electronics

Article

Context-Dependent Multimodal Sentiment Analysis Based on a Complex Attention Mechanism

Lujuan Deng [1], Boyi Liu [1,*], Zuhe Li [1], Jiangtao Ma [1] and Hanbing Li [2]

[1] School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; denglujuan@zzuli.edu.cn (L.D.); zuheli@zzuli.edu.cn (Z.L.); majiangtao@zzuli.edu.cn (J.M.)
[2] Songshan Laboratory, Zhengzhou 450000, China; anatoly-li@foxmail.com
* Correspondence: 332107040627@email.zzuli.edu.cn

Abstract: Multimodal sentiment analysis aims to understand people's attitudes and opinions from different data forms. Traditional modality fusion methods for multimodal sentiment analysis concatenate or multiply various modalities without fully utilizing context information and the correlation between modalities. To solve this problem, this article provides a new model based on a multimodal sentiment analysis framework based on a recurrent neural network with a complex attention mechanism. First, after the raw data is preprocessed, the numerical feature representation is obtained using feature extraction. Next, the numerical features are input into the recurrent neural network, and the output results are multimodally fused using a complex attention mechanism layer. The objective of the complex attention mechanism is to leverage enhanced non-linearity to more effectively capture the inter-modal correlations, thereby improving the performance of multimodal sentiment analysis. Finally, the processed results are fed into the classification layer and the sentiment output is obtained using the classification layer. This process can effectively capture the semantic information and contextual relationship of the input sequence and fuse different pieces of modal information. Our model was tested on the CMU-MOSEI datasets, achieving an accuracy of 82.04%.

Keywords: sentiment analysis; deep learning; complex attention mechanism

Citation: Deng, L.; Liu, B.; Li, Z.; Ma, J.; Li, H. Context-Dependent Multimodal Sentiment Analysis Based on a Complex Attention Mechanism. *Electronics* **2023**, *12*, 3516. https://doi.org/10.3390/electronics12163516

Academic Editor: Chiman Kwan

Received: 14 July 2023
Revised: 11 August 2023
Accepted: 18 August 2023
Published: 20 August 2023

1. Introduction

Sentiment analysis is a technique that uses computers to automatically analyze and infer human emotional expressions [1,2]. Initially, sentiment analysis work mainly focused on text data [3]. Text sentiment analysis aims to analyze, mine, and reason about the underlying sentiment in text. However, with the rapid development of social networks, the ways for people to express their emotions on the platform have become more and more abundant [4]. Extending from text expression to multimodal expression including pictures and videos, a single-text sentiment analysis technology cannot adapt to the diverse expression environment in online social media. Multimodal data have more information than unimodal data. For example, to detect a sarcastic sentence, such as "Great job, you've managed to make yourself even more unlikable", if only the textual information of the first half of the sentence is considered, it is easy to classify it as a positive sentiment [5]. But if supplemented by some information from the visual modality, it is easy to detect the speaker's unpleasant gesture or expression and classify it as a negative emotion. Similarly, in some cases, audio features (laughter, sighs, and screams) are used as supplementary information to regular textual modality information to verify the accuracy of emotion prediction. However, efficiently integrating information from these different modalities is a crucial task.

Videos provide an excellent source for extracting multimodal information. In addition to a visual modality, speech and text information is also provided [6]. Furthermore, the speaker can utter multiple utterances in one video, and these utterances can have different

emotions. The emotional information of an utterance is often interdependent with other contextual utterances.

We propose a multimodal sentiment analysis model based on a complex attention mechanism recurrent neural network. We apply complex attention to neighboring utterances to help the network model learn in a better way. Traditional attention mechanisms usually use single-matrix multiplication to compute attention weights to represent the influence of different parts of the input on the output. However, in multimodal tasks, different input sources have different representations and may require different transformations and fusions. Therefore, complex attention mechanisms use multiple matrix multiplications to compute attention weights to account for interactions and correlations among multiple sources. Among them, complex attention and occupation are composed of a simple attention head and a complex attention head. In the simple attention head, the calculation method of the attention matrix is consistent with the original multi-head attention mechanism; in the complex attention head, both Q and K are cut into two parts, and each part is subjected to linear transformation and activation function processing to obtain different "keys" and "values". Then, we use a "query" vector to calculate the similarity with each "key" and pass the result into a softmax function for normalization. The attention weights from the simple and complex attention heads are combined, weighted, and summed with each "value" vector to obtain the final output vector. Via the complex attention mechanism, the context information is fully integrated to further improve the accuracy rate. Our experiment is also based on the CMU-MOSEI dataset.

The structure of this article is as follows: In Section 2, the current state of research on multimodal sentiment analysis and related work is described; in Section 3, we describe our model; and in Section 4, we describe the composition of the dataset we use and how to extract different modalities. The eigenvectors of the results obtained on the CMU-MOSEI dataset are described in Section 5, and the conclusions and directions for future improvement are summarized in Section 6.

