} \in \mathbb{R}^{B\times C\times H\times W}$, and the number of neurons on this channel is $N = H \times W - 1$. Then the importance of neurons can be expressed as:

$$y = \frac{\left(v - \mu\right)^2}{4\left(\frac{\sigma^2}{N} + \varepsilon\right)} + 0.5 \tag{2}$$

Here, $\mu = \frac{1}{N}\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W} v_{ij}$, $\sigma^2 = \frac{1}{N}\sum\limits_{i=1}^{H}\sum\limits_{j=1}^{W} (v_{ij} - \mu)^2$, $\varepsilon = 0.0001$. Finally, according to Hillyard et al. (1998), attentional modulation in the mammalian brain usually manifests itself as a gain effect on neuronal responses:

$$Output = v \odot sigmiod\,(y) \tag{3}$$

In RFAM, the RFB forms a residual-free connected rectilinear network by reserving and merging operation, and the THAB also does not introduce additional parametric quantities, i.e., it does not change the structure of the RFB. So the RFAM is able to extract visual modal features better while maintaining fast inference speed.

3.2. Multi-scale hierarchical context-aggregated audio feature processing module

To improve the effectiveness of the acoustic modality in multimodal sentiment analysis tasks, we designed a multi-scale hierarchical context-aggregation audio feature processing module. This approach addresses the observed limitations of the acoustic modality when compared to the other two modalities. Inspired by the success of the Transformer architecture in computer vision, we utilized its key component, Self-Attention(SA) Vaswani et al. (2017), to create feature representations. In SA, the attention scores between a Query and all Keys are computed through an interaction operation, and then aggregated to generate the final representation. Despite its effectiveness, the traditional SA has a critical limitation: its computational complexity grows quadratically with the length of the input sequence, due to the need to compute attention weights between all tokens. This makes it computationally expensive and potentially limits its application to long sequences. To address this challenge, we drew inspiration from the concept of focal attention Yang et al. (2024), which performs multi-level aggregation to capture both fine and coarse-grained audio contexts. This method allows for effective context-aggregation without the high computational overhead of standard SA. The structure of our audio feature extraction module, as illustrated in Fig. 4, employs focal attention to improve the computational efficiency while maintaining robust context-aggregation. This approach enables our model to process longer sequences with greater efficiency and contributes to a more balanced multimodal sentiment analysis by enhancing the role of the acoustic modality.

The initial spectrum is obtained by passing the audio signal through a set of Mel-scale filter banks. Following this, the resulting spectrogram, with dimensions $H \times W$, serves as the input $\mathbf{X}$:

$$\mathbf{y}_i = Q\,(x_i) \odot CA\,(i, \mathbf{X}) \tag{4}$$

where $Q\,(\cdot)$ is the Query for each spectrogram token and $\odot$ is the Hadamard product element-wise. $CA\,(\cdot)$ represent the Context Aggregation Function, whose output is the modulator. $Q\,(\cdot)$ retains the most valuable information from each spectrum token, while $CA\,(\cdot)$ extracts coarse-grained contextual information; they are decorrelated but combined with the modulator.

The specific process of $CA\,(\cdot)$ involves applying LN to the input $\mathbf{x}_i$ to project it into the new feature space $\mathbb{R}^{H\times W\times C}$. Then a depth-separable convolution stack with L convolution kernel sizes k_i is passed, because of its lower number of parameters and computational cost compared to normal convolution, followed by the GeLU Hendrycks and Gimpel (2016) activation function. The above depth-separable convolution as well as the GeLU function obtains a hierarchical representation of the context. At this point the receptive field is much larger compared to the convolutional kernel k_i. To capture the global context of the entire input audio, we then perform global average pooling. As a result, we get a total of i + 1 feature maps, which collectively capture the context information at different granularity levels.

$$\mathbf{O}_i = AvgPool\left(GeLU\left(DWConv\left(LN\left(\mathbf{x}_{i-1}\right)\right)\right)\right) \tag{5}$$

G_i denotes the gating weight at position i. We use the gating mechanism to control how much each query is aggregated from different levels, specifically, we use a linear layer to obtain spatial and level-aware gating weights G_i. We then perform weighted sums by multiplying by elements to obtain individual feature map $\mathbf{S}_O$ of the same size as the input $\mathbf{x}_i$. We find that focus modulation can adaptively learn contextual information from different focuses; for tokens with audio information highlighting emotional features, it focuses more on fine-grained local structure at low focus levels, while tokens in other contexts need to be perceived from higher levels of focus.

