Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

Multimodal Emotion Analysis Model based on Interactive Attention Mechanism

Bainan Zhou [1], Xu Li [2]

[1] School of Information Science and Engineering, Dalian Polytechnic University, Dalian, China
[2] Engineering training center, Dalian Polytechnic University, Dalian, China

Abstract: In traditional multi-modal sentiment analysis, feature fusion is usually achieved by simple splicing, and multi-modal sentiment analysis is only trained as a single task, without considering the contribution of inter-modal information interaction to sentiment analysis and the correlation and constraint relationship between multi-modal and single-modal (text, video and audio) tasks. Therefore, a multi-task model based on interactive attention mechanism is proposed in this paper, which uses inter-modal attention mechanism and single-modal self-attention mechanism to train multi-modal sentiment analysis and single-modal sentiment analysis together, so as to make full use of inter-modal and inter-task information sharing, mutual complement, and reduce noise to improve the overall recognition performance. Experiments show that the proposed model performs well on MOSI and MOSEI common data sets for multimodal sentiment analysis.

Keywords: Multimodal sentiment analysis; Attention mechanism; Multi-mission joint training.

1. Introduction

Traditional sentiment analysis usually refers to the text sentiment analysis task [1], but with the development of social media, the types and forms of data are diversified. In recent years, multimodal sentiment analysis has become a hot research topic due to the limitations of traditional text sentiment analysis only using text data to analyze emotional information. Typically, multimodal data consists of three parts: text, video, and audio. This paper mainly studies the multimodal sentiment analysis task of three modes.

Multimodal data representation and multimodal feature fusion are two important types of work in multimodal sentiment analysis [2]. In recent years, many research methods have been proposed for these two types of important work. For multi-modal data representation, neural networks are usually used to represent data of different modes respectively. Poria et al. [3] proposed a model based on multi-core learning, which uses deep convolutional neural networks to extract text features and integrate them with other (video and audio) modes in the form of splicing. Cambria et al. [4] proposed a general multi-modal sentiment analysis framework, which consists of feature learning within modes and feature splicing of multi-modes, and laid a foundation for future research. Zadeh et al. [5] introduced a multimodal dictionary in order to better understand the interaction between face, gesture and audio in expressing emotion, and proposed a MOSI multimodal emotion analysis dataset. Later, Zadeh et al. [6] proposed tensor fusion network to learn features within and between modes. And the precision of multimodal sentiment analysis is improved on MOSI data set. In order to consider more context information, Poria et al. [7] developed a framework based on LSTM, which uses context information to capture the dependency between modes. In the later work of Zadeh et al. [8], multiple attention blocks were used to obtain the information between the three modes and have good performance on different data sets. After the emergence of Transformer [9], Tsai et al. [10] proposed a cross-modal Transformer model to strengthen the feature learning of target modes through the cross-modal attention mechanism. However, the above research on multimodal data representation does not consider the correlation between different modal data in multimodal data representation, and only focuses on the representation of different modal data respectively.

In the study of multi-modal sentiment analysis, the multi-modal feature fusion usually adopts the fusion way of early fusion (fusion after feature learning). Poria et al. [3] and Cambria et al. [4] represent the multi-modal data by neural network, and then use a simple splicing way to carry out feature fusion. Later, Liu et al. [19] proposed a low-rank multi-mode fusion method, using low-rank vector to represent multi-mode data, and then using Cartesian product to achieve feature fusion. However, the simple fusion method does not consider the contribution degree of different modal data to the results of sentiment analysis, and the common multi-modal sentiment analysis task is usually a single-task training method, and does not consider the correlation and constraint relationship between multi-modal sentiment analysis and each modal sentiment analysis task.

To solve the above problems, this paper proposes a multi-task emotion analysis model based on attention mechanism [12], which uses the correlation between single-mode and multi-mode emotion analysis tasks to achieve emotion analysis tasks. In natural language tasks, context information has a great impact on the overall recognition performance [13], so in single-mode emotion classification tasks, in this paper, the self-attention mechanism is used to consider the semantic information of the single mode itself, and the interactive attention mechanism is used to achieve more fully multi-mode feature fusion.

The innovation of this paper mainly includes two parts: attention mechanism and multi-task joint training. The attention mechanism consists of two parts: intermodal attention mechanism and unimodal self-attention mechanism.

