Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.042150

ARTICLE

Multimodal Sentiment Analysis Based on a Cross-Modal Multihead Attention Mechanism

Lujuan Deng, Boyi Liu* and Zuhe Li

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

*Corresponding Author: Boyi Liu. Email: 332107040627@email.zzuli.edu.cn

Received: 19 May 2023 Accepted: 29 November 2023 Published: 30 January 2024

ABSTRACT

Multimodal sentiment analysis aims to understand people's emotions and opinions from diverse data. Concatenating or multiplying various modalities is a traditional multi-modal sentiment analysis fusion method. This fusion method does not utilize the correlation information between modalities. To solve this problem, this paper proposes a model based on a multi-head attention mechanism. First, after preprocessing the original data. Then, the feature representation is converted into a sequence of word vectors and positional encoding is introduced to better understand the semantic and sequential information in the input sequence. Next, the input coding sequence is fed into the transformer model for further processing and learning. At the transformer layer, a cross-modal attention consisting of a pair of multi-head attention modules is employed to reflect the correlation between modalities. Finally, the processed results are input into the feedforward neural network to obtain the emotional output through the classification layer. Through the above processing flow, the model can capture semantic information and contextual relationships and achieve good results in various natural language processing tasks. Our model was tested on the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) and Multimodal EmotionLines Dataset (MELD), achieving an accuracy of 82.04% and F1 parameters reached 80.59% on the former dataset.

KEYWORDS

Emotion analysis; deep learning; cross-modal attention mechanism

1 Introduction

Sentiment analysis means the use of computer automatic analysis to determine the emotions that people wish to express [1]. Much of the early sentiment analysis studies focused on text data. Text sentiment analysis focuses on the analysis, mining and reasoning of emotions embedded in a text. However, with the numerous changes occurring in various forms of social media, the expressions people present on the platform range from written expressions to rich and detailed graphics and videos expressing various emotions and opinions, single text sentiment analysis cannot adapt to the new multi-modal environment.

Although multimodal data can provide richer feature information, multimodal heterogeneity creates new problems and challenges for multimodal sentiment analysis tasks [2]. First, different

modalities must use corresponding methods to process raw data information and extract their respective feature representations. Second, both the shared information between modalities and the unique information within a modality are important sources of information in multimodal emotion analysis tasks. Making good use of the shared information between modalities becomes the key to performing multimodal sentiment analysis tasks [3].

In previous multimodal fusion research applications [4], three fusion methods were used: early fusion, late fusion, hybrid fusion which combines both of them. Early fusion models can obtain the correlation of different modalities between multimodal features, can then better complete the required sentiment analysis. In the early stage, the features of different modalities were fused after the features were extracted separately, only one classifier needs to be learned. However, the feature information of different levels are not completely independent of each other; therefore, early fusion may result in some low-level feature information being lost, which may cause some important information to be ignored during the fusion process. In the late fusion model, different modal features are independently extracted and classified in different ways, after the local decision results are obtained, the fusion of local decision results is performed to obtain a final decision. However, hybrid fusion is a combination of both methods and aims to exploit the advantages of both while overcoming the disadvantages of both methods.

However, most tasks using the abovementioned fusion methods perform simple splicing or mathematical operations on the relevant modes [5]. That is, a simple linear transformation of the text modality is converted into the feature vector size required by the image modality, concatenating text feature vectors with image feature vectors. This method neither fully utilizes the text information nor effectively realizes the text image [6]. From the research, it can be observed that simple feature splicing or mathematical models do not make the connection between different modalities more apparent [7]. However, when we use the cross-modal attention mechanism, we do not need to splice the feature vectors of different modalities. Our model incorporates a multihead attention mechanism specially designed for cross-modal fusion, which can more effectively integrate different modalities in sentiment analysis. This attention mechanism allows the model to capture the interdependencies and relationships between modalities, resulting in improved performance compared to traditional fusion methods. At the transformer layer, a cross-modal attention consisting of a pair of multi-head attention modules are employed to reflect the correlation between modalities. Finally, the processed results are input into the feedforward neural network to obtain the emotional output through the classification layer. The experiments show that the accuracy rate reaches 82.04%, which determined that our multimodal fusion method is effective. Our contributions are as follows:

- We propose a deep interactive framework for multi-modal sentiment analysis. This framework can obtain a richer comprehension of multi-source data by capturing intricate relationships and dependencies across modalities.
- We introduce a multi-modal fusion module with text modality as the main driver, which effectively integrates information from different modalities.
- We propose a new cross-modal multi-head attention mechanism. The role of cross-modal attention mechanism is to effectively capture and integrate correlated information between different modalities in multi-modal data analysis.