2. Related Works

Sentiment analysis is a natural language processing technology designed to automatically identify and extract the emotion or emotional color contained in text, audio, images, and other data sources [7]. The significance of sentiment analysis is that it can help people better understand and analyze a large amount of content on social media, such as users' attitudes and opinions on a certain product, brand, political event, etc. It can also be applied to many fields such as advertising, market research, brand management, and public opinion monitoring [8].

The development of sentiment analysis can be traced back to the 1960s, when research was mainly based on sentiment lexicons and grammar rules. With the continuous development of machine learning and deep learning technology, the method of sentiment analysis gradually shifts from a rule-based method to a data-driven method [9]. In addition, with the expansion of sentiment analysis application scenarios and the enrichment of data sources, sentiment analysis has gradually expanded from a single modality to a multimodal approach.

The purpose of multimodal sentiment analysis is to obtain sentiment information from multiple data sources and comprehensively consider the relationship between different data sources to identify and understand sentiment more accurately. For example, text and audio data can provide different emotional information and thus can complement and corroborate each other [10], thereby improving the accuracy of sentiment analysis. The application fields of multimodal sentiment analysis include audio and video content analysis, social media analysis, human–computer interaction, intelligent customer service, and other fields. The value of studying multimodal sentiment analysis lies in improving the accuracy and practicality of sentiment analysis, and it also helps people better understand and utilize the relationship between different data sources, thus promoting the development of interdisciplinary research. With the continuous exploration and research on multimodal

sentiment analysis tasks, multimodal sentiment analysis tasks can be subdivided into two subtopics: 1. conversational multimodal sentiment analysis; 2. narrative multimodal sentiment analysis.

A. Conversational Multimodal Sentiment Analysis

Conversational multimodal sentiment analysis refers to sentiment analysis based on multiple modal data sources (such as text, voice, video, etc.) for a situation or scene involving multiple rounds of interaction. This task enables better understanding of and coping with emotional changes in conversations by continuously identifying and tracking emotional states during conversations.

Interactive multimodal sentiment analysis usually includes three tasks, namely emotion recognition, emotion expression, and emotion interaction. The goal of emotion recognition is to identify the emotional state of the participants and infer the emotional state of the speaker via the analysis of multiple modal data sources, such as voices, facial expressions, and text. Emotional expression is analyzed to understand the interactive behavior of the participants and infer the way they express their emotions, such as anger, happiness, and other emotions and expressions. The goal of emotional interaction is to analyze the emotional transmission and interaction between different participants, such as analyzing the transmission, transfer, and empathy of emotions in a dialogue.

B. Narrative Multimodal Sentiment Analysis

Narrative multimodal sentiment analysis refers to sentiment analysis based on multiple modal data sources (such as text, image, video, etc.) for a static multimodal dataset (such as a video or a set of images). The goal of this task is to analyze the emotional information shown in these data in order to better understand the meaning and emotional color behind it. Differently from conversational multimodal sentiment analysis, narrative multimodal sentiment analysis does not need to consider multiple rounds of interaction. It pays more attention to the connection of contextual information within the data samples and the fusion of information from different modalities.

Zadeh et al. (2018) [11] proposed a new model, the TFN (Tensor Fusion Network), which can learn end-to-end intra- and inter-modal dynamics while aggregating interactions between unimodal, bimodal, and trimodal inputs. The model fuses the features of different modalities using mathematical matrix operations and uses the tensor outer product between modalities to calculate the correlation between elements of different modalities. In the calculation process, as the product of the matrix continues to expand, the dimensionality of the eigenvector will also increase greatly, which eventually makes the model too large to be trained. Shankar et al. (2022) [12] proposed a new multimodal fusion architecture, that is, progressive fusion (Profusion), which improves the problem that the tensor fusion network is difficult to train due to feature vector dimensions that are too large. Connecting late fusion representations to unimodal feature generators via backlinks enables early layers to provide multimodal information, revealing the advantages of early fusion and late fusion.

Chen M et al. (2017) [13] proposed an application of gated multimodal embedded long short-term memory (LSTM) with temporal attention for the word-level fusion of multimodal inputs. Gated multimodal embeddings ease the difficulty of fusion, while LSTM(A) with temporal attention performs word-level fusion. Agarap A F et al. (2018) [14] conducted a sentiment analysis on reviews using a bidirectional recurrent neural network (RNN) with long short-term memory. After analyzing the dataset, irrelevant features were removed, but due to the problem of gradient disappearance in the neural recurrent network, the time series of the neural recurrent network could not be too long, only short-term information could be remembered, and long-term information could not effectively be used.