1) Attention mechanism between modal: through two attention mechanism between modal corresponds to a weight of the other two kinds of mode, give full consideration to the modal semantic association between the information interaction, making models focus on the more important

information, achieve the purpose of reducing noise.

2) Single-mode state since the attention mechanism: by using single mode state context information, in order to realize the purpose of information gain.

3) Multitasking joint training: learn by multitasking to multimodal emotional analysis tasks and single-mode state analysis task joint training in order to realize the related tasks associated with constraints, and the weight given dynamic loss for different mission, make model can focus on loss value larger task, optimizing the training process, speed up the model convergence.

2. Method

2.1. Problem definition

In the multimodal sentiment analysis task, the main task is to determine the emotional polarity or intensity of the video segment, so generally, multimodal sentiment analysis can be regarded as either a classification task or a regression task. In this paper, multimodal emotion analysis is regarded as a regression task to judge the emotional intensity of video clips.

The model input in this paper is unimodal original sequence $X_m \in \mathbb{R}^{l_m \times d_m}$, where X_m represents the original sequence of input, l_m represents the length of the sequence, and d_m represents the dimension represented by the vector of the input sequence. In this paper, $m \in \{a, v, t\}$, where a is an audio sequence, v is an image sequence, and t is a text sequence. Output for multiple modal emotional intensity $\hat{y}_m \in R$, in the stage of training this model with the other three output $\hat{y}_n \in R$ where $n \in \{a, v, t\}$, this paper uses only $\hat{y}_m$ as the final prediction results.

2.2. Integral structure

The overall structure of the model in this paper is shown in Figure 1. The model is mainly divided into two parts: multi-modal sentiment analysis and three independent single-modal sentiment analysis. Multi-task [11] learning mode is adopted. Multi-modal sentiment analysis is used as the main task in multi-task learning, and three independent single-modal sentiment analysis is used as the auxiliary task in multi-task learning. The learning mode of sharing underlying representation is adopted among different tasks.