The present article is structured in the following manner: Section 2 describes the articles that also use the CMU-MOSEI dataset, Section 3 explains our model, Section 4 describes the dataset we used and extracted the feature vectors for each modality, Section 5 describes the results obtained for various datasets, Section 6 summarizes and elaborates upon future improvements that may be performed.

2 Related Works

As an important branch of natural language processing sentiment analysis, the multimodal emotion analysis task has been considerably developed in recent years. The task of multimodal sentiment analysis can be divided into two categories: interactive and narrative multimodal emotion analyses [8].

Interactive multimodal emotion analysis refers to the analysis of human verbal- and non-verbal-interaction behaviors to infer the emotional state of the participants. Unlike traditional single-modal emotion analysis, it is a requirement for interactive multimodal emotion analysis to consider multimodal information, including voice, facial expression, body language, text, tone and other different forms of information. In interactive scenarios, these different sets of information interact with each other, an individual's emotional state may be expressed through multiple modalities and the transmission and interaction of emotions may also be conducted between different modalities.

The task of interactive multimodal sentiment analysis typically involve three aspects: emotion recognition, emotional expression and emotional interaction. Emotion recognition usually involves the analysis of data in different modalities such as text, speech, images, videos, etc., to identify and classify emotions [9]. Emotional expression refers to expressing one's emotional state through various forms, such as whether a person is expressing certain emotions, such as anger or happiness and what their method of expression is [10]. Emotional interaction aims to make computer systems more humane and emotionally intelligent by identifying, generating, and simulating emotions to provide a richer and more intimate human-computer interaction experience.

Extracting sentiment information from multiple modalities and combining them to attain a comprehensive analysis and inference is the fundamental requirement of narrative multimodal sentiment analysis to classify subjective attitudes into positive, negative and neutral categories. In addition to feature learning, multimodal narrative sentiment analysis must also be subjected to the process of information fusion, which integrates and analyzes the data obtained from different modalities (such as a text, image, or video). This process is called multimodal interaction or multimodal state fusion [11].

In the early research, the problem of multimodal fusion gained too much attention from the researchers and the key to solving multimodal fusion problems is to improve the fusion method of different modalities. During the ongoing process of improving fusion methods, some better models have emerged so far that are worth our attention.

The tensor fusion network (TFN) [12] is a newly built model that can learn intramodal and intermodal dynamics in an end-to-end manner, aggregating unimodal, bimodal and trimodal interactions. However, the model is a typical multimodal network performing feature fusion activity through mathematical matrix operations. The model obtains the correlation of each modal element by calculating the tensor outer product of different modes. However, it drastically increases the rank of the feature tensor, making the model too large to train.

In Shankar et al.'s study [13], a new form of multimodal fusion architecture, profusion was proposed, which improves the shortcomings of TFN networks by connecting late-stage fused representations to single-modal feature generators via reverse links.

The recurrent neural network (RNN) [14] was first used to analyze the non-text censorship features obtained from the relevant dataset and then used RNN with long short-term memory to analyze the comment sentiment after removing any irrelevant features. However, due to the vanishing-gradient problem of the RNN, the RNN can only remember short-term information and finds it difficult to address long-term dependence issues.

In their study [15], this model combines lexical information with an attention LSTM model and utilizes deep neural networks to create a more stable framework without the need for additional labeled data.

Delbrouck et al. [16] used the transformer model in their study to solve multimodal sentiment analysis tasks and proposed a transformer-based joint-encoding mechanism (TNJE). Using the attention mechanism to perform modality fusion, better results were obtained.

Moreover, Zadeh et al. used mathematical models to splice or multiply the features of different modalities for modality fusion, while other experiments used deep learning models to fuse different modalities. This paper presents a fusion mechanism based on cross-modal attention to achieve better emotion analysis; by utilizing the three key values of Q (query), K (key) and V (value) in the attention mechanism, K and V of the audio mode are applied to the text mode and the attention mechanism is then used to fuse the differences between the different modes with each other. Ultimately, its effectiveness is demonstrated through relevant experiments.

3 Method

As a variant of the traditional transformer model, the interactive transformer layer learns representations for different modalities through the use of an encoding framework. Unlike sequence encoding and recursive models, this model obtains the global relationship from start to finish through attention mapping and a feedforward neural network. In comparison to the sequential long short-term memory network, a transformer allows for parallel processing and can better handle the global information it acquires [17]. As shown in Fig. 1.

