CCDA: A Novel Method to Explore the Cross-Correlation in Dual-Attention for Multimodal Sentiment Analysis

Peicheng Wang , Liu Shuxian * , Chen Jinyan

Posted Date: 7 December 2023

doi: 10.20944/preprints202312.0472.v1

Keywords: Multimodality; Sentiment Analysis; Attention Mechanism

Article

CCDA: A Novel Method to Explore the Cross-Correlation in Dual-Attention for Multimodal Sentiment Analysis

Peicheng Wang, Shuxian Liu * and Jinyan Chen

School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
* Correspondence: liushuxian@xju.edu.cn

Abstract: With the development of the Internet, The content people share contains types of text, images, and videos, and utilizing these multimodal data for sentiment analysis has become an important area of research. Multimodal sentiment analysis aims to understand and perceive emotions or sentiments in different types of data. Currently, the realm of multimodal sentiment analysis faces various challenges, with a major emphasis on addressing two key issues: 1) Inefficiency when modeling the intra-modality and inter-modality dynamics and 2) Inability to effectively fuse multimodal features. In this paper, we proposed the CCDA(Cross-Correlation in Dual-Attention) model, a novel method to explore dynamics between different modalities and fuse multimodal features efficiently. We capture dynamics at intra- and inter-modal levels by using two types of attention mechanisms simultaneously. Meanwhile, the cross-correlation loss is introduced to capture the correlation between attention mechanisms. Moreover, the relevant coefficient is proposed to integrate multimodal features effectively. Extensive experiments were conducted on three publicly available datasets, CMU-MOSI, CMU-MOSEI, and CH-SIMS. The experimental results fully confirm the effectiveness of our proposed method, and compared with the current optimal method (SOTA), our model shows obvious advantages in most of the key metrics, proving its better performance in multimodal sentiment analysis.

Keywords: multimodality; sentiment analysis; attention mechanism

1. Introduction

Multimodal sentiment analysis(MSA) is an important branch in the field of artificial intelligence. It aims to capture and understand human sentiment or emotion contained in text, speech, images, or other types of data, usually including positive, negative, neutral, or more specific emotional states such as joy, sadness, and anger [1]. In recent years, with the popularity of online social platforms, a large amount of multimodal data has emerged on the Internet. By analyzing data containing multiple modalities, computers can perceive human sentiment in the data [2]. Multimodal sentiment analysis has attracted widespread attention and it is widely applied in social media analysis [3,4], market research [5,6], and human-computer interaction [7,8]. Nevertheless, there are two main challenges in current multimodal sentiment analysis research. The first one is inefficiency in modeling the intra-modality and inter-modality dynamics. Multimodal sentiment analysis requires processing data from different modalities and correlating them to capture sentiment. It also needs to deal with sentiment dependencies within a single modality to help the model understand sentiment more accurately. Effective integration of features from different modalities can improve the accuracy and robustness of the model, which is crucial for the reliability of sentiment analysis in practical applications.

In early studies of multimodal sentiment analysis, researchers have used two main approaches to process multimodal data: The first one is simply concatenating multimodal features [9–14] (commonly referred to as early fusion), and the second one is late fusion [15–18]. Although these two methods were relatively simple, when dealing with modal features, the model is unable to capture intra- and inter-modality dynamics efficiently, which may lead to poor model performance. Subsequently, researchers proposed more complex methods such as hybrid fusion [19–21] and model-level fusion [22–

 doi:10.20944/preprints202312.0472.v1

25] to further explore the correlations between different modalities. Additionally, attention mechanisms have been introduced into multimodal sentiment analysis to compute intra- and inter-modality correlations by using attention mechanisms [26,27]. With the invention of Transformer [28] and its outstanding performance in the field of natural language processing, Transformer has been widely used in other research areas such as multimodal sentiment analysis. For example, [29–32] leverage the Transformer encoder to model correlation information between different modalities and have achieved good results in multimodal sentiment analysis. Some scholars have used tensor-based fusion methods [33–35] to solve the problem of fusion of multimodal features, and there are other researchers have adopted other methods such as self-supervised learning [36], contrastive learning [37], multi-task learning [38], etc.