Bao L et al. (2019) [15] proposed a long short-term memory network (LSTM) with an attention mechanism. This model combines lexical information with the attention LSTM model and uses a deep neural network to make the framework more stable; thus, no additional tag data are required. Compared with the RNN network, the LSTM network

introduces a gating mechanism to control the circulation and loss of features, which can effectively solve the gradient explosion problem of the RNN network and support the memory of long sequences. However, LSTM requires a lot of computation during training due to the numerous parameters of LSTM and the computational complexity between each gate.

Chung J et al. [16] proposed a gated recurrent neural network model (GRU). They simplified the structure and training parameters of LSTM, and the GRU only uses two gating switches, which reduces the risk of overfitting with too many parameters and greatly reduces the amount of calculation during training. The GRU still has the problem of not being able to perform parallel computing.

Graves A et al. [17] were the first to use a bidirectional LSTM model to solve the multimodal sentiment analysis problem. They obtained better results than the LSTM model. However, due to the bidirectional LSTM's greater number of parameters, the calculation cost is greater, and the calculation time required is longer. Hamborg F et al. used a bidirectional GRU model to embed language models and knowledge from external sources. Compared with the bidirectional LSTM model, the parameters of the BiGRU model are relatively few, the calculation time is short, and it can effectively solve problems that cannot be calculated in parallel. When dealing with problems with less data, it can obtain better results than the BiLSTM model.

Our proposed approach differs from existing work in that our model framework assigns larger weights to adjacent utterances, thereby exploiting contextual information for utterance-level sentiment prediction. We use the multimodal complex attention mechanism framework to exploit the contributed features across multiple modalities and adjacent utterances for sentiment analysis to obtain better sentiment analysis results.

3. Method

This section mainly introduces the framework structure of the model. In our proposed framework, we aim to leverage multiple modalities and contextual information for utterance sentiment prediction. Our proposed framework inputs multilingual formal information (including textual, visual, and acoustic information) of a series of utterances into three independent bidirectionally gated recurrent units.

The Bi-GRU is a bidirectional gated recurrent neural network (bidirectional gated recurrent unit), which is composed of GRUs (gated recurrent units) in two directions, one from left to right and one from right to left, which can perform sequential-data-processing two-way modeling. The Bi-GRU is widely used in natural language processing, time series data analysis, and other fields, and it is an excellent sequence modeling tool. In the Bi-GRU, the GRU is a special recurrent neural network (RNN), which can memorize and process the input sequence and generate a hidden state vector for the next step of prediction. Compared with a traditional RNN, the GRU has fewer parameters, faster convergence speed, and can better obtain the information of longer sequences. Compared with LSTM (long short-term memory), it has fewer parameters and a faster calculation speed. On the basis of the GRU, the Bi-GRU adds the ability of two-way modeling, which can better use contextual information for prediction and improves the performance and accuracy of the model.

3.1. Attention

Multimodal sentiment analysis in natural language processing usually includes sentiment analysis tasks on multiple modal data sources such as text, images, and audio [18]. Attention mechanism is a technique widely used in multimodal sentiment analysis, which can improve the performance of the model [19].

In multimodal sentiment analysis, attention mechanisms can be used to select the most relevant features in text, images, and audio to better capture emotional information in different modal data. In terms of specific implementation, the common attention mechanisms are as follows:

1. The text attention mechanism can be used to select the most relevant features from multiple text features. It is usually based on an attention score, which is determined by computing the similarity between text features and sentiment labels. A common implementation is to use an attention-based mechanism to compute a weighted sum of text features to generate the final sentiment representation.
2. The image attention mechanism can be used to pick out the most relevant image regions. It is usually determined based on the attention distribution of visual features. A common implementation is to use a convolutional neural network (CNN) to extract image features and then to use an attention mechanism to select the most relevant features and combine them into a final emotion representation.
3. The audio attention mechanism can be used to select the most relevant audio features. It is usually based on an attention score, which is determined by computing the similarity between audio features and sentiment labels. A common implementation is to use an attention-based mechanism to compute a weighted sum of audio features to generate the final emotion.

3.2. Complex Attention Mechanism

We use a complex attention mechanism to fuse three modalities (text, audio, and vision). Complex attention mechanisms are variants of attention mechanism that can capture different levels of semantic information. Compared with the traditional simple attention mechanism, the complex attention mechanism divides the input into different parts, uses different linear transformations for calculation, and then combines them to produce the final attention weights. This approach can better represent different levels of semantic information and thus enhance the performance of the model.

As shown in Figure 1, after the original data are extracted by different methods, they are input into the recurrent neural network, and the output results are multimodally fused using the complex attention mechanism layer. The workflow of the complex attention mechanism is as follows: First, linearly transform the input tensor X and map it to three tensors: query, key, and value. Next, the Q, K, and V tensors are divided into simple attention heads and complex attention heads, and the number of each head is set to 2. Specifically, each tensor is divided into simple attention heads and complex attention heads. In the next step, the attention weights are calculated for the simple attention head and the complex attention head.