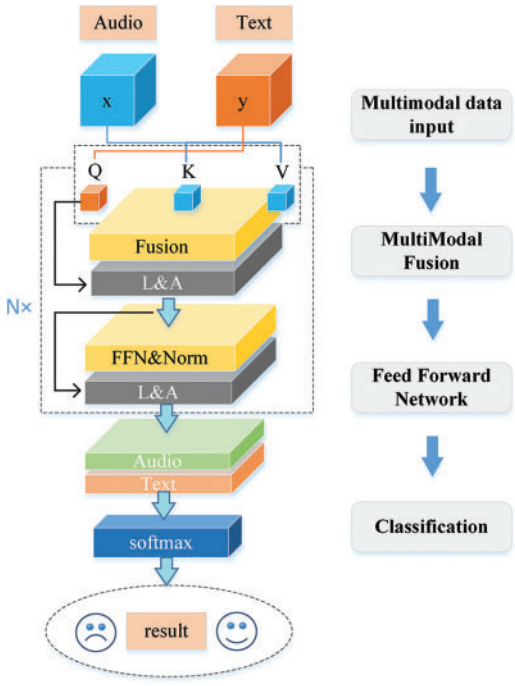

Figure 1: Overview of the proposed framework

3.1 Attention

A traditional transformer encoder comprises multiple same layers, where each layer includes two sublayers and a transformer encoder is formed by stacking these layers. Residual connections are evident around the two sublayers whose role is to add the input vector to the output of each sublayer to transfer information and avoid information loss. The layer-normalization process performs a normalization operation on the output of each sublayer in order to enhance the model's stability and reduce the occurrence of internal covariate shifts. The output expression of the sublayer is y = LayerNorm(x + Sublayer(x)), where the function of sublayer(x) is to realize the function of the sublayer itself [18]. In a traditional transformer [19], these two sublayers are the multi-head attention (MHA) and the feedforward neural network (FFN) [20,21]. As shown in Fig. 2.

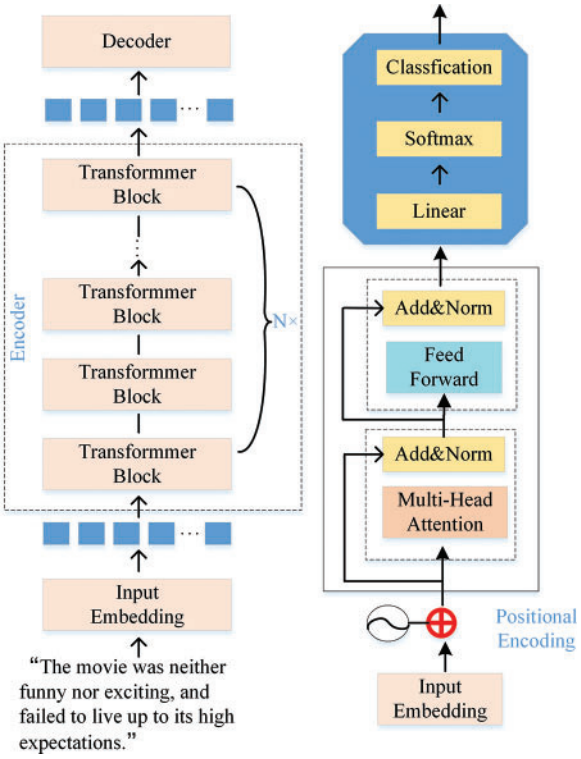

Figure 2: Transformer coding details and overall process

The self-attention mechanism [22,23]: when human beings observe objects, they cannot observe everything in front of them carefully and at the same time and can only focus on a certain section of the object. Usually, after understanding the scene in front of us, our brain can rapidly focus on the most valuable parts for careful observation to perform effective judgments [24]. Attention is therefore achieved through the following formula:

$$Attention(Q, K, V) = softmax\frac{QK^T}{\sqrt{d_k}}V \tag{1}$$

where QK^T is the attention square matrix, $\sqrt{d_k}$ is the scaling factor and V is the matrix.

If the input size is N × M, the operation generates an attention square matrix and $\sqrt{d_k}$ is a scaling factor. The attention calculation result of the current time step can be understood as the accumulation

of a group of coefficients multiplied by the feature vector value of each time step. These coefficients are obtained by the inner product of the query of the current time step and the key corresponding to other time steps. This process is equivalent to querying the keys of other time steps with your own query, judging the similarity and deciding on what proportion should inherit the information of the corresponding time step [25]. Multihead attention (MHA) superimposes multiple self-attention modules while paying attention to information available in different positions and subspaces. In other words, it transforms the vector calculation form of self-attention into a matrix calculation form, where multiple word vectors are placed into a matrix to calculate the output of all-time steps simultaneously.

$$MHA(Q, K, V) = \sum_{i=1}^{h} head_i W_o \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

MHA represents the multihead attention mechanism and W_o is a matrix form of multiple word vectors.