In this paper, we use a transformer-based approach to capture sentiment information and extract dynamics within and between modalities, and we introduce the relevant coefficient for the fusion of multimodal features. In addition, we propose a new cross-correlation loss function for investigating the correlations between different levels of attention mechanisms. Specifically, we obtain the inter-modality dynamics between the global representation and unimodal representation by using the cross-attention mechanism, which is the component of the Transformer. So that they can strengthen themselves by learning about each other in this process. At the same time, we obtain the intra-modality information by using the self-attention mechanism for three unimodal features respectively. In addition, in our research, we hypothesized that there is some correlation between different levels of attention mechanisms, so we proposed the cross-correlation loss to assess the interrelationship between cross-attention and self-attention. The contributions of this paper can be summarized as follows:

- We propose CCDA, a hierarchical model that studies intra- and inter-modality correlations by using self-attention and cross-attention, respectively. Moreover, we introduce a new method to fuse multimodal features efficiently.
- We innovatively introduce a new cross-correlation loss function to study the correlation between different levels of attentional mechanisms in more depth. The objective function is minimized to cut down redundant information, which can help our model to better perceive sentiment information.
- Extensive experiments demonstrated the effectiveness of our proposed methodology. Our model achieves comparable results to the state-of-the-art (SOTA) approach in all evaluation metrics on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets.

2. Related works

Multimodal sentiment analysis aims to obtain sentiment information from different types of data. It provides additional sources of information for affective computing and enables computers to understand and perceive human sentiment more accurately [1–4]. A key challenge in this area is how to efficiently fuse data from different modalities so that the model can recognize sentiment precisely. This section presents related works on multimodal sentiment analysis, including early fusion, late fusion, hybrid fusion, word-level fusion, tensor-based fusion, attention-based methods, and other recent research approaches.

Early fusion combines all of the features from different modalities(text, audio, and visual) into a single feature vector, which is then used for sentiment prediction using a classification algorithm or model [9,12–14,26,39,40]. The advantage of this approach is that it can take into account the correlation between different modality features at the early stage. However, there may be asynchrony in the way features are extracted for different modalities, leading to temporal inconsistency between features. Additionally, early fusion may contain a lot of redundant and conflicting information, which reduces the learning efficiency of the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

Table 1. Related works in multimodal sentiment analysis.

Method type	Description	Advantages	Flaws
Early fusion	Combines all of the features from different modalities into a vector.	Realizes modal interactions at the early stage.	Time asynchrony and information redundancy
Late fusion	Employs independent classifiers separately for each modality.	Helps model to better integrate semantic information	Usually involves more complex model structures.
Hybrid fusion	Combines the advantages of early fusion and late fusion	Balance the moddl's complexity.	Inefficiency
Word-level fusion	Fuses word representation in the temporal dimension.	Helps model to understand the intrinsic relation of multimodal data.	Insufficient generalization
Tensor-based	Utilizes various tensor-based methods to integrate information from different modalities.	Integrate multimodal data effectively and address the complexity and noise issues.	Excessive computation and lack of interpretability.
Attention-based	Learns the semantic and relevant information using different attention mechanisms.	More flexible and accurate in processing multimodal data	Models multimodal information inefficiently in some cases.
Transformer-based	Models the correlation between different modalities using transformer	Ability to process temporal information and capture interactions between different modalities.	Closely related to the depth of the model.

In contrast to early fusion, late fusion employs independent classifiers separately for each unimodal data and then fuses the outputs of each model to generate the final multimodal representation, or votes on the results of each model [15–18]. While late fusion helps the model to better integrate semantic information, it usually involves more complex model structures.

Hybrid fusion combines the advantages of early fusion and late fusion, capitalizing on their strengths and compensating for their weaknesses respectively. The core idea of this approach is to allow features to be fused at different stages of the model while avoiding some of the potential problems of early fusion and late fusion [19–21,41]. In addition, Hybrid fusion is able to better balance model complexity.

Word-level fusion is a method that fuses word representations in the temporal dimension to capture the interrelationships between different modalities. This approach emphasizes word-level information interactions and helps to understand the intrinsic structure and semantic relatedness of multimodal data in more detail [22–25]. Word-level fusion improves the model's ability to process and model multimodal sentiment data by fusing information from different modalities in word representations.

Tensor fusion utilizes various tensor-based methods to integrate information from different modalities. These methods can effectively integrate multimodal data and address the complexity and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

noise issues in the data [33–35,42–46]. However, these methods may suffer from excessive computation and lack of interpretability.

Attention mechanism plays a significant role in multimodal sentiment analysis, it helps models better understand and leverage the interconnections and semantic information between different modalities, and to be more flexible and accurate in processing multimodal data [13,26,31,47,48]. Google proposed the Transformer [28] in 2017, which has been very successful in natural language processing. Subsequently, Transformer was introduced into various domains including multimodal sentiment analysis, and spawned a series of significantly innovative approaches. The Transformer model can model temporal information in the data and process unimodal data through the self-attention mechanism, and Transformer can also achieve the interaction between different modalities [29–32,48–52]. Furthermore, the Transformer exhibits strong generalization capabilities, making it suitable for different types of multimodal sentiment analysis tasks.