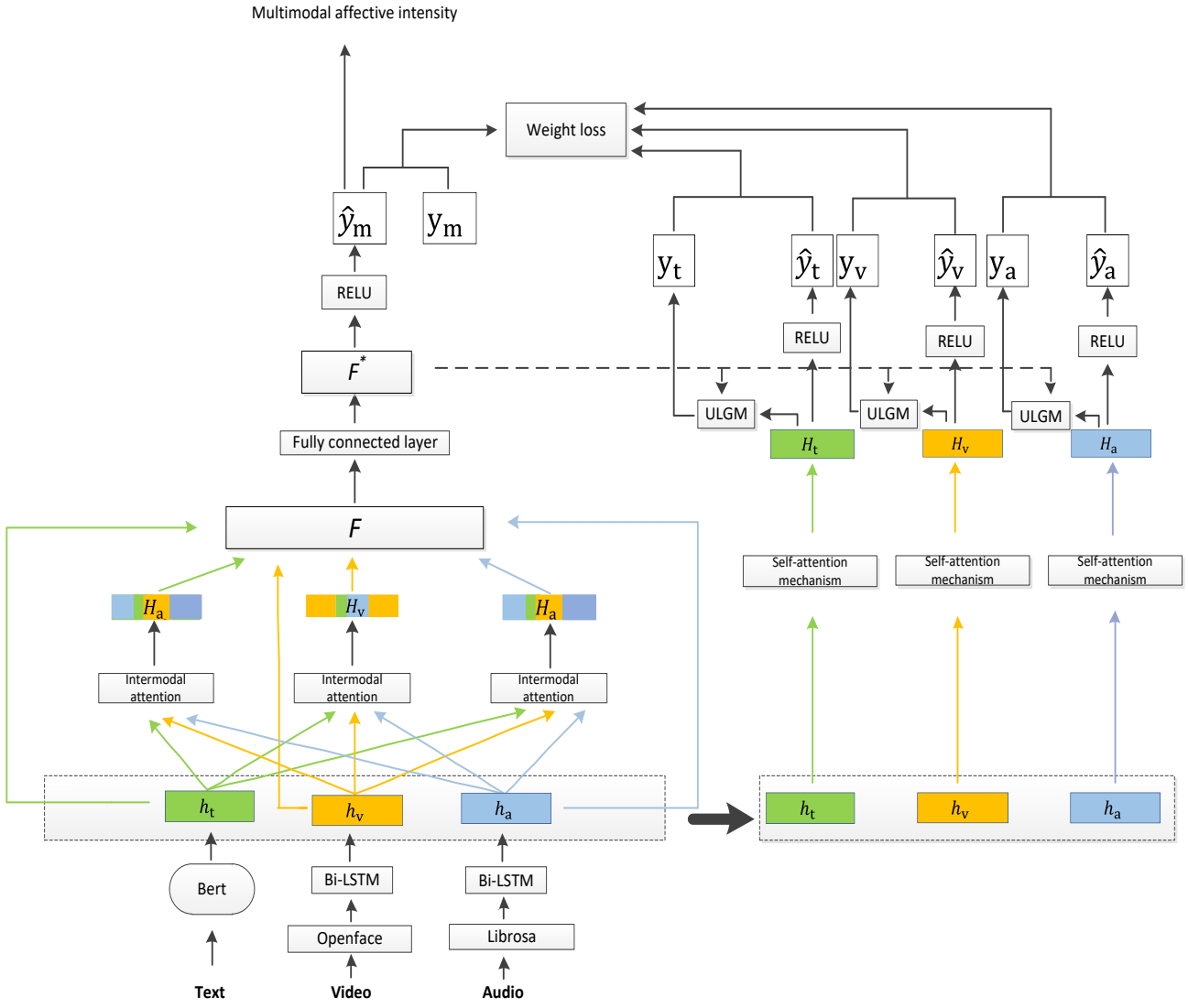

Figure 1. Overall model structure

2.3. Multimodal data representation

Firstly, the data information of three modes (text, video and audio) is extracted from the video segment, and the input sequence X_m of the three modes is encoded into a vector representing h_m, where $m \in \{a, v, t\}$. Specific implementation method: For text-like data, BERT [13] is used in this paper to encode the input sequence, and the vector representation of the [CLS] flag bit in the last layer is taken

as the vector representation of the whole sentence. For audio data and image data, bidirectional LSTM [16] is used to encode non-text data features according to previous work [14,15]. The specific expression is as formula (1), formula (2).

$$\mathrm{h_t} = \mathrm{BERT}(\mathrm{X_t}; \theta_{\mathrm{t}}^{\mathrm{BERT}}). \quad (1)$$

$$\mathrm{h_m} = \mathrm{biLSTM}\left(\mathrm{X_m}; \theta_{\mathrm{m}}^{\mathrm{biLSTM}}\right) \mathrm{m} \in \{\mathrm{v}, \mathrm{a}\}. \quad (2)$$

h_{t} is the text data representation, and h_m is the non-text data representation? Three types of data representation are the underlying representation shared by multitasking learning.

2.4. Single-mode self-attention mechanism

Single-mode attention mechanism is applied to the auxiliary task part of multi-task learning. It mainly uses the context information of single-mode and only considers the importance of single-mode to emotion classification, without considering the relationship between modes. Concrete implementation way: with specific data representation of single-mode states $h_n \in \mathbb{R}^{l_n \times \mathrm{d}_n}$ (h_{t} text data representation and h_m non-text data representation) which l_{n} sequence length, says d_n said input sequence data representation vector dimensions.

The original input unimodal data represents h_n. By taking advantage of its own semantic correlation and transposing the dot product with its own unimodal data representation, the semantic correlation matrix M is calculated, and higher weight is given to the part with greater semantic correlation. After the normalization of the semantic correlation matrix, the unimodal self-attention weight matrix A is obtained.

$$M = h_n \cdot h_n^T. \quad (3)$$

$$A = softmax(M). \quad (4)$$

The dot product of the attention weight matrix A containing semantic relevance and the original unimodal data representation h_n is obtained to obtain a new unimodal data representation H_n containing contextual information.

$$H_n = A \cdot h_n. \quad (5)$$

Finally, the resulting new data representation is used in the auxiliary task section.