3.2 Cross-Modal Attention

In the previously performed multimodal emotion analysis tasks [26], the process of data integration and analysis from different modalities (such as a text, image, or video), that is, multimodal fusion, mainly uses three methods: early, late and hybrid, which fuses both early and late fusion methods. However, these fusion methods do not effectively perform the fusion between modalities. The model is able to capture the interdependencies and relationships between modalities through this attention mechanism, resulting in improved performance compared to traditional fusion methods. At the transformer layer, a cross-modal attention consisting of a pair of multi-head attention modules are employed to reflect the correlation between modalities. Finally, the processed results are fed into the feedforward neural network and the sentiment output is obtained through the classification layer. We combined the two modes by utilizing the key and value of the audio mode and the query of the text mode. We hypothetically designed two modes, α and β, where the sequences of each mode were denoted as X_α and X_β. Inspired by language translation, we believed that modality fusion performed by using cross-modal attention mechanism was a good way to achieve the desired results, i.e., β - α:

$$Y_\alpha = CM_{\beta\to\alpha}\left(X_\alpha, X_\beta\right) \tag{4}$$

$$Y_\alpha = softmax\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}} V_\beta\right) \tag{5}$$

$$Y_\alpha = softmax\left(\frac{X_\alpha W_{Q\alpha} W_{K\beta}^T X_\beta^T}{\sqrt{d_k}}\right) X_\beta W_{V\beta} \tag{6}$$

The cross-modal attention mechanism modulates the self-attention of the modality through the main modality x and adds residual links to form a new attention sublayer:

$$y = LayerNorm\,(y + MHA(x, y, y)) \tag{7}$$

where y is the current input and MHA(x,y,y) is the output after using the multihead attention mechanism.

To ensure that the operation and residual connection functioned properly following the addition of QK^T, the feature sizes of x and y need to be consistent. The adjustment of the two features to the same size can be solved using the MHA module.

Finally, the end of the modality block consisted of a SoftMax layer; the weighted sum output by the FFN layer was used as the input for the SoftMax layer, projecting the modality into a new space. The SoftMax layer consists of G identical blocks and their output values are stacked.

$$h = W_m M \tag{8}$$

$$a_i = softmax(v_i^{aT} h) \tag{9}$$

$$Z = \sum_{j=0}^{N} a_{ij} M_j \tag{10}$$

where W_m is the transformation matrix with a size of 2k × k embedding matrix M at a higher dimension, the size of V_i^{aT} is 1 × 2k, M is the output of the FFN layer and Z is a vector of size k, which consists of G identical blocks and the stacked output. This can be expressed as

$$G_m = Stacking(Z_1, \ldots, Z_m) \tag{11}$$

3.3 Classification Layer

A total of 8 hidden transformer blocks of size 512 are typically used during experiments; the self-attention mechanism had 4 multiheads. The decay rate we applied to the output of each block is set to 0.1, the coefficient of weight decay used for the input value of the classification layer is set to 0.5.

We truncated the relevant features with a spatial dimension of 80 as a treatment for the acoustic features. Additionally, we set the SoftMax value to 80.

After the calculation of all the model components, the modality enters the SoftMax layer of size 1 to obtain the vector value. The various forms of vectors are then added element by element.

$$y \sim p = W_a \left(LayerNorm\left(s_i\right)\right) i \in [L, A] \tag{12}$$

4 Data Preparation

4.1 Mosei Dataset

We validated our experimental results using this dataset [27]. The dataset contains over 2000 speaker opinion video clips, each video clip is essentially a monolog and contains three different modalities: spoken language communicated in a text format, language communicated visually through percept gestures and deadpan expressions and language communicated through voice. Each sentence is labeled with various tags. The Likert scale was used to categorize the emotions presented by the subjects. An additional scale, using Ekman's Atlas of Emotions (Happiness, Despair, Anger, Outrage, Revulsion, Surprise), was created based on a Likert scale of [0,3] for denoting Sentiment z (0: no proof for z, 1: weak z, 2: z, 3: strong z). Among them, the training set: contains more than 2000 video clips for model training and parameter adjustment. Validation set: contains approximately 250 video clips used to evaluate model performance and hyperparameter selection during training. Test set: contains about 1000 video clips for final evaluation of the performance of the model on sentiment analysis tasks. The label distribution is presented in the Fig. 3 below.