In addition, there are other methods in multimodal sentiment analysis, such as multi-task contrastive learning [38], dynamic filtering mechanism [53], bidirectional multimodal dynamic routing mechanism [54], cross-modal hierarchical graph contrastive learning strategy [37], supervised contrastive learning [22], and dynamic refined sentiment words [55], etc.

3. Methodology

3.1. Problem Definition

Multimodal sentiment analysis is a task that utilizes multiple modalities for the study of human sentiment. Typically, it includes three modalities: text, speech, and images. We define three modality feature sequences, $X_m = \{x_{m,1}, x_{m,2}, \dots x_{m,n}\}$, and sample labels $Y = \{y_1, y_2, \dots y_n\}$, where the modality is represented as $m \in \{t, a, v\}$(t stands for text, a stands for audio, and v stands for visual) and n represents the number of samples in the dataset. Our goal is to input modality features $X_m \in \mathbb{R}^{T_m \times d_m \times n}$ into a model to obtain an accurate sentiment prediction label $y \in \mathbb{R}^1$, where T_m and d_m represent the sequence length and the dimension of modality features separately.

3.2. Model Structure

In this section, we will provide a detailed overview of the architecture of the CCDA (Cross-Correlation in Dual-Attention) model, as shown in Figure 1. We first use three unimodal encoders to obtain the utterance representation U_m and embedding F_m for each modality separately, which $m \in \{t, a, v\}$. This helps the model understand the semantic and sentiment information in each modality.

Next, we will delve into the Dual-Attention mechanism(which contains self-attention and cross-attention), a core component of CCDA. By utilizing self-attention and cross-attention, CCDA can capture sentiment information and dynamics within a single modality (Intra-modality) and across different modalities (Inter-modality), respectively. This Dual-Attention mechanism enables the model to comprehensively analyze multimodal data and sentiment information, thereby improving the accuracy of sentiment analysis.

Following that, CCDA fuses the unimodal and multimodal representations obtained from these two attention mechanisms to generate the final sentiment representation. It is worth noting that while obtaining information about the intra-modality and inter-modality dynamics, CCDA also calculates cross-correlation losses between the embeddings generated by the two attention mechanisms. This contributes to the indirect interaction between the two attention mechanisms and thus improves the model's performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

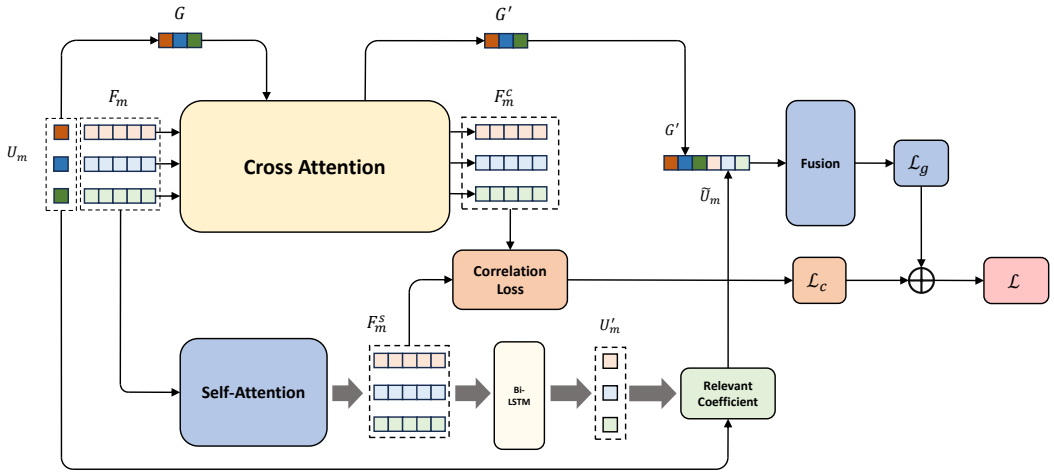

Figure 1. The structure of CCDA. The global representation G consists of three unimodal representations $\{U_t, U_a, U_v\}$. The model processes the global representation G and the unimodal features F_m using the Dual Attention to obtain new global and unimodal representations $\{G', \widetilde{U}_t, \widetilde{U}_a, \widetilde{U}_v\}$ and fuses these representations for sentiment prediction. The unimodal features $\{F_t^S, F_a^S, F_v^S, F_t^C, F_a^C, F_v^C\}$ generated during this process are used to learn the correlation between the two attention mechanisms. The final objective function consists of the prediction loss $\mathcal{L}_g$ and the cross-correlation loss $\mathcal{L}_c$.