2.5. Intermodal attention mechanism

Intermodal attention mechanism is applied to the main task of multi-task learning to better learn the hidden information between modes according to the information interaction between modes. Taking text data as an example, the specific implementation method is as follows: After the text data representation h_{t} is obtained, the semantic correlation between modes is used to calculate the correlation degree of text corresponding to audio and video respectively (denoted by semantic correlation matrix M_1 and M_2, and the part corresponding to audio or video correlation degree is given high weight), and the two semantic correlation matrices are normalized. The attention weight matrices A_1 and A_2 corresponding to video and audio are obtained.

$$M_1 = h_{\mathrm{t}} \cdot (h_{\mathrm{a}})^{\mathrm{T}}. \quad (6)$$

$$M_2 = h_{\mathrm{t}} \cdot (h_{\mathrm{v}})^{\mathrm{T}}. \quad (7)$$

$$H_n = softmax(M_1). \quad (8)$$

$$A_2 = softmax(M_2). \quad (9)$$

The attention weight matrix A_1 and A_2 are dot product with the original single-mode text data representation, so that the new data representation contains the semantic relevance weight information of text data, audio data and video data, and two data representations corresponding to video and audio text are obtained.

$$H_{\mathrm{TV}} = A_1 \cdot h_{\mathrm{t}}. \quad (10)$$

$$H_{\mathrm{TA}} = A_2 \cdot h_{\mathrm{t}}. \quad (11)$$

In order to make the text data representation contain more intermodal information, the two data representations are fused in the way of concatenation, and the new text data representation H_{T} is obtained.

$$H_{\mathrm{t}} = H_{\mathrm{TV}} + H_{\mathrm{TA}}. \quad (12)$$

Similarly, for audio and video data and text data processing the same way, respectively by computing the semantic correlation matrix and weight matrix, attention to get new data respectively said H_{a} and H_{a}, finally said the new data for main task.

2.6. Multi-mission joint training

For multi-task joint training, multi-modal sentiment analysis is taken as the main task, single-modal sentiment analysis as the auxiliary task, and the auxiliary task only exists in the training stage. The main task and the auxiliary task share the underlying data representation part, and the loss of the main task and the loss of the auxiliary task are taken as the total loss of the model to optimize together.

As for the chief officer, after obtaining the single mode data representation h_n, due to the different dimension of feature vectors extracted from the multi-mode data, this paper puts three multi-mode feature vectors through the full connection layer to avoid the influence of spatial features of vectors on the calculation, and then makes use of semantic correlation between modes to make inter-modal attention mechanism for the three modal data respectively. The new data representation H_n of three modes is obtained, where $n \in \{\mathrm{a}, \mathrm{v}, \mathrm{t}\}$, and the new single-mode data representation is connected with the original data representation to achieve multi-mode feature fusion to get$F$.

$$F = H_{\mathrm{a}} + H_{\mathrm{t}} + H_{\mathrm{v}} + h_{\mathrm{t}} + h_{\mathrm{a}} + h_{\mathrm{v}} \quad (13)$$

The fused feature vector F contains two parts of information: the first is the original data representation, which contains the context information of the three modes after the time series modeling of the three-modal data by using the bidirectional LSTM neural network. On the other hand, using the intermodal attention mechanism, the new data representation includes semantic correlations between the three modes. Therefore, the feature vector F can contain more information and reduce noise compared with the previous fusion method.

Relu activation function is sparse, which enables the sparse model to better mine relevant features and fit training data. Therefore, Relu activation function is selected as the activation function in this paper. The fused data represents F through the Relu activation function to get F^*.

$$F^* = Relu(W_1^T F + b_1). \quad (14)$$

Where $W_1 \in R^{l_{\mathrm{n}} \times 2(d_{\mathrm{a}} + d_{\mathrm{v}} + d_{\mathrm{t}})}$, F^* is used to predict the emotional intensity. Where $W_2 \in R^{2(d_{\mathrm{a}} + d_{\mathrm{v}} + d_{\mathrm{t}}) \times 1}$.

$$\hat{\mathrm{y}}_{\mathrm{m}} = \mathrm{W_2^T F^*} + \mathrm{b_2}. \quad (15)$$

$$y_n = \mathrm{ULGM}(y_{\mathrm{m}}, F^*, H_n). \quad (16)$$

For auxiliary tasks, in order to reduce the internal noise of single-mode data, a new single-mode data representation H_n is obtained by using the single-mode self-attention mechanism. Similarly, Relu activation function is used in this paper to make the single-mode data representation non-linear. The results of activation function are used to predict emotion. $\hat{y}_n = W_2^T \cdot [\mathrm{Relu}(W_1^T h_n + b_1)] + b_2)$, $W_1 \in R^{l_n \times \mathrm{d}_n}$、 $W_2 \in R^{\mathrm{d}_n \times 1}$ where $n \in \{\mathrm{a}, \mathrm{v}, \mathrm{t}\}$, according to the YU's work before [13], The self - generated label method is used to generate single - mode label.