$$\mathbf{S}_O = \sum\left(\mathbf{G}_i \odot \mathbf{O}_i\right) \tag{6}$$

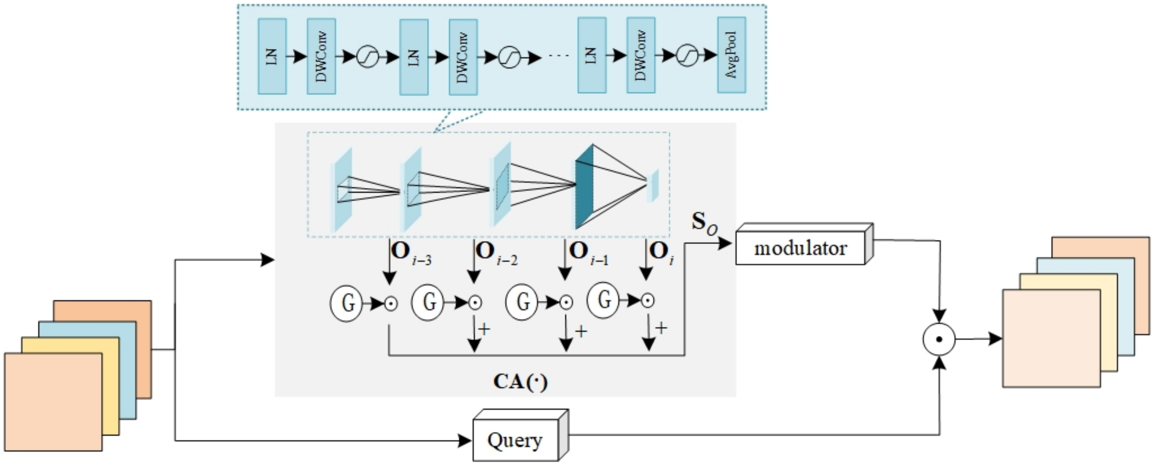

Figure 4: multi-scale hierarchical context aggregation audio feature processing module.

4. Experiments and Analyses

4.1. Experimental details

Experimental Environment The experiment was conducted using NVIDIA GeForce RTX 4090 GPU and based on PyTorch v1.8.0. Furthermore, training of the model was facilitated by the Adam optimizer Kingma and Ba (2014). Given the multi-class and multi-label nature of both datasets, binary cross-entropy loss function was employed. In our experiments, the learning rate was set to 4.5e-6, the number of epochs to 30, and the batch size to 8.

Datasets We use two datasets in the experiment: IEMOCAP Busso et al. (2008) and CMU-MOSEI Zadeh et al. (2018a). The IEMOCAP dataset is a multimodal emotional recognition data set consisting of 151 videos in which two professional actors engage in parallel dialogue in English. The data is labelled into six emotion categories: anger, happiness, excitement, sadness, frustration and neutral. Each data sample consists of three modalities: audio data sampled at 16 kHz, text transcriptions, and image frames sampled from the video at 30 Hz. There are 7380 data samples in total, and we randomly allocated 70%, 10%, and 20% of the data to the training, validation, and test sets, respectively.

CMU-MOSEI consists of 3837 videos from 1000 different speakers with six emotion categories: happy, sad, angry, fearful, disgusted, and surprised. Each data sample in this dataset consists of three modalities: audio data sampled at a rate of 44.1 kHz, text transcripts, and image frames sampled from the videos at a frequency of 30 Hz. In total, there are 23259 samples in the dataset.

In subsequent experiments, the same dataset partitioning method was also used for all the baselines.

In this paper, we use the following five state-of-the-art models in previous work as the baselines.

LF-LSTM: Lost-fusion using LSTM.

LSTMLF-TRANS: Lost-fusion using transformers.

MulT Tsai et al. (2019): The structure of cross-modal transformer was used to construct the relationship between different modalities, and after obtaining the multimodal fusion information, the three sets of features are combined for prediction.

FE2E Dai et al. (2021): Visual and audio features are extracted using VGG16 and text features are extracted using Albert.

V2EM Wei et al. (2023): The hierarchical attention spectrum computing module was used to obtain detailed spectral information, visual features were extracted by RepVGG, and text features were extracted by Albert.