In the following parts, we elaborate on the three main components of CCDA: unimodal encoders[Section 3.2.1], dual-Level attention[Section 3.2.2], and fusion and prediction units[Section 3.2.3].

3.2.1. Unimodal Encoders

Similar to EMT [32], we employ the pre-trained BERT model to encode textual tokens into context-aware word embeddings. Specifically, we notice that the [CLS] token of the BERT model contains a sequential representation of the text modality. Therefore, we use this token as the utterance representation for the text sequence, denoted as $u_t \in \mathbb{R}^{d_t}$. For the audio and visual modalities, we use LSTM recurrent neural networks to extract temporal information from the feature sequences. Ultimately, we select the hidden state of the last time step of the LSTM network for both the audio and visual modalities as their respective utterance representations: $u_a \in \mathbb{R}^{d_a}$ and $u_v \in \mathbb{R}^{d_v}$. Simultaneously, we need to process other tokens output by the BERT model and hidden states from LSTMs at different time steps for later use in Self-Attention and Cross-attention mechanisms. These representations are denoted as $F_m \in \mathbb{R}^{T_m \times d_m}, m \in \{t, a, v\}$, representing the text, audio, and visual modalities, respectively.

$$\begin{aligned} F_t &= BERT(X_t) \\ F_a &= LSTM(X_a) \\ F_v &= LSTM(X_v) \end{aligned} \tag{1}$$

3.2.2. Dual-Level Attention

Attention mechanisms help the model better understand multimodal sentiment data and perceive emotional information. They enable the model to capture dynamics within a single modality or between different modalities during the multimodal sentiment processing. The Transformer [28] is a language model in the field of natural language processing, it is based on dot-product self-attention mechanisms. It employs self-attention to infuse global semantic information and consider long-range dependencies for every word in the sequence. Furthermore, the multi-head mechanism allows the model to learn different subspaces of semantics.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

In simple terms, the Transformer processes the input sequence $H \in \mathbb{R}^{T \times d}$ with positional encoding, and it defines Query as $Q = HW_Q$, Key as $K = HW_K$, and Value as $V = HW_V$, where W represents the weight matrices during the feature sequence mapping process. Therefore, self-attention can be represented as equation 2:

$$Self\text{-}Attention(H) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

In MulT [29], the Query and K-V pair in the self-attention computation process come from different modalities. Thus, MulT captures the interaction between the two modalities. MulT combines three modality pairs and calculates bidirectional modality interactions for each pair. As shown in equation 3, For two modality feature sequences H_1 and H_2, MulT defines Query as $Q_1 = H_1W_Q$, Key as $K_2 = H_2W_K$, and Value as $V_2 = H_2W_V$. It calculates cross-modal attention in two directions between a pair of modalities:

$$\begin{aligned} Cross\text{-}Attention(H_1 \rightarrow H_2) &= softmax(\frac{Q_1K_2^T}{\sqrt{d_k}})V_2 \\ Cross\text{-}Attention(H_2 \rightarrow H_1) &= softmax(\frac{Q_2K_1^T}{\sqrt{d_k}})V_1 \end{aligned} \tag{3}$$

EMT [32] concatenates three unimodal utterance representations into a multimodal global representation. Inspired by EMT [32], we concatenate the utterance representations from each modality u_m as the global representation $G = Concat(u_t, u_a, u_v)$ during the Cross-Attention stage, where $m \in (t, a, v)$. Subsequently, we utilize a Transformer to calculate inter-modality information between the modality feature sequences $F_m \in \mathbb{R}^{len \times d}$and the global representation $G \in \mathbb{R}^{3 \times d}$. As shown in Figure 2 and Equation 4.

$$\begin{aligned} Attention(G \rightarrow F_m) &= Cross\text{-}Attention(G \rightarrow F_m) \\ Attention(F_m \rightarrow G) &= Cross\text{-}Attention(F_m \rightarrow G) \end{aligned} \tag{4}$$

On the other hand, we utilize modality-specific Transformer encoder layers, denoted as L_s, to capture intra-modality information for each modality individually (using Equation 2). After encoding each modality, we use the self-attention mechanism in Transformer to process the unimodal feature sequences separately, in which the embedding at each position is able to learn the semantic and emotional information contained in the sequences.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