2.7. Weight loss allocation

Finally, multi-modal sentiment analysis is trained jointly with three single-modal sentiment analysis. The absolute value of the difference between the label generated by single mode and the predicted label is taken as the loss value of single mode, and the loss of single-modal sentiment analysis subtask is assigned a weight w_s^i, where $s \in \{\mathrm{a}, \mathrm{v}, \mathrm{t}\}, w_s^i = tanh(|y_s^{(i)} - y_{\mathrm{m}}|)$ said the subtasks s first I samples, when the tag model focuses more on multimodal emotional labels are different parts. The loss of the single-mode sentiment analysis subtask is:

$$L_s = w_s^i \times \left|\hat{y}_s^{(i)} - y_s^{(i)}\right|. \quad (17)$$

For the loss of multimodal sentiment analysis of the main task, this paper adopts the mean value of the total loss, where N represents the total number of training samples.

$$L_{\mathrm{m}} = \frac{1}{N} \times \sum_i^n \left|\hat{y}_{\mathrm{m}}^i - y_{\mathrm{m}}^i\right|. \quad (18)$$

For the loss of each task, this paper calculates a loss weight w_L_s where $s \in \{\mathrm{a}, \mathrm{v}, \mathrm{t}, \mathrm{m}\}$ and assigns each task separately.

$$w\_L_s = \frac{L_s}{\sum_s^{\{\mathrm{a,v,t,m}\}} L_{\mathrm{s}}} L_{\mathrm{s}}. \quad (19)$$

It indicates that the larger weight loss is given to the part with larger loss to accelerate the convergence of the model and accelerate the training process. The total loss of the model is:

$$L = \sum_s^{\{\mathrm{a,v,t,m}\}} w\_L_s. \quad (19)$$

3. Experimental Setting

3.1. Dataset

The datasets used in this paper are MOSI and MOSEI common data sets for multimodal sentiment analysis. The basic information of the two data sets is shown in Table 1.

Table 1. Basic dataset information

Dataset	Train	Valid	Test	All
MOSI	1284	299	686	2199
MOSEI	16323	1871	4659	22856

MOSI data set: MOSI data set is a commonly used benchmark data set for multimodal sentiment analysis. It is a video blog mainly about film review videos on YouTube collected by Zadeh et al. [35]. A total of 93 videos including 2199 video clips were randomly collected. The labels of these videos were marked and averaged by five markers from Amazon's crowdsourcing platform, and marked as seven emotional tendencies ranging from -3 (negative) to +3 (positive).

MOSEI dataset: The MOSEI [36] dataset expands the content of the data, including 3228 videos, 23,453 sentences, 1000 narrators, 250 topics, with a total duration of 65 hours and a total of 22,856 video segments. This data set provides data annotation for emotion classification of 2, 5 and 7 categories and also marks emotion intensity, which is marked as seven emotional tendencies from -3 (negative) to +3 (positive).

3.2. Contrast model

In order to fully verify the performance of the model, the recent multimodal sentiment analysis benchmark model is compared with the proposed model under the same data set. The relevant benchmark model is as follows:

TFN [6]: tensor fusion network, by learning the data representation of three modes respectively, carries out feature fusion through the form of Cartesian product, and finally realizes the task of multi-mode sentiment analysis through classifier.

LMF [19]: Low-rank fusion network, which is an improvement of tensor fusion network, reduces the computation amount of Cartesian product of tensor fusion network through low-dimensional vector representation.

MFN [20]: Memory fusion network, which uses LSTM to learn single-mode characteristic information and then uses the memory attention mechanism to realize the interaction between modes to achieve multi-mode emotion analysis.

MISA [14]: multi-modal data is mapped to two subspaces, and then feature vectors of different subspaces are fused.

REVEN [21]: An attention-based model adjusts word embedding in textual data through non-textual data.

MULT [8]: Two-way transformer is used to realize the representation of one mode corresponding to another mode, and then the feature fusion process is carried out.

Mag-bert [22]: A multi-modal adaptation gate (MAG) is designed to optimize the fusion process of textual and non-textual data.