Evaluation Metrics On the IEMOCAP dataset, we use Accuracy and F1 scores to evaluate our proposed model; on the CMU-MOSEI dataset, we use Weighted Accuracy ($WA_{CC.}$) instead of Accuracy because this dataset contains much more negative samples than positive samples on each emotion category if Normal Accuracy is used, the model will still get a good score when predicting all samples as negative. The formula for weighted accuracy is:

$$WA_{CC.} = \frac{TP \times N/P \ + TN}{2N} \tag{7}$$

where P is the overall positive sample size, TP is the true positive sample size, N is the overall negative sample size, and TN is the true negative sample size.

4.2. Experimental results and analyses

4.2.1. Accuracy and F1

Comparison Experiments Table 1 shows the results of various models on the IEMOCAP dataset. Compared with other models, the accuracy of the model proposed in this paper has been significantly improved (16.42% and 10.94% compared with LF-LSTM and LF-TRANS, 9.84% compared with MuIT, 4.65% compared with FE2E), which indicates the structural excellence of the model proposed in this paper. Because V2EM has the best performance among existing SOTA models, and the visual mode processing method it uses is to reparameterize the multi-branch topology to RepVGG similar to VGG(single branch) during deployment, we choose V2EM as the reference model. The model proposed in this paper is significantly better than V2EM, with an accuracy increase of 3.81%. F1 scores increased by 4.34%. In Table 2, we further evaluate the results of each model on the CMU-MOSEI dataset. In terms of weighted accuracy, the model proposed in this paper is also significantly superior compared with other models (10.74% compared with LF-LSTM, 8.87% compared with LF-TRANS, 8.11% compared with MuIT, 7.03% compared with FE2E). 5.86% improvement compared to the current SOTA model V2EM, with the same trend in terms of F1 scores.

Table 1: Results on the IEMOCAP dataset. We report the mean $A_{CC.}$ and F1 scores for the six emotion categories, and the speed is the total time(s) to test the 1481 samples.

Model	$A_{CC.}$	F1	Speed(s)
LF-LSTM	73.25	47.55	-
LF-TRANS	76.87	49.46	-
MuIT Tsai et al. (2019)	77.64	56.87	-
FF2E Dai et al. (2021)	81.49	55.76	146.62
V2EM Wei et al. (2023)	82.15	55.73	70.45
Ours	**85.28**	**58.15**	**66.35**

Table 2: Results on the CMU-MOSEI dataset. We list the mean $WA_{CC.}$ and F1 scores for the six emotion categories, and the speed is the total time(s) to test the 4188 samples.

Model	$WA_{CC.}$	F1	Speed(s)
LF-LSTM	66.84	44.62	-
LF-TRANS	67.99	45.35	-
MuIT Tsai et al. (2019)	68.47	45.28	-
FF2E Dai et al. (2021)	69.16	**47.46**	142.13
V2EM Wei et al. (2023)	69.92	44.83	111.69
Ours	**74.02**	46.37	**69.99**

Ablation Experiments In order to test the reliability of the proposed multi-scale hierarchical context aggregation audio feature processing module and the residual-free attention module for visual modal feature extraction, we also conducted ablation experiments. We compare with the existing SOTA model V2EM. Table 3 and 4 show the results of each model on the IEMOCAP dataset and the CMU-MOSEI dataset respectively. The multi-scale hierarchical context aggregation audio feature processing module (V+A(Ours)+T) proposed in this paper improves the accuracy of IEMOCAP data set by 1.19% and F1 score by 2.96% compared with the hierarchical concern spectrum computing module (V+A+T) in V2EM. On the CMU-MOSEI dataset, the weighted accuracy is increased by 3.56%, and the F1 score is increased by 2.74%. It can be seen that the multi-scale hierarchical context aggregation audio feature processing module proposed in this paper has greatly improved the accuracy, indicating that the module can perform multi-level aggregation to capture fine-grained and coarse-grained audio context as scheduled. More abundant audio feature information can be obtained, so that the inflection change of voice intonation in audio can better predict emotion. At the same time, the performance of the V+A(Ours)+T model on the two data sets is not much different or even better than that of V(Ours)+A+T, indicating that the contribution of audio modes is not at a disadvantage in the model proposed in this paper. Compared with the RepVGG-based single branch inference module in V2EM, the residual-free attention module V(Ours)+A+T proposed in this paper improves the accuracy of IEMOCAP data set by 1.61% and F1 score by 0.9%. Accuracy improved by 1.36% on the CMU-MOSEI dataset. Through the analysis of the above results, it can be seen that although the residual-free attention module proposed in this paper has a small improvement in accuracy, the original intention of this paper is to improve the processing speed of the model, and the above results are not inconsistent with the original intention of this paper.