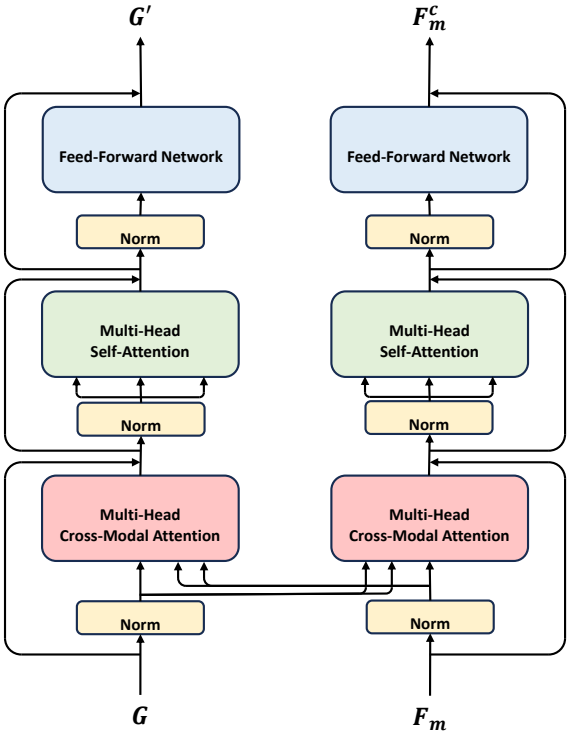

Figure 2. The structure of cross-attention. Cross-attention is used to capture dynamics between the global representation G and unimodal representations F_m.

The use of dual attention allows the model to process and analyze multimodal data at two different levels, Inter-modality and Intra-modality, for a more comprehensive understanding and interpretation of multimodal sentiment data.

3.2.3. Modality Fusion

After passing through the Cross-Attention stage, the model obtains Inter-Modality information, which is reflected through the global representation G'. While in the Self-Attention stage, to maintain consistency with the global representation, we employ Bi-LSTMs to process the three single-modal feature sequences individually, obtaining each unimodal representation. Meanwhile, in order to prevent the unimodal representation from being affected by the depth of the network, we propose the relevant coefficient, which is computed based on the relationship between the modal representation and the initial representation.

To be more specific, after learning Intra-Modality information in the Self-Attention stage, the model utilizes Bi-LSTMs to transform unimodal feature sequences into feature representations $U'_m \in \mathbb{R}^{b\times d}$, which are specific to each modality. Subsequently, we calculate relevant coefficients based on the correlation between this representation and the initial modal representations $U_m \in \mathbb{R}^{b\times d}$:

$$r_m = \sum(Diag(tanh(U'_m) \bigotimes tanh(U_m)) - 1)^2 \tag{5}$$

Where $\bigotimes$ denotes matrix multiplication, and $Diag(\cdot)$ represents all the diagonal elements of a square matrix. After obtaining the relevance coefficient r_m for each modality, we multiply it with U'_m to obtain the single-modal representation:

$$\widetilde{U}_m = r_m \times U'_m \tag{6}$$

Here, r_m is the relevance coefficient specific to each modality, and U'_m represents the feature representation of the corresponding modality obtained through Bi-LSTMs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

After obtaining the representations for both inter-modality and intra-modality $\{G', \widetilde{U}_l, \widetilde{U}_a, \widetilde{U}_v\}$, we concatenate the unimodal representations $\{\widetilde{U}_l, \widetilde{U}_a, \widetilde{U}_v\}$ with the global representation G' to create the representation for the sample. Finally, we employ several linear layers in combination with activation functions to make predictions for the ultimate result.

$$y = Pred(Concat(G', \widetilde{U}_l, \widetilde{U}_a, \widetilde{U}_v)) \tag{7}$$

3.3. Cross-Correlation Loss

Most of the current research uses attention mechanisms to capture relevant information from both intra-modality and inter-modality, but few scholars consider the relationship between these two different attention levels. In order to extract this relationship in dual attention, we propose a Cross-Correlation loss to obtain relevant information. By adding it to the objective function, the model is able to accomplish an undirected interaction between two different kinds of attention.

As shown in Figure 3, we use linear projectors to expand the feature sequence dimensions of the two different attention mechanisms and perform modality-specific matrix multiplication to obtain a set of matrices with a shape of $(batch, length, length)$.

$$C_m = F_m^S \bigotimes F_m^C \tag{8}$$

where C_m represents the cross-correlation matrix of the m modality's feature sequences in two different attention mechanisms, $m \in \{t, a, v\}$. The diagonal elements in this matrix represent the correlation between the corresponding positions of the two feature sequences, while the off-diagonal elements represent the redundant information.

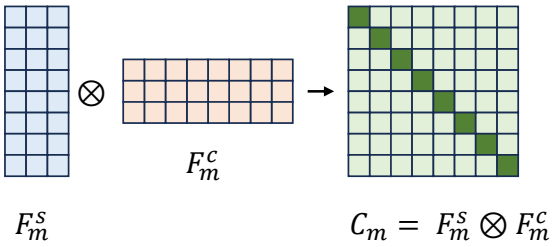

Figure 3. The Cross-Correlation matrix in Dual-Attention. We perform modality-specific matrix multiplication on the two types of unimodal feature sequences to obtain a cross-correlation matrix, and we use the diagonal elements of the matrix to represent the indirect interaction between these two feature sequences.