SELF-MM [15]: Multi-modal sentiment analysis task is realized by using multi-task learning and single-modal label automatic generation.

3.3. Hyperparameter and evaluation index

This article uses the Adam as optimizer, Bert vector of the initial $5e^{-5}$, dropout value of 0.1, two-way LSTM dropout value is 0, vector other parameters as shown in table 2, batch_size (batch_size=32) was used for both MOSI and MOSEI data sets.

Table 2. Relevant learning rates of MOSI and MOSEI datasets

video data learning rate	audio data learning rate	Learning rate of other parameters	video data learning rate
MOSI	$0.6e^{-4}$	$0.6e^{-3}$	$0.6e^{-3}$
MOSEI	$0.6e^{-4}$	0.005	$0.6e^{-3}$

According to previous work [14,22], this paper evaluates the performance of the proposed model in two ways: classification and regression. For classification problems, accuracy (seven classification tasks and two classification tasks) and F1-score were used to evaluate the model performance. Specifically, for MOSI and MOSEI datasets, binary accuracy and F1 scores are calculated in two ways: negative/non-negative (without excluding zero) (Zadeh et al. [6]) and negative/positive (excluding zero) (Tsai et al. [14]). The average absolute error and correlation coefficient are used as evaluation indexes for regression tasks.

3.4. Experimental result

The specific experimental results are shown in Table 3 and Table 4, where the data in the table comes from the experimental data of Wei et al. [23], and the MODEL is the experimental results of the model in this paper. For binary classification accuracy and F1 score, the left side is the negative/non-negative experimental results, and the right side is the negative/positive experimental results (excluding zero). The overall experimental results are based on the percentage system.

The training loss curves of MOSI data set and MOSEI data set are shown in Figure. 2 and Figure. 3.

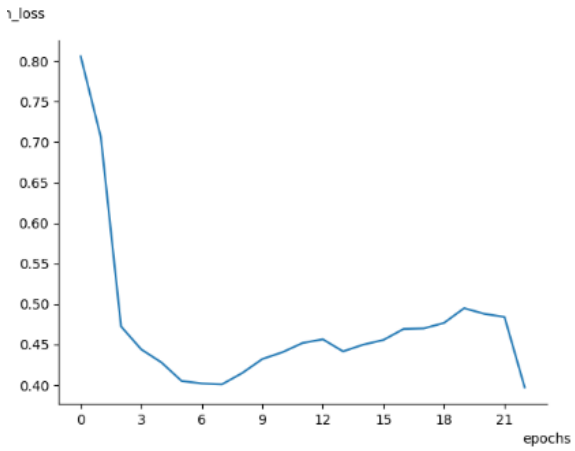

Figure 2. Loss of MOSI training

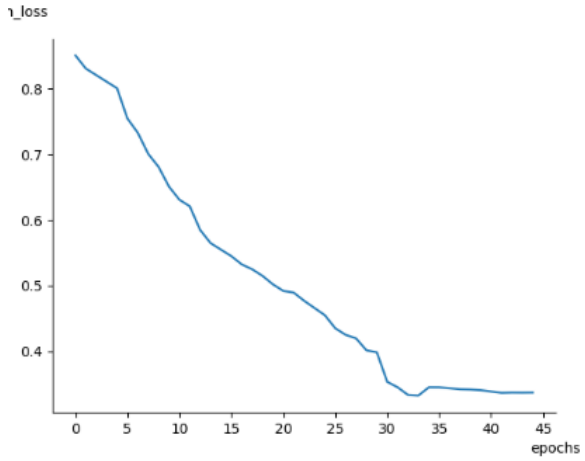

Figure 3. Loss of MOSEI training

Table 3. Experimental results of MOSI dataset

Model	MOSI				
	MAE	**Corr**	**Acc7**	**Acc2**	**F1**
TFN	90.10	96.80	34.90	-/80.80	-/80.70
LMF	91.70	69.50	33.20	-/82.50	-/82.40
MFM	87.70	70.60	35.40	-/81.70	-/81.60
RAVEN	86.20	71.40	39.00	-/83.00	-/83.00
MULT	86.10	71.10	-	81.50/84.10	80.60/83.90
MISA	80.40	76.40	-	80.79/82.10	80.77/82.03
MAGBERT	72.70	78.10	43.62	82.37/84.43	82.50/84.61
SELF-MM	71.20	79.50	45.79	84.14/84.77	82.68/84.91
MODEL	72.10	79.60	45.33	83.21/85.18	83.11/85.15