4.2 Meld Dataset

In this experiment, we also used the MELD dataset [28]. The data presented in this dataset were obtained from a multi-person dialog and were annotated with seven emotions, namely, anger, repulsion, fear, joy, neutral, sorrow and surprise, where anger, repulsion, fear and sorrow were perceived as negative emotions, joy was perceived as a positive emotion and the remaining emotions, unlike the

abovementioned emotions, were regarded as neutral, thereby converting fine-grained emotions into coarse-grained emotions. Among them, the training set: contains more than 1000 dialogues, which are used for model training and parameter adjustment. Validation set: contains about 110 conversations and is used to evaluate the performance of the model and perform hyperparameter selection during training. Test set: contains about 280 dialogues and is used to finally evaluate the performance of the model on sentiment analysis tasks. Fig. 4 presents an introduction to the MELD dataset.

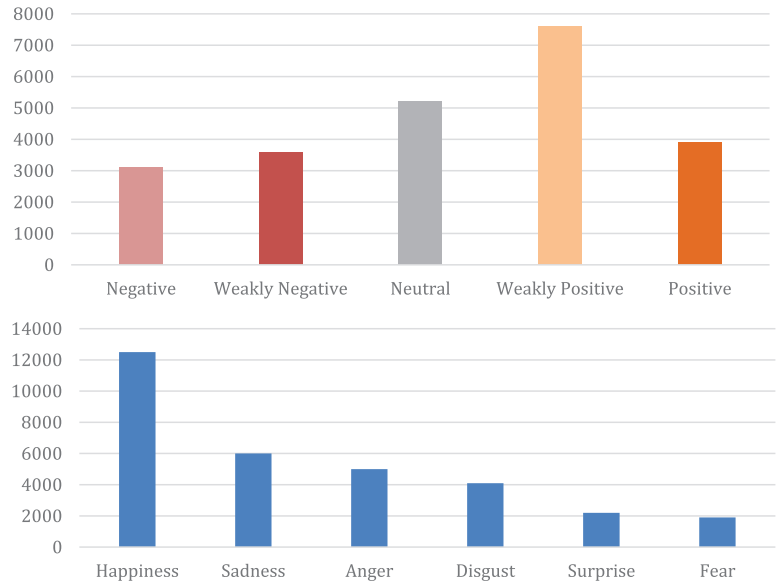

Figure 3: MOSEI statistics obtained from the author's paper [27]

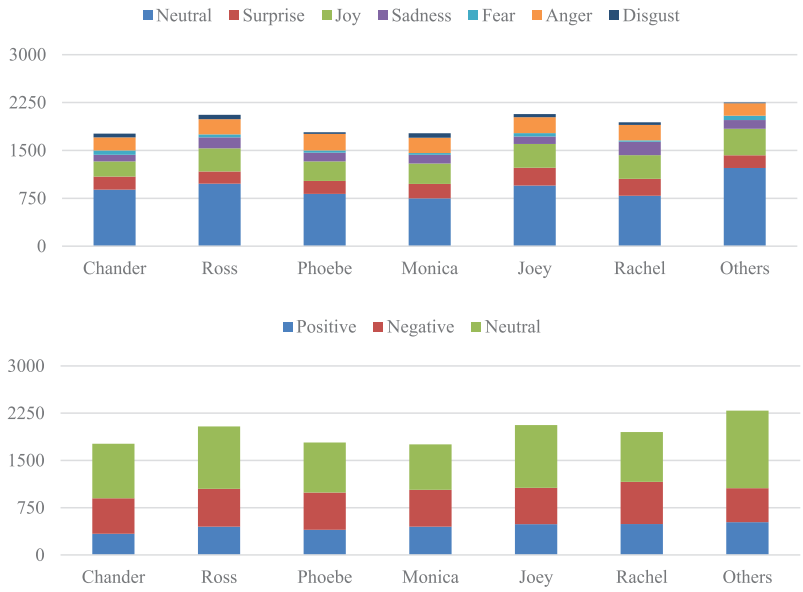

Figure 4: MELD statistics obtained from the author's paper [28]

4.3 Feature Extraction

The following text describes the feature extraction process used for the different modalities.

4.3.1 Linguistic Feature

First, all the relevant statements were converted into lowercase letters, removing special characters and any punctuation. A vocabulary database of 14,176 unique words was built and every word was embedded in a 300-dimensional tensor. Words in the validation or test set that do not appear in the vocabulary were labeled as unknown.

4.3.2 Acoustic Feature

The information transmitted through an audio means is often more complex, such as nonverbal expressions (laughter, gasping, sighing) and rhythmic characteristics (speech rate and intonation). The common audio-extraction features we considered were as follows: 1. Number of times the signal crosses points on the time axis (zero-crossing rate); 2. Features (melody) that must be obtained through a specific model; and 3. The characteristics of sound (MFCCs). When extracting relevant audio features, to ensure that the data were not affected by noise, we first removed any irrelevant sounds and then focused on the sound produced by the human voice [29].