Our goal is to maximize the diagonal elements of the cross-correlation matrix (improving the correlation in Dual-Attention) while minimizing the off-diagonal elements (reducing the redundant information in Dual-Attention) during model training. As shown in Equation 9.

$$\mathcal{L}_{Corr} = \frac{\lambda}{M} \cdot \sum_{m}^{M} (\sum_{i=j}^{n} (c_{ij} - 1)^2 + \sum_{i \neq j}^{n} c_{ij}^2) \tag{9}$$

The term $\sum_{i=j}^{n} (c_{ij} - 1)^2$ in $\mathcal{L}_{Corr}$ is the correlation term, which denotes the correlation between the sequence of modality features of m in different attention mechanisms, and the other term $\sum_{i \neq j}^{n} c_{ij}^2$ is the redundancy term. Intuitively, the model increases the correlation between different attentional mechanisms by making the diagonal elements of the cross-correlation matrix close to 1. At the same time, it reduces the redundancy term by making the off-diagonal elements of the cross-correlation matrix close to 0.

Since the cross-correlation loss is calculated for all elements in the cross-correlation matrix, setting the weight of the cross-correlation loss too high in the objective function can cause the two attention mechanisms to lose their specificity and thus reduce the model performance. Therefore, we set a scaling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

factor λ in the cross-correlation loss according to the expansion of the feature sequence dimension. We conducted ablation experiments on different scaling weights on two datasets, as shown in Section 4.3.

4. Experiment

4.1. Preparations

4.1.1. Datasets

Multimodal dataset collects information from different modalities, such as text, speech, and vision, providing researchers with opportunities to gain a deeper understanding and analysis of sentiment expression. Three publicly available datasets are used in this article, including CMU-MOSI, CMU-MOSEI, and CH-SIMS.

CMU-MOSI [56](Multimodal Opinion Level Sentiment Intensity) is a multimodal dataset with character subjective sentiment and sentiment intensity annotations. It contains 2,199 multimodal samples from 93 YouTube videos, with each video ranging from 2-5 minutes and featuring 89 different speakers. Each video has been annotated with sentiment intensity, ranging from strong positive to strong negative on a scale from -3 to 3.

Another dataset is CMU-MOSEI [23](CMU Multimodal Opinion Sentiment and Emotion Intensity), an upgraded version of the CMU-MOSI dataset and one of the largest sentiment analysis datasets covering multiple fields, including sentiment recognition. CMU-MOSEI contains 23,453 manually annotated video clips from 5,000 videos on YouTube, including 1,000 different speakers and 250 different topics, covering almost all topics in daily life. CMU-MOSEI uses the same annotation method as CMU-MOSI.

In addition, considering the research on multimodal sentiment analysis in the Chinese community, we also used CH-SIMS [57], a refined Chinese multimodal dataset. It contains 2,281 samples from 60 videos collected from movies, TV shows, and variety shows. Compared to the first two datasets, it not only includes multimodal sentiment labels but also provides independent fine-grained single-modality sentiment labels for each sample. Each label in this dataset is manually annotated from -1 (strongly positive) to 1 (strongly negative). The statistical information of these three datasets is shown in Table 2.

Table 2. Statistics of MOSI, MOSEI, and SIMS datasets.

Dataset	Train	Validation	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

4.1.2. Data Processing

We targeted the different modalities for processing. For the text modality, we used the BERT-based-uncased model to encode the CMU-MOSI and CMU-MOSEI datasets. In addition, for the Chinese multimodal sentiment dataset CH-SIMS, we used the BERT-based-Chinese model for text encoding. This step helps to transform text data into vector representations with rich semantic information.

When processing the speech modality, we used the COVAREP tool to extract audio features, including pitch, glottal source parameters, and 12 Mel-frequency cepstral coefficients (MFCCs). These features capture sound frequencies, voice source properties, and acoustic features in speech, providing important information for sentiment analysis. For the CH-SIMS dataset, we used the Librosa toolkit in Python to extract speech features such as log fundamental frequency, constant-Q chromatograms, and 20 MFCCs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

For visual modality, we used the Facet tool to extract 35 facial features for the CMU-MOSI and CMU-MOSEI datasets, which record facial muscle movements related to sentiment. For the Chinese sentiment dataset CH-SIMS, we used the OpenFace 2.0 toolkit to extract 17 facial action units, 68 facial landmarks, and some features related to head posture and eye movements. These facial features capture information related to facial expressions in sentiment expression, providing important visual data for multimodal sentiment analysis.