Table 3: Results on the IEMOCAP dataset. The base V+A+T model uses the Hierarchical Attention Spectral Computing module to obtain spectral information, RepVGG to extract visual features, and Albert to extract textual features; the Multi-scale Hierarchical Contextual Aggregation module to process audio features in V+A(Ours)+T; and the Residual-Free Attention module to process visual modalities in V(Ours)+A+T. $A_{CC.}$ and F1 scores as evaluation metrics.

Modalities	$A_{CC.}$	F1	Speed(s)
V+A+T	82.15	55.73	70.45
V+A(Ours)+T	83.13	57.38	**65.72**
V(Ours)+A+T	83.47	56.23	67.09
V(Ours)+A(Ours)+T	**85.28**	**58.15**	66.35

4.2.2. Rate of inference

Comparison Experiments Table 1 and 2 also show the reasoning speed of each model. It can be seen from the table that the reasoning speed of the proposed model is the fastest. Compared with the existing SOTA model V2EM, the reasoning speed of the proposed model is increased by 5.82% on the IEMOCAP data set and 37.34% on the CMU-MOSEI data set, indicating that the proposed model has reached the expected goal. It can better adapt to the fast-paced short video environment and achieve fast and effective reasoning.

Table 4: Results on the CMU-MOSEI dataset. The base V+A+T model uses the Hierarchical Attention Spectral Computing module to obtain spectral information, RepVGG to extract visual features, and Albert to extract textual features; the Multi-scale Hierarchical Contextual Aggregation module to process audio features in V+A(Ours)+T; and the Residual-Free Attention module to process visual modalities in V(Ours)+A+T. $WA_{CC.}$ and F1 scores as evaluation metrics.

Modalities	$WA_{CC.}$	F1	Speed(s)
V+A+T	69.92	44.83	111.69
V+A(Ours)+T	72.41	46.06	92.54
V(Ours)+A+T	70.87	44.62	87.47
V(Ours)+A(Ours)+T	**74.02**	**46.37**	**69.99**

Ablation Experiments As shown in Table 3 and 4, the application of the residual-free attention module proposed in this paper effectively improves the reasoning efficiency of the two data sets. Specifically, on the IEMOCAP dataset, V(Ours)+A+T required 67.09s in the test set, an increase of 4.77% compared to the baseline. Similarly, on the CMU-MOSEI dataset, V(Ours)+A+T required 87.47s in the test set, an increase of 21.69% compared to the baseline. The above results show that the residual-free attention module proposed in this paper has advantages in processing speed.

4.2.3. Number of parameters

Table 5 shows the comparison between the existing SOTA model V2EM and the model proposed in this paper in terms of the number of parameters in the three modes of text, audio and vision. It can be seen from the table that the number of parameters in the audio and visual modes of the proposed model is less than that of the V2EM model, which indicates that the proposed model can better adapt to the fast-paced short video environment. Meet the growing demand for video content processing.

Table 5: Comparison of the number of V2EM and Ours model parameters.

Modality	V2EM	Ours
V	1.81M	0.73M
A	67.25M	49.89M
T	11.68M	11.68M

5. Conclusion

In this paper, we propose a residual-free attention module to process visual modality, in which the residual-free module and the 3-D attention module effectively improve the speed and efficiency of emotion analysis while maintaining the recognition accuracy. In addition, we also propose a multi-scale hierarchical contextual aggregation audio feature processing module, which introduces the idea of 'aggregation before interaction' into the audio feature extraction process, so as to extract richer audio feature information and increase the contribution of audio modality in the emotion analysis task. Finally, through experiments on two datasets, we find that the model achieves a better balance between speed and accuracy. In summary, the model proposed in this paper has some academic value and practical application value.

However, the proposed model still has some limitations.The video scenes in the IEMOCAP dataset and the CMU-MOSEI dataset are relatively homogeneous, i.e., the environment in the videos is relatively quiet and almost noiseless, which is different from some of the other datasets that were collected in the field. However, in real life such quiet scenes are after all a minority, and most scenes are still noisy. In addition, our model is not fully end-to-end. In the future, we will continue to optimise our model to take visual feature extraction for complex backgrounds, audio feature extraction in noisy environments into account, and form a fully end-to-end sentiment analysis system to better adapt to the fast-paced rhythm of short videos.

Acknowledgement

This work received support from the National Science Foundation of China (No.61671310) and the Aviation Science Foundation under Grant 2019ZC054004.