First, N sampling points were assembled into one observation unit. The time covered was approximately 20∼30 ms. To avoid the presence of excessive changes in the two frames, an overlapping area, including H sampling points, was set, where H was approximately 1/2 to 1/4 of N. Noise was removed, we obtained a feature that presented 80 dimensions.

4.3.3 Visual Feature

The good performance of the convolutional neural network (CNN) caused us to use it to preprocess the relevant visual features. The CNN with a 3D convolution kernel was used to process video to extract features. However, due to the noise produced by the visual features when using them, our experiments did not obtain better results; therefore, we abandoned the use of visual features.

5 Performance Evaluation

In the fifth section, we present the experimental results we obtained in our study. The experimental results were obtained from datasets CMU-MOSEI and MELD.

5.1 Experimental Setting

This study used the Adam optimizer [30] to optimize the model. The learning rate was set to $1e^{-4}$ to produce a better performance and the batch size was set to 64 [31]. If the accuracy of the validation set in a prespecified batch did not increase, we preset 0.2 as the decay coefficient after updating each the learning rate and stopped training after obtaining three batches.

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t + \varepsilon}} m_t \tag{13}$$

5.2 The Results of the CMU-MOSEI Dataset

The CMU-MOSEI dataset was used to validate our designed model and compare its results with those obtained by other models. Therefore, the two-category sentiment results are presented in Table 1 and Fig. 5. Additionally, our experiments achieved the best results after we performed a comparison.

The caveat we need to mention is that we used standard accuracy instead of weighted accuracy variables. Therefore, we only performed comparisons with a model that used the same method as ours. It should also be noted that the model using only language + acoustic was our optimal model and adding visual information did not improve the result. How to improve the use of visual information has become the goal of our continued efforts. As shown in Tables 2 and 3.

Table 1: Two classification results based on the CMU-MOSEI dataset

Model	Accuracy (%)	F1-score (%)
Graph-MFN (T + A + V) [32]	76.90	77.00
B2 + B4 w/multimodal fusion (T + A + V) [33]	81.14	78.53
Multilogue-Net (T + A) [34]	80.18	79.88
TBJE (T + A + V) [16]	81.50	/
HFFN [35]	80.20	80.30
Our approach (T + A)	82.04	80.59

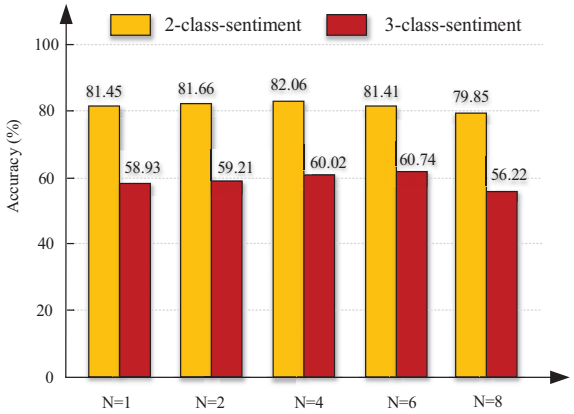

Figure 5: Accuracy results based on the number of transformer blocks

Table 2: Seven classification results based on the CMU-MOSEI dataset

Model	Accuracy (%)	F1-score (%)
Graph-MFN (T + A + V) [32]	45.00	/
TBJE (T + A + V) [16]	44.40	/
Our approach (T + A)	45.20	41.36

Table 3: Six-class emotions of the CMU-MOSEI dataset

Model	Accuracy (%)	F1-score (%)
Graph-MFN (T + A + V) [32]	/	/
TBJE (T + A + V) [16]	80.68	/

(Continued)

Table 3 (continued)		
Model	Accuracy (%)	F1-score (%)
Our approach (T + A)	81.03	80.07

5.3 The Results of the MELD Dataset

We continued to evaluate the model on the MELD dataset. However, the effect we achieved was not very significant. The reason for this was that the compositions of the MELD and CMU-MOSEI datasets were different. The MELD dataset focused on multiperson dialog scenes. Taking acoustics as an example, in multiperson dialogue scenes, there is often accompanied by noise and personal emotions are affected by others. This factor affected our model. The results of the MELD dataset are presented in Table 4 and Fig. 6.