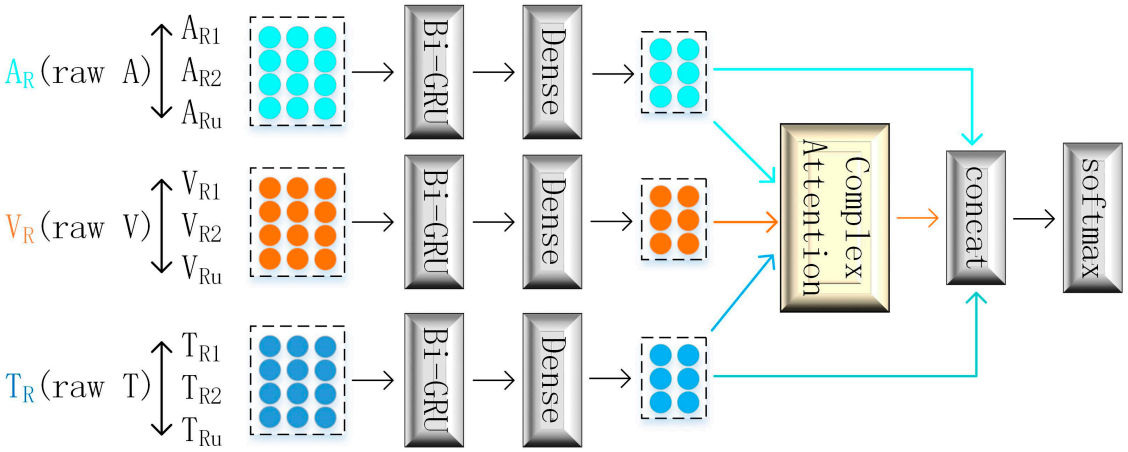

Figure 1. Overview of the proposed framework.

In the simple attention head, the input tensor will be transformed by three matrices to obtain three low-dimensional tensors: Q, K, and V. Next, for each simple attention head, calculate the attention weight matrix A—each element of A is obtained by the inner product of Q and K—and then multiply A and V to obtain the final output. The attention weight matrix A of the simple attention head is computed by the inner product of Q

and K. In the complex attention head, the input tensor will also be transformed by three matrices to obtain three low-dimensional tensors: Q, K, and V. The difference is that the complex attention head uses an extra linear layer to compute the inner product of Q and K, which introduces a stronger nonlinear capability. As shown in Figure 2. Therefore, the attention weight matrix A of the complex attention head is obtained by performing product calculation on the linearly transformed Q and K instead of directly performing the inner product calculation on Q and K. This transformation enables the model to better capture high-level features in the input tensor. The attention weight is calculated by first multiplying the Q tensor and the K tensor and then dividing by the scaling factor, which is an adjustable parameter to scale the calculation result.

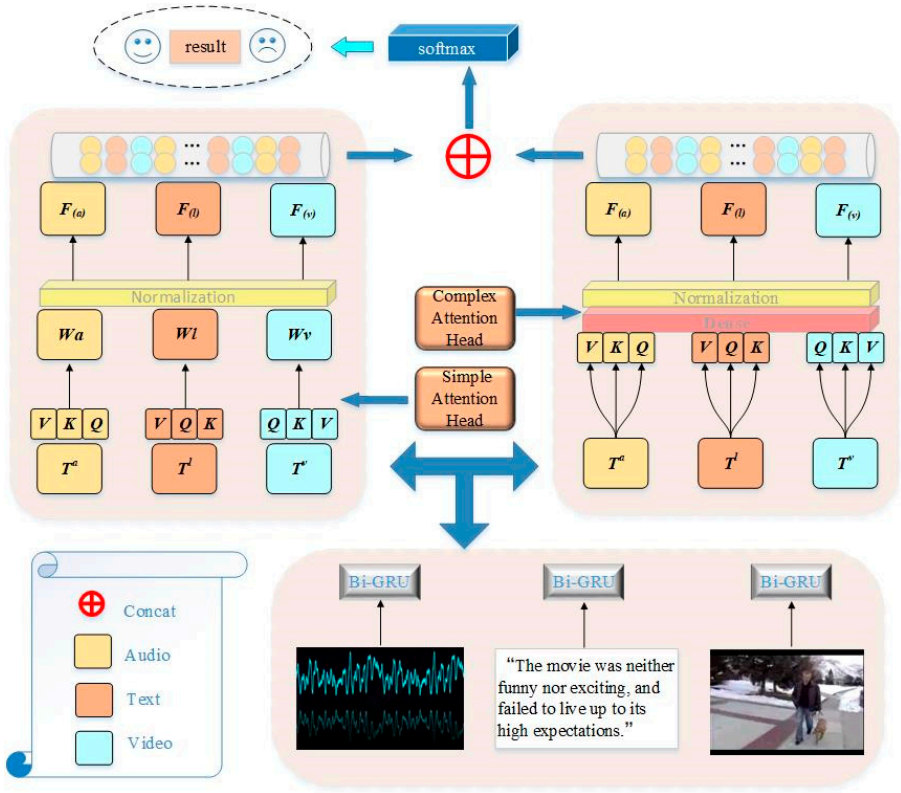

Figure 2. Overview of the complex attention head.