Table 4. Experimental results of MOESI dataset

Model	MOSEI				
	MAE	**Corr**	**Acc7**	**Acc2**	**F1**
TFN	59.30	70.00	50.20	-/82.50	-/82.10
LMF	62.30	67.70	48.00	-/82.00	-/82.10
MFM	56.80	71.70	51.30	-/84.40	-/84.30
RAVEN	56.50	71.30	-/81.70	-/84.20	-/84.20
MULT	58.00	70.30	-	-/82.50	-/82.30
MISA	56.80	72.40	-	82.59/84.23	82.53/85.30
MAGBERT	54.30	75.50	52.67	82.51/84.82	82.77/84.71
SELF-MM	52.90	76.70	53.46	82.68/84.96	82.95/84.93
MODEL	52.40	77.20	54.31	82.18/85.30	83.01/85.24

3.5. Analysis of experimental results

According to the data results in Table 3 and Table 4, for classification tasks, compared with the earlier TFN, LFN and other reference models, the model proposed in this paper has a 3%-5% improvement in accuracy and F1 score, which is still better than the reference model that also uses BERT as the feature extraction of text data. For regression tasks, the model proposed in this paper is better than the benchmark model in terms of mean absolute error MAE and correlation Corr. In terms of the two common multimodal sentiment analysis data sets, as the amount of data in MOSEI data set is much larger than that in MOSI data set, experimental results show that the model proposed in this paper has improved in correlation, accuracy and F1 score compared with before in MOSI data set, while its effect in average absolute error is not so good. It indicates that a small amount of data will affect the overall fitting effect of the data and lead to larger overall error. However, for the MOSEI data set, the data set has a relatively large amount of data, so the improvement effect is obvious in all aspects. To sum up, the experimental improvement effect of the model proposed in this paper on the MOSEI data set is better than that on the MOSI data. It can be shown that data is one of the important factors affecting the model effect.

3.6. Ablation experiment

According to the experimental results, it can be found that the performance of the model is improved by introducing intermodal self-attention mechanism, single-mode self-attention mechanism and loss weight allocation. The following ablation experiments were conducted for these three aspects, and the experimental results are shown in Table 5 and Table 6. Note: bi represents the intermodal attention

mechanism, uni represents the unimodal attention mechanism, and weight represents the weight loss mechanism. Other data descriptions are the same as in Table 3. The whole experimental results are based on the percentage system.

Table 5. Ablation study results of MOSI dataset

Method	MOSI				
	MAE	Corr	Acc7	Acc2	F1
bi+uni	72.70	76.40	45.54	82.54/84.36	82.54/84.39
bi+weight	72.40	79.40	44.92	82.32/84.42	82.59/84.44
uni+weight	73.60	78.80	44.75	82.68/84.39	82.65/84.41
bi	72.70	76.40	45.28	82.62/84.27	82.60/84.30
uni	72.30	79.10	46.30	82.98/84.85	82.9/84.84
weight	73.60	79.00	44.05	82.80/84.61	82.75/84.61
MODEL	72.10	79.60	45.33	83.21/85.18	83.11/85.15

Table 6. Ablation study results of MOSI dataset

Method	MOSI				
	MAE	Corr	Acc7	Acc2	F1
bi+uni	53.50	76.00	53.35	81.27/84.52	81.67/84.44
bi+weight	52.60	77.10	53.84	77.92/84.16	78.81/84.29
uni+weight	53.10	76.40	53.91	81.95/85.00	82.34/84.92
bi	53.20	76.40	53.18	81.76/84.65	82.08/84.51
uni	53.60	76.00	53.41	81.95/84.63	82.22/84.45
weight	53.40	76.20	53.40	77.72/83.14	78.52/83.24
MODEL	52.40	77.20	54.31	82.18/85.30	83.01/85.24

3.7. Analysis of ablation results

According to the experimental results of ablation implementation in Table 5 and Table 6, for classification tasks, using only the attention mechanism on a relatively small MOSI dataset can better improve the accuracy of classification. In addition, the mean absolute error MAE and correlation values in the regression task are only 0.2%~0.7% difference. For MOSEI dataset with large data set size, the effect of the combined use of attention mechanism and weight loss mechanism was significantly improved. For classification tasks, the accuracy of seven classification tasks, the accuracy of two classification tasks and F1 score were increased by 1%-4%. In regression tasks, attention mechanism and weight loss mechanism were used together. Mean absolute error MAE and correlation are also better than the two alone. As a whole, the attention mechanism and weight loss mechanism proposed in this paper show obvious performance on MOSEI with large data scale. It can be seen that when the attention mechanism is removed, the overall classification effect decreases significantly in the accuracy of binary classification, reaching 1.2%~4.4%. It fully demonstrates the importance of attention mechanism in natural language tasks. It can effectively help the model understand semantic and contextual information.