Table 4: The results of the MELD dataset

Model	3-Class-sentiment accuracy (%)	7-Class-sentiment accuracy (%)
Text-CNN (T) [36]	64.25	55.02
BcLSTM (T+A) [37]	66.68	59.25
Our approach (T + A)	60.74	55.13

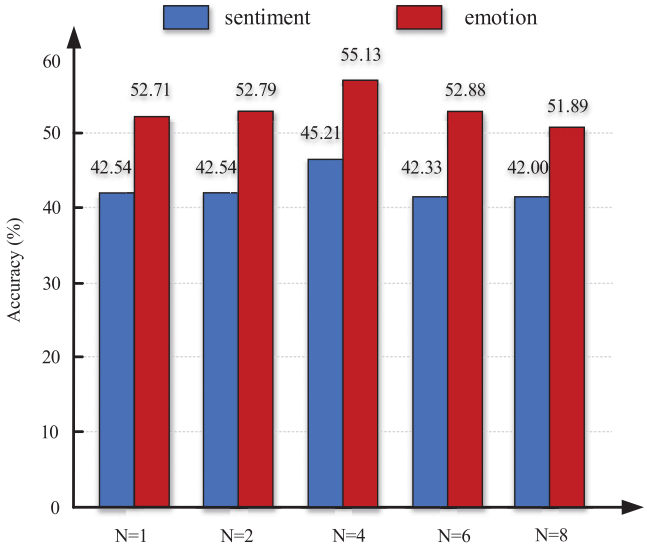

Figure 6: Seven-class sentiment and emotion accuracy values according to the number of blocks per interactive transformer

6 Conclusions

In this study, we propose a cross-modal multi-head attention mechanism by utilizing the key values of Q, K and V in the attention mechanism. Specifically, we apply the K and V from the audio

modality to the text modality, leveraging the attention mechanism to integrate different modalities for improved accuracy in multi-modal sentiment analysis. The proposed cross-modal attention mechanism effectively utilizes information from different modalities, outperforming traditional approaches such as modality concatenation or modality-wise inner product. In our experiments on the CMU-MOSEI and MELD datasets, we did not use visual information. This is because the visual modality includes not only the visual features of the characters but also a significant amount of background information, which introduces irrelevant noise unrelated to the emotional states of the characters and degrades the performance of the model when visual modality information is incorporated. Therefore, we excluded any visual information. However, utilizing visual features and integrating visual information into the experiments remains a potential direction for future research. Next, we plan to focus more on the facial expression region of individuals and remove irrelevant information to reduce noise when extracting visual features, leading to further experiments and investigations.

Acknowledgement: This paper was supported by the National Natural Science Foundation of China, the Henan Provincial Science and Technology Research Project, the Research and Practice Project of Higher Education Teaching Reform in Henan Province, the Research and Practice Project of Higher Education Teaching Reform in Henan Province, the Academic Degrees & Graduate Education Reform Project of Henan Province.

Funding Statement: This paper was supported by the National Natural Science Foundation of China under Grant 61702462, the Henan Provincial Science and Technology Research Project under Grants 222102210010 and 222102210064, the Research and Practice Project of Higher Education Teaching Reform in Henan Province under Grants 2019SJGLX320 and 2019SJGLX020, the Undergraduate Universities Smart Teaching Special Research Project of Henan Province under Grant Jiao Gao [2021] No. 489-29, the Academic Degrees & Graduate Education Reform Project of Henan Province under Grant 2021SJGLX115Y.

Author Contributions: Conceptualization, L.D.; Data curation, L.D.; Formal analysis, L.D. and Z.L.; Investigation, Z.L.; Methodology, L.D. and B.L.; Software, Z.L.; Supervision, Z.L.; Writing–review & editing, B.L.

Availability of Data and Materials: The data presented in this study are openly available in (CMU-MOSEI) at (10.18653/v1/P18-1208) and (MELD) at (1810.02508v6) reference numbers [28] and [29].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Poria, E. Cambria, R. Bajpai and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[2] I. Kansizoglou, L. Bampis and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 756–768, 2019.

[3] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L. P. Morency *et al.,* "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. of ICMI*, Montréal, QC, Canada, pp. 6–15, 2021.

[4] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2022.