Next, the attention weights of the simple and complex attention heads are spliced along the head dimension to obtain a merged tensor of shape (num_heads+num_complex_heads, batch_size, seq_len, seq_len), and then the total attention weight is combined with V, and the quantities are multiplied to obtain a new tensor. Add the above tensors to the residual connection, that is, add them to the input tensor X, and finally, perform normalization. Output the processed tensor as the result of complex attention mechanism:

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

$$W_{head} = \sum_{i}^{n} Head_i \tag{2}$$

where $\sqrt{d_k}$ is an adjustable parameter used to scale the calculation results, and W_{head} is the total attention weight after the simple attention head weight and the complex attention head weight are spliced.

3.3. Classification Layer

After obtaining the result of the complex attention mechanism, a softmax layer with a size of 1 will be used to obtain a vector result, and the vectors of various forms will be summed element by element. The formula is as follows:

$$y \sim p = W_a(LayerNorm(S_i))^{i\in[L,A]}. \tag{3}$$

Among them, y is the final output result, S_i is the tensor after complex attention mechanism, and W_α is the weight parameter.

3.4. Single-Sentence Complex Attention Framework (SSCAF)

Under the SSCAF, a complex attention mechanism is used to compute a single sentence. This framework does not consider utterance information from other attention levels but utilizes multimodal information of a single utterance for prediction. Differently from the simple attention mechanism, the complex attention mechanism divides the input vector into multiple parts and uses different weight matrices to calculate the relationship between different parts. Suppose we have a unimodal input tensor X, X ∈ {batch_size, seq_len, input_dim}, where input_dim indicates the number of elements or features in the input vector. Use num_head complex attention heads, and each head uses three parameter matrices for calculation: W^Q, W^K, and W^V.

First, we divide the input tensor X into num_heads subvectors, that is, X = [X_1, X_2, . . ., X_{num_heads}]. Then, we use W^Q, W^K, and W^V to linearly transform each subvector to obtain $Q_i = X_iW^Q$, $K_i = XiW^K$, and $V_i = X_iW^V$. Note that the weight matrices W^Q, W^K, W^V here are shared for the entire input tensor. Then, we calculate the attention weight matrix A_i for each subvector Q_i, K_i, and V_i and achieve A_i = softmax ($Q_iK_i^T$/sqrt{d_k}). Finally, we multiply the attention weight matrix A_i with the corresponding subvector Vi and splice to achieve the final output tensor Y, namely Y = [Y_1, Y_2, . . ., Y_{num_heads}] W_O, where W_O is a parameter matrix for linear transformations. Note that it is necessary to ensure that the dimensions of the subvectors are consistent during splicing, which can be achieved by adding zero vectors at the end of shorter subvectors. The following is the complex attention mechanism formula for computing a single sentence:

$$\begin{aligned}
&MultiHead(Q,K,V) = concat(head_1,\ldots,head_h)W^o\\
&head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)\\
&Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V\\
&W_i^Q \in R^{d_{model}\times d_k}, W_i^K \in R^{d_{model}\times d_k}, W_i^V \in R^{d_{model}\times d_v}, W^O \in R^{hd_v\times d_{model}}
\end{aligned} \tag{4}$$

where W_i^Q, W_i^K, and W_i^V represent the weight matrix of the Q, K, and V matrices of the *i*-th header, respectively; W_O is the weight matrix of the output matrix; h is the header; d_k and d_v are the dimensions of the key matrix and value matrix; and dmodel is the dimension of the hidden layer of the model.

3.5. Context-Complex Attention Framework (CCAF)

Under the CCAF, a complex attention mechanism is applied to the utterances of each modality, based on which the classification is performed. Differently from the SSCAF, the CCAF uses the contextual information of the utterance to calculate its Q, K, and V matrices for each modality:

$$\begin{aligned}
Q_i^m &= Q^m W_{Q,i}^m\\
K_i^m &= K^m W_{K,i}^m\\
V_i^m &= V^m W_{V,i}^m
\end{aligned} \tag{5}$$

where, i ∈ {1, 2, . . ., h}; h represents the number of heads; and WQim, WKim, and WVim represent the weight matrix used to divide the Q, K, and V matrices, respectively.