4. Conclusion

This paper introduces a multi-task and multi-mode emotion analysis model based on attention mechanism, which can be divided into two types: intermodal attention mechanism and single-mode self-attention mechanism. The intermodal attention mechanism can fully consider the intermodal relationship and complement each other. Compared with the previous work, more feature information can be learned. For the single-mode self-attention mechanism, the context information of single-mode can be fully learned, and the information gain between modes can be fully utilized to reduce noise through the combination of multi-task learning.

However, the model in this paper still needs to be improved. The model proposed in this paper can only deal with multi-modal emotion task segmentation, and has no adaptive ability to single mode and double mode emotion analysis task, and the multi-task weight loss is a part without parameter learning. It needs to be improved in the future.

Acknowledgment

This paper was supported by "Seedling Cultivation" Project for Young Scientific and Technological Talents of Education Department of Liaoning Province, No. J2020113.

References

[1] Yang Li-gong, Zhu Jian, TANG Shi-ping. A review of text sentiment Analysis [J]. Journal of Computer Applications, 2013, 33(6):1574-1607.

[2] Liu Jiming, Zhang Peixiang, Liu Ying, et al. Multimodal sentiment Analysis [J]. Journal of Computer Science and Exploration, 2021, 15(7): 1165.

[3] Poria S, Chaturvedi I, Cambria E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016: 439-448.

[4] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. 2017.Benchmarking multimodal sentiment analysis. In International Conference on Computational Linguistics and Intelligent Text Processing, pages 166–179.Springer.

[5] Zadeh A , Zellers R , Pincus E , et al. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages[J]. IEEE Intelligent Systems, 2016, 31(6):82-88.

[6] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis.arXiv preprint arXiv: 1707. 07250.

[7] Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]//Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). 2017: 873-883.

[8] Zadeh A, Liang P P, Poria S, et al. Multi-attention recurrent network for human communication comprehension[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[9] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.

[10] Tsai, Y .-H. H.; Bai, S.; Liang, P . P .; Kolter, J. Z.; Morency,L.-P .; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for Computational Linguistics. Meeting, volume 2019, 6558. NIH Public Access.

[11] Yu Z , Qiang Y . A Survey on Multi-Task Learning[J]. 2017.

[12] Davoodi E , Kosseim L , Mongrain M . On the Influence of Contextual Features for the Identification of Complex Words[J]. International Journal of Semantic Computing, 2017, 11(04):497-511. V, et al. Channel models for fixed wireless applications. IEEE 802.16a cont. IEEE 802.16.3c-01/29r4, 2003

[13] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. 2018.

[14] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, pages 1122–1131.

[15] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu.2021. Learning modality-specific representa-tions with self-supervised multi-task learning for multimodal sentiment analysis. arXiv preprintarXiv: 2102.04830.

[16] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[17] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 31(6):82–88.

[18] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria,Erik Cambria, and Louis-Philippe Morency. 2018.Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246.

[19] Liu, Z.; Shen, Y .; Lakshminarasimhan, V . B.; Liang, P . P .;Zadeh, A. B.; and Morency, L.-P . 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (V olume 1: Long Papers), 2247–2256.

[20] Zadeh, A.; Liang, P . P .; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P . 2018a. Memory fusion network for multi-view sequential learning. arXiv preprintarXiv:1802.00927 .

[21] Wang, Y .; Shen, Y .; Liu, Z.; Liang, P . P .; Zadeh, A.; and Morency, L.-P . 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 7216–7223.

[22] Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A. B.; Mao, C.;Morency, L.-P .; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2359–2369.

[23] Han W , Chen H , Poria S . Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis[J]. 2021. arXiv preprintarXiv: 2102. 04830.2109.0041.