References

Hynek Boril, Seyed Omid Sadjadi, Tristan Kleinschmidt, and John H. L. Hansen. Analysis and detection of cognitive load and frustration in drivers' speech. In *Interspeech*, 2010. URL https://api.semanticscholar.org/CorpusID:16011794.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008. URL https://api.semanticscholar.org/CorpusID:11820063.

Marisa Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525, 2011. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2011.04.012. URL https://www.sciencedirect.com/science/article/pii/S0042698911001544. Vision Research 50th Anniversary Issue: Part 2.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.417. URL https://aclanthology.org/2021.naacl-main.417.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv: Learning*, 2016. URL https://api.semanticscholar.org/CorpusID:125617073.

Steven A. Hillyard, Edward K. Vogel, and Steven J. Luck. Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 353 1373:1257–70, 1998. URL https://api.semanticscholar.org/CorpusID:3570968.

Haiyun Huang, Qiuyou Xie, Jiahui Pan, Yanbin He, Zhenfu Wen, Ronghao Yu, and Yuanqing Li. An eeg-based brain computer interface for emotion recognition and its application in patients with disorder of consciousness. *IEEE Transactions on Affective Computing*, 12(4):832–842, 2021. doi: 10.1109/TAFFC.2019.2901456.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2019. URL https://api.semanticscholar.org/CorpusID:202888986.

Lin Li and Turghun Tayir. Multimodal machine translation enhancement by fusing multimodal-attention and fine-grained image features. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 267–272, 2021. doi: 10.1109/MIPR51284.2021.00050.

Weihuang Liu, Jinhao Qian, Zengwei Yao, Xintao Jiao, and Jiahui Pan. Convolutional two-stream network using multi-facial feature fusion for driver fatigue detection. *Future Internet*, 11(5), 2019. ISSN 1999-5903. URL https://www.mdpi.com/1999-5903/11/5/115.

Fanxu Meng, Hao Cheng, Jia-Xin Zhuang, Ke Li, and Xing Sun. Rmnet: Equivalently removing residual connection from networks. *ArXiv*, abs/2111.00687, 2021. URL https://api.semanticscholar.org/CorpusID:240353976.

Jiahui Pan, Qiuyou Xie, Haiyun Huang, Yanbin He, Yuping Sun, Ronghao Yu, and Yuanqing Li. Emotion-related consciousness detection in patients with disorders of consciousness through an eeg-based bci system. *Frontiers in Human Neuroscience*, 12, 2018. ISSN 1662-5161. doi: 10.3389/fnhum.2018.00198. URL https://www.frontiersin.org/articles/10.3389/fnhum.2018.00198.

Cheng Peng, Chunxia Zhang, Xiaojun Xue, Jiameng Gao, Hongjian Liang, and Zhengdong Niu. Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. *Tsinghua Science and Technology*, 27(4):664–679, 2022. doi: 10.26599/TST.2021.9010055.

Wan sen Wang and Jiabin Wu. Notice of retractionemotion recognition based on cso&svm in e-learning. *2011 Seventh International Conference on Natural Computation*, 1:566–570, 2011. URL https://api.semanticscholar.org/CorpusID:4807587.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Fan Wang, Shengwei Tian, Long Yu, Jing Liu, Junwen Wang, Kun Li, and Yongtao Wang. Tedt: Transformer-based encoding-decoding translation network for multimodal sentiment analysis. *Cognitive Computation*, 15(1):289–303, JAN 2023. doi: 10.1007/s12559-022-10073-9.

Shui-Hua Wang, Preetha Phillips, Zheng-Chao Dong, and Yu-Dong Zhang. Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm. *Neurocomputing*, 272:668–676, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2017.08.015. URL https://www.sciencedirect.com/science/article/pii/S0925231217313644.

Ben S. Webb, Neel T. Dhruv, Samuel G. Solomon, Chris Tailby, and Peter Lennie. Early and late mechanisms of surround suppression in striate cortex of macaque. *Journal of Neuroscience*, 25(50):11666–11675, 2005. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3414-05.2005. URL https://www.jneurosci.org/content/25/50/11666.

Qinglan Wei, Xuling Huang, and Yuan Zhang. Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference. *IEEE Transactions on Broadcasting*, 69(1):10–20, 2023. doi: 10.1109/TBC.2022.3215245.

Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:235825945.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Annual Meeting of the Association for Computational Linguistics*, 2018a. URL https://api.semanticscholar.org/CorpusID:51868869.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018b. URL https://arxiv.org/abs/1802.00923.