4.1.3. Baseline

In the field of multimodal sentiment analysis, there exists a series of different baseline models, each with its own characteristics. In order to comprehensively verify the performance of the method proposed in this paper, we compared it with many current methods, which mainly include:

TFN [33]. Tensor Fusion Network is a tensor fusion-based method that computes the triple Cartesian product between three modalities to explicitly capture intra-modality and inter-modality dynamic information. It utilizes tensor operations to capture the interaction and fusion of multimodal information.

LMF [34]. Similar to TFN, Low-rank Multimodal Fusion also relies on tensor operations, but it cleverly uses modality-specific low-rank factors to more efficiently compute multimodal representations, improving fusion efficiency while ensuring information quality.

MulT [29]. Multimodal Transformer adopts a bidirectional cross-modal attention mechanism to calculate the relation between two different modalities separately. The method is based on Transformer architecture, which can better capture dynamic information between different modalities.

MISA [58]. Modality-invariant and-specific Representations for Multimodal Sentiment Analysis. MISA uses a subspace learning approach to map each modality to two different subspaces for learning, providing a comprehensive view of multimodal representation learning and achieving better fusion results.

Self-MM [21]. Self-Supervised Multi-task Multimodal sentiment analysis network designs an unimodal label generation module based on self-supervised learning to obtain independent unimodal representations. It utilizes self-supervised learning to improve model performance. Also, it jointly trains multimodal and unimodal tasks to learn modal consistency and variability.

AMML [59]. Adaptive Multimodal Meta-learning uses a meta-learning approach to train unimodal networks and applies them to multimodal inference. This method focuses on network adaptability and optimizes unimodal representations through adaptive learning rate adjustment for better multimodal fusion.

MMIM [40]. MultiModal InfoMax proposes a hierarchical maximization of mutual information framework, which improves the consistency and information density of multimodal representations by maximizing mutual information and preserves task-relevant information through multimodal fusion.

EMT [32]. Efficient Multimodal Transformer proposes an efficient network based on the Transformer architecture for integrating multimodal information. This network utilizes unimodal encoders to obtain multimodal representations and enables mutual learning between multimodal global representations and unimodal feature sequences.

4.1.4. Hyper-Parameter Setting

We use the Pytorch in deep learning to build our model and optimize it with the Adam optimizer, and we adopt an early-stop strategy. Table 3 shows the parameter settings for CCDA trained on CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets. In the Cross-Attention section, we adopt the same hyperparameter settings as EMT, and in the Self-Attention section, we use the Transformer parameter settings in MulT. To reflect the accuracy of the results, we conducted five experiments and averaged each metric in the experimental results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

Table 3. CCDA hyperparameter settings on three datasets.

Hyper-Parameter	CMU-MOSI	CMU-MOSEI	CH-SIMS
Batch size	32	16	32
Early stop(epochs)	16	8	16
Learning rate	1e-3	1e-4	1e-3
Optimizer	Adam	Adam	Adam
Dimension of feature and representation	128	128	128
Transformer layers in Cross-Attention	3	2	4
Cross-Attention heads	4	4	4
Transformer layers in Self-Attention	2	2	2
Attention dropout	0.1	0.1	0.1
Stacked LSTM layers for Self-Attention	2	2	2
Stacked LSTM dropout	0.1	0.1	0.1
λ in Cross-Correlation Loss	5e-5	5e-5	1e-3
Projector dims in Cross-Correlation Loss	1024	1024	256

4.2. Result Analysis

4.2.1. Evaluation Metrics

In regression tasks, we mainly use two metrics to measure model performance: Mean Absolute Error (MAE) and Person Correlation Coefficient (Corr). MAE is used to measure the average absolute error between the model's predicted values and the true labels, with lower values indicating better model performance. Corr is used to measure the correlation between the model's predicted results and the true labels, with values closer to 1 indicating better model performance. Additionally, we also convert the model's output results into classification task metrics, including Acc-k and F1-score. Acc-2, Acc-5, and Acc-7 on the CMU-MOSI and CMU-MOSEI datasets and Acc-2, Acc-3, and Acc-5 on CH-SIMS are used to evaluate the model's accuracy in multi-classification tasks, with larger values indicating better model performance. F1-score represents the harmonic mean of precision and recall and is used to evaluate the balance between positive and negative categories. Higher F1 indicates better model performance in classification tasks.

4.2.2. Quantitative analysis

The experimental data for TFN, LMF, MulT, MISA, Self-MM, and MMIM comes from [40]. For the other models, we conducted five experiments on each of the three datasets using publicly available source code and averaged the experimental results for each model. In all evaluation metrics, except for MAE, larger values indicate better model performance. The experimental results are compared in Tables 4–6.