Concatenate the divided Q, K, and V matrices to obtain the final Q, K, and V matrices:

$$\begin{aligned} Q_i &= [Q_i^1, Q_i^2, \ldots, Q_i^m] \\ K_i &= [K_i^1, K_i^2, \ldots, K_i^m] \\ V_i &= [V_i^1, V_i^2, \ldots, V_i^m] \end{aligned} \tag{6}$$

where m represents the number of modalities. Then, calculate its corresponding attention weight matrix for each head:

$$Attention_i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}}) \tag{7}$$

where d_k represents the dimension of the key matrix. For each head, the output result of the head is calculated according to its corresponding attention weight matrix and value matrix, and the output results of all heads are spliced to obtain the final output result, where W_O means to combine all. The weight matrix of the concatenation of the output results in the head. The resulting O is the multimodal data:

$$O_i = Attention_i V_i \tag{8}$$

$$O = [O_1, O_2, \ldots, O_h] W_O \tag{9}$$

4. Data Preparation

4.1. CMU-MOSEI Dataset

We validate our experiments using the CMU-MOSEI dataset, which contains a collection of 2084 speaker video clips, each of which is essentially a monologue, containing three forms: spoken language in text form, vision in the movements and facial expressions; voice in intonation and rhythm. Each sentence is annotated with a variety of tags. First, the Likert scale is used to mark the emotion, which is divided into seven categories: [highly negative, negative, weakly negative, neutral, weak positive, positive, and highly positive]. Ekman emotions {happiness, sadness, anger, fear, disgust, and surprise} are also annotated on a Likert scale of [0, 3] for denoting sentiment x: [no evidence for x, weak x, x, strong x] [20].

4.2. Linguistic Feature

When extracting text features, all sentences are lowercased first, and special characters and punctuation marks are removed. Build a vocabulary containing only unique words and use the unsupervised word embedding Glove model to embed each word in a 300-dimensional vector [21]. Label the words that are not in the vocabulary in the validation or test set as "unk".

4.3. Acoustic Feature

Audio is one of the important ways by which human beings can obtain information. Among them, nonverbal information, such as laughter, gasps, and sighs, and the rhythmic characteristics of speech, such as speech rate and intonation, often convey more complex information. These nonverbal expressions are also used as data in emotion classification tasks [22]. In terms of audio feature extraction, common methods include the following: 1. the zero-crossing rate, that is, the number of intersection points of the signal on the time axis; 2. spectral quality, that is, the center position of the signal in the frequency domain; 3. the features obtained by a specific model; 4. MFCCs, that is, the features of the sound extracted by simulating the characteristics of human hearing. When extracting audio features, in order to reduce the influence of noise, it is first needed to remove irrelevant sounds and then to focus on vocals. For this, we use Mel-Frequency Cepstral Coefficients (MFCCs) to extract acoustic features. Combine N sampling points into one observation unit, usually about 20–30 ms. In order to avoid too-large changes between two frames,

we set an overlapping region containing M sampling points, where M is about 1/2 to 1/3 of N. Each sampled frame is filtered using a Mel filter, and then inverse discrete Fourier transform is performed to obtain features containing 80 dimensions.

4.4. Visual Feature

We chose to use a convolutional neural network to preprocess features. We used a CNN with 3D convolutional kernels to process the temporal and spatial information of videos to extract features for sentiment analysis tasks. The model has been preprocessed on Sports-1M and Kinetics. In this model, we take a video clip of 32 RGB frames as input and slide between the 32 RGB frames with a stride of 8 frames to obtain a feature vector for the entire video.

5. Experiments

In this section, we present the results of model experiments on a dataset based on our CMU-MOSEI dataset.

5.1. Experimental Setting

We built a complex attention mechanism model based on the Tensorflow deep learning framework and trained the model on an NVIDIA Tesla V100 GPU. We incorporated an early stopping mechanism with a default value of 3 and set the dropout rate to 0.1. All experiments are performed without using pretrained models.

We use the BiGRU with 300 neurons, each followed by a dense layer of 100 neurons. With dense layers, we project the input features of all three modalities to the same dimension.

Optimize the training of the model using the Adam optimizer and set the learning rate to make the training converge to better performance. Set the random number seed to ensure that the random number in the tensor of each run remains unchanged. The batch size is set to 32, the early stop mechanism is added, and the patience value is set to 5; that is, if the accuracy on the validation set does not increase over the prespecified batches, then the training will stop after five batches. The updated formula of the optimizer is as follows:

$$\theta = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v} + \varepsilon}} \hat{m}_t \tag{10}$$

In the experiments, we used two simple attention heads and two complex attention heads to construct the complex attention mechanism. We adopt cross-entropy as the loss function, and the model weights are set as time weights, which were used to adjust sample weights at different time steps in order to more effectively utilize contextual information in the sequence. In time weighting, earlier time steps are usually given lower weights, and newer time steps are given higher weights. This weight assignment reflects the importance of time in sequence data, because in many sequence data, information closer to the current time step tends to have a greater impact on the task. By using time weights, the model can pay more attention to the newer information to better capture the contextual relationship and timing patterns in the sequence data so that the model can better adapt to the characteristics of the sequence data.