[5] L. Zhu, Z. Zhu, C. Zhang, Y. Xu and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion*, vol. 95, pp. 306–325, 2023.
[6] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *Association for Computing Machinery*, vol. 19, no. 1, pp. 1–19, 2023.
[7] Z. Yu, J. Yu, Y. Cui, D. Tao and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. of IEEE/CVF*, Long Beach, CA, USA, pp. 6281–6290, 2019.
[8] R. Das and S. D. Thoudam, "Multimodal sentiment analysis: A survey of methods, trends and challenges," *ACM Computing Surveys*, vol. 55, no. 270, pp. 1–38, 2023.
[9] J. Zhang, Z. Yin, P. Chen and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.
[10] I. Kansizoglou, E. Misirlis, K. Tsintotas and A. Gasteratos, "Continuous emotion recognition for long-term behavior modeling through recurrent neural networks," *Technologies*, vol. 10, no. 3, pp. 59, 2022.
[11] Y. Cimtay, E. Ekmekcioglu and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.
[12] A. Zadeh, M. Chen, S. Poria, E. Cambria and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," arXiv:1707.07250, 2017.
[13] S. Shankar, L. Thompson and M. Fiterau, "Progressive fusion for multimodal integration," arXiv: 2209.00302, 2022.
[14] A. F. Agarap, "Statistical analysis on E-commerce reviews, with sentiment classification using bidirectional recurrent neural network (RNN)," arXiv:1805.03687, 2018.
[15] L. Bao, P. Lambert and T. Badia, "Attention and lexicon regularized LSTM for aspect-based sentiment analysis," in *Proc. of ACL*, Florence, Italy, pp. 253–259, 2019.
[16] J. B. Delbrouck, N. Tits, M. Brousmiche and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," arXiv:2020.15955, 2006.
[17] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency *et al.,* "Multimodal transformer for unaligned multimodal language sequences," in *Proc. of ACL*, Florence, Italy, vol. 2019, pp. 6558–6569, 2019.
[18] S. Mai, Y. Zeng, S. Zheng and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, pp. 2276–2289, 2022.
[19] F. Wang, S. Tian, L. Yu, J. Liu, J. Wang *et al.,* "TEDT: Transformer-based encoding-decoding translation network for multimodal sentiment analysis," *Cognitive Computation*, vol. 15, pp. 289–303, 2023.
[20] L. Yang, J. C. Na and J. Yu, "Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis," *Information Processing & Management*, vol. 59, no. 5, pp. 103038, 2022.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.,* "Attention is all you need," in *Advances in Neural Information Processing Systems*. Long Beach, CA, USA, vol. 30, 2017.
[22] C. Xi, G. Lu and J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proc. of ICMLSC*, Haiphong City, Viet Nam, pp. 34–39, 2020.
[23] K. Kim and P. Sanghyun, "Aobert: All-modalities-in-one BERT for multimodal sentiment analysis," *Information Fusion*, vol. 92, pp. 37–45, 2023.
[24] W. Dai, S. Cahyawijaya, Z. Liu and P. Fung, "Multimodal end-to-end sparse model for emotion recognition," arXiv:2103.09666, 2021.
[25] Q. Chen, G. Huang and Y. Wang, "The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2689–2695, 2022.
[26] Y. Du, Y. Liu, Z. Peng and X. Jin, "Gated attention fusion network for multimodal sentiment classification," *Knowledge-Based Systems*, vol. 240, pp. 108107, 2022.
[27] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria and L. P. Morency, "Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph," in *Proc. of ACL*, Melbourne, Australia, pp. 2236–2246, 2018.
[28] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria *et al.,* "MELD: A multimodal multi-party dataset for emotion recognition in conversations," arXiv:1810.02508, 2018.

[29] Q. Qi, L. Lin, R. Zhang and C. Xue, "MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis," *IEEE Access*, vol. 10, pp. 28750–28759, 2022.
[30] A. Joshi, A. Bhat, A. Jain, A. V. Singh and A. Modi, "COGMEN: COntextualized GNN based multimodal emotion recognition," arXiv:2205.02455, 2022.
[31] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Transactions*, vol. 19, no. 1, pp. 1–19, 2023.
[32] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria *et al.,* "Memory fusion network for multi-view sequential learning," in *Proc. of AAAI*, Washington DC, USA, no. 691, pp. 5634–5641, 2018.
[33] A. Kumar and J. Vepa, "Gated mechanism for attention based multi-modal sentiment analysis," in *Proc. of ICASSP*, Barcelona, Spain, pp. 4477–4481, 2020.
[34] F. Lin, S. Liu, C. Zhang, J. Fan and Z. Wu, "StyleBERT: Text-audio sentiment analysis with Bi-directional style enhancement," *Information Systems*, vol. 114, pp. 102147, 2023.
[35] X. Li and M. Chen, "Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language," in *Proc. of CCL*, Hainan, China, vol. 19, pp. 359–373, 2020.
[36] K. Yoon, "Convolutional neural networks for sentence classification," in *Proc. of EMNLP*, Doha, Qatar, pp. 1746–1751, 2014.
[37] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh *et al.,* "Context-dependent sentiment analysis in user-generated videos," *Proc. of ACL*, vol. 1, pp. 873–883, 2017.