5.2. Experimental Results

We validated our model on the CMU-MSOEI dataset and compared it with the results obtained by other models. We experimented with all valid combinations of textual, visual, and auditory inputs, including unimodal (only one modality at a time) and trimodal (all three modalities at a time) inputs, and measured accuracy using an evaluation index.

For the MOSEI dataset, we achieve better performance using text. Subsequently, we construct the input with three modalities and feed it to the network. For unimodal input, the text in the SSCAF has the highest accuracy, followed by visual and audio inputs. The accuracies are 80.12%, 79.92%, and 79.79%, respectively, and the result obtained by

simultaneously inputting the three modes is 81.78%. The above results are the average results obtained from five experiments. As shown in Table 1.

Table 1. Classification results using various modality combinations on the CMU-MSOEI dataset.

Method	Modality			Contextual LSTM [23]	MU-SA and MMUU-SA [24]	B2+B4 [25]	Proposed
	T	V	A				
Unimodal	√			76.75%	78.23%	/	80.12%
		√		71.84%	74.84%	/	79.92%
			√	70.94%	75.88%	/	79.79%
Trimodal	√	√	√	77.64%	79.80%	81.14%	81.78%

In an RNN, each input sequence may be of a different length. This can present some challenges when trying to assign weights to each sample as some samples may be more important than others. Therefore, the mode assignment weights of different sample weights can assign different weights for each time step to consider the timing of samples and better train the model. As shown in Table 2.

Table 2. Effect of weight on network classification performance.

Modality	Time-Wise	Sample-Wise	None	Temporal
T	78.72%	78.99%	78.19%	80.12%
V	77.53%	78.86%	76.59%	79.92%
A	74.34%	77.39%	71.01%	79.79%
T+V+A	81.38%	80.85%	80.72%	81.87%

The optional sample weight assignment mode [26] has the following four classifications: 1. "Temporal" assigns different weights to each time step and trains the model considering the timing of samples. 2. "Time-wise" assigns a weight to each time step. The weight of the same sample at different times can be different, but the weight of different samples at the same time is the same. 3. "Sample-wise" assigns a weight to each sample, and this weight is the same during the training process. 4. "None" assigns no sample weights, and the weights of all samples are set to 1, which is the default value.

In the attention mechanism, the weight of each neuron is calculated according to the relationship between it and other neurons, so the interaction relationship between different modalities can be better handled. We use the complex attention mechanism to fuse the three modalities [13].

The self-attention mechanism is an application of the attention mechanism, which realizes the understanding and representation of sequences by connecting the connections between different positions in a sequence. In the self-attention mechanism, the representation of each position is obtained by a weighted summation of the representations of all other positions in the sequence, where the weight of each position to other positions is calculated by a series of linear transformation and softmax operation.

Multi-head attention is an extension of the self-attention mechanism [27], which allows the model to perform self-attention calculations in multiple different subspaces. Specifically, after the multi-head attention mechanism linearly transforms the input, it divides the transformed results into multiple heads, performs independent self-attention calculations on each head, and finally combines all heads. The outputs are stitched together and subjected to a final linear transformation. The benefit of this is that it allows the model to focus on multiple subspaces with different levels of attention, thus gaining a better understanding of the input.

Complex attention is an extension of the multi-head attention mechanism, which can handle inputs with different modalities. In the complex attention mechanism, the input of each modality is first linearly transformed into different subspaces, and then an

independent attention computation is performed in each subspace. Finally, the outputs of all modalities are concatenated together, and the final output is obtained via linear transformation. This method can handle multiple input types, such as text, images, and speech, and has good expressiveness when dealing with multimodal input. As shown in Table 3.

Table 3. Effect of attention on network classification performance.

Method	Modality			Self-Attention	Multi-Head Attention	Proposed
	T	V	A			
Unimodal	√			78.32%	78.99%	80.12%
		√		77.13%	78.32%	79.92%
			√	76.86%	77.79%	79.79%
Trimodal	√	√	√	80.05%	80.85%	81.78%

6. Conclusions

We propose a novel variant of the attention mechanism, referred to as the complex attention mechanism. Firstly, we employ the Bi-GRU model to capture the sequential patterns within the original data [28]. The Bi-GRU is extensively used in natural language processing, time series data analysis, and other fields, serving as an exceptional tool for sequence modeling. Comprising two directions of gated recurrent units (GRU), the Bi-GRU models time series data bidirectionally—from left to right and from right to left. GRU's lower parameter count and faster convergence speed enhance its ability to comprehend lengthy sequences of information. Building upon the GRU, the Bi-GRU incorporates bidirectional modeling, effectively leveraging contextual information for improved predictive accuracy. Subsequently, the output from the Bi-GRU model is channeled into a complex attention mechanism, which comprises both simple and complex attention heads. By introducing an additional linear layer, certain attention heads within the multi-head attention mechanism are transformed into complex attention heads. The inner product of Q and K is computed, generating the attention weight matrix for the complex attention head via linear transformation. In contrast to the multi-head attention mechanism, the complex attention mechanism introduces heightened non-linearity, facilitating robust contextual association and thereby enhancing the accuracy of multimodal sentiment analysis. Following this, the output of the complex attention mechanism is passed through a softmax layer to obtain the final sentiment analysis result. We conducted experiments on all valid combinations of textual, visual, and auditory inputs using the CMU-MOSEI dataset. This encompassed both unimodal and trimodal inputs, with accuracy serving as the evaluation metric. Our approach yielded superior results, underscoring its effectiveness in enhancing sentiment analysis across different modalities.

Author Contributions: Conceptualization, L.D.; data curation, L.D. and H.L.; formal analysis, L.D. and Z.L.; investigation, Z.L.; methodology, L.D. and B.L.; software, B.L. and Z.L.; supervision, Z.L. and J.M.; validation, B.L. and J.M.; writing—review and editing, B.L; writing—original draft preparation, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the Henan Provincial Science and Technology Research Project under Grants 232102211006 and 232102210044, the Songshan Laboratory Pre-research Project under Grant YYJC012022023, 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, and the Academic Degrees and Graduate Education Reform Project of Henan Province under Grant 2021SJGLX115Y.

Data Availability Statement: The data presented in this study are openly available in [CMU-MOSEI] at [10.18653/v1/P18-1208] reference number [20].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
2. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A survey of multimodal sentiment analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
3. Verma, S.; Wang, C.; Zhu, L.; Liu, W. Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
4. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [CrossRef]
5. Biesialska, K.; Biesialska, M.; Rybinski, H. *Sentiment Analysis with Contextual Embeddings and Self-Attention*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 32–41.
6. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394. [CrossRef]
7. Al-Absi, A.A.; Kang, D.K.; Al-Absi, M.A. Sentiment Analysis and Classification Using Deep Semantic Information and Contextual Knowledge. *CMC—Comput. Mater. Contin.* **2023**, *74*, 671–691.
8. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
9. Yadav, A.; Vishwakarma, D.K. A deep multi-level attentive network for multimodal sentiment analysis. *ACM* **2023**, *19*, 1–19. [CrossRef]
10. Lin, F.; Liu, S.; Zhang, C.; Fan, J.; Wu, Z. StyleBERT: Text-audio sentiment analysis with Bi-directional Style Enhancement. *Inf. Syst.* **2023**, *114*, 102147. [CrossRef]
11. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
12. Shankar, S.; Thompson, L.; Fiterau, M. Progressive Fusion for Multimodal Integration. *arXiv* **2022**, arXiv:2209.00302.
13. Chen, M.; Wang, S.; Liang, P.P.; Baltrušaitis, T.; Zadeh, A.; Morency, L.P. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 163–171.
14. Agarap, A.F. Statistical analysis on E-commerce reviews, with sentiment classification using bidirectional recurrent neural network (RNN). *arXiv* **2018**, arXiv:1805.03687.
15. Bao, L.; Lambert, P.; Badia, T. Attention and lexicon regularized LSTM for aspect-based sentiment analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Florence, Italy, 28 July–2 August 2019; pp. 253–259.
16. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
17. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM networks. In Proceedings of the IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; pp. 2047–2052.
18. Das, R.; Thoudam, S.D. Multimodal sentiment analysis: A survey of methods, trends and challenges. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; MIT Press: Cambridge, UK, 2017.
20. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. 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, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2236–2246.
21. Zang, Y.; Qi, F.; Yang, C.; Liu, Z.; Zhang, M.; Liu, Q.; Sun, M. Word-level textual adversarial attacking as combinatorial optimization. *arXiv* **2019**, arXiv:1910.12196.
22. Wu, T.; Peng, J.; Zhang, W.; Zhang, H.; Tan, S.; Yi, F.; Huang, Y. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl.-Based Syst.* **2022**, *235*, 107676. [CrossRef]
23. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 873–883.
24. Delbrouck, J.B.; Tits, N.; Brousmiche, M.; Dupont, S. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv* **2020**, arXiv:2006.15955.
25. Kumar, A.; Vepa, J. Gated mechanism for attention based multi modal sentiment analysis. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 4477–4481.

26. Chen, M.; Li, X. Swafn: Sentimental words aware fusion network for multimodal sentiment analysis. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 13–18 September 2020; pp. 1067–1077.
27. Kim, K.; Sanghyun, P. Aobert: All-modalities-in-one BERT for multimodal sentiment analysis. *Inf. Fusion* **2023**, *92*, 37–45. [CrossRef]
28. Ghosal, D.; Akhtar, M.S.; Chauhan, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Contextual inter-modal attention for multi-modal sentiment analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3454–3466.