Table 4 shows the model's results on the CMU-MOSI dataset. Compared to the EMT model, CCDA improved by 0.009 on the regression metrics MAE and Corr. In terms of classification task metrics, CCDA improved by 0.6% on Acc-2 and Acc-5 and 0.7% on Acc-7 and achieved a 0.6% improvement in F1 score over the best model. Similarly, as shown in Table 5, CCDA's performance on CMU-MOSEI improved by 0.003 on MAE, 0.006 on Corr, 0.5% on Acc-7, 0.4% on Acc-5, 0.6% on Acc-2, and 0.7% on F1 compared to EMT. Table 6 shows the experimental results of the model on CH-SMIS, where CCDA achieved better results on some metrics, such as 0.006 on MAE, 0.005 on Corr, 1.4% on Acc-3, 1.2% on Acc-2, and 0.9% on F1. However, its performance on the 5-classification task was slightly worse than that of the EMT model. We believe that while CCDA improves coarse-grained sentiment classification, but it does not improve much for fine-grained classification.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

Table 4. Experiments on CMU-MOSI

Models	MAE(↓)	Corr(↑)	Acc-7(↑)	Acc-5(↑)	Acc-2(↑)	F1(↑)
TFN [33]	0.901	0.698	34.9	-	80.8	80.7
LMF [34]	0.917	0.695	33.2	-	82.5	82.4
MulT [29]	0.846	0.725	40.4	46.7	83.4	83.5
MISA [58]	0.804	0.764	-	-	82.1	82.0
Self-MM [21]	0.717	0.793	46.4	52.8	84.6	84.6
MMIM [40]	0.712	0.790	46.9	53.0	85.3	85.4
AMML [59]	0.723	0.792	46.3	-	84.9	84.8
EMT [32]	0.705	0.798	47.4	54.1	85.0	85.0
Ours	**0.696**	**0.807**	**48.0**	**54.8**	**85.7**	**85.6**

Table 5. Experiments on CMU-MOSEI

Models	MAE(↓)	Corr(↑)	Acc-7(↑)	Acc-5(↑)	Acc-2(↑)	F1(↑)
TFN [33]	0.593	0.700	50.2	-	82.5	82.1
LMF [34]	0.623	0.677	48.0	-	82.0	82.1
MulT [29]	0.564	0.731	52.6	54.1	83.5	83.6
MISA [58]	0.568	0.724	-	-	84.2	84.0
Self-MM [21]	0.533	0.766	53.6	55.4	85.0	85.0
MMIM [40]	0.536	0.764	53.2	55.0	85.0	85.1
AMML [59]	0.614	0.776	52.4	-	85.3	85.2
EMT [32]	0.527	0.774	54.5	56.3	86.0	86.0
Ours	**0.524**	**0.780**	**55.0**	**56.7**	**86.6**	**86.7**

Table 6. Experiments on CH-SIMS

Models	MAE(↓)	Corr(↑)	Acc-5(↑)	Acc-3(↑)	Acc-2(↑)	F1(↑)
TFN [33]	0.437	0.582	-	-	77.1	76.9
LMF [34]	0.438	0.578	-	-	77.4	77.4
MulT [29]	0.442	0.581	40.0	65.7	78.2	78.5
MISA [58]	0.447	0.563	-	-	76.5	76.6
Self-MM [21]	0.411	0.601	43.1	66.1	78.6	78.6
MMIM [40]	0.422	0.597	42.0	65.5	78.3	78.2
AMML [59]	0.437	0.583	41.2	64.2	78.0	78.1
EMT [32]	0.396	0.623	**43.5**	67.4	80.1	80.1
Ours	**0.393**	**0.628**	43.3	**68.3**	**81.1**	**81.0**

4.3. Ablation Study

To validate the multimodal fusion strategy used in the CCDA model and the impact of cross-correlation loss on model performance, we conducted ablation experiments on two datasets, CMU-MOSI and CH-SIMS. First, we examined the influence of relevant coefficients in the multimodal fusion strategy. Secondly, we verified that the cross-correlation loss helps the model capture the correlations in the dual attention during the training process.

4.3.1. Multi-Modal Fusion Strategy

Before performing multi-modal fusion in the model, we adjusted the unimodal representations based on the relevant coefficients computed between unimodal representations and their respective initial modality representations. Subsequently, these representations were concatenated with the global multimodal representation. To validate the effectiveness of our proposed fusion method, we conducted experiments on both Chinese and English datasets. We compared the performance of models with and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

without considering unimodal relevant coefficients, where the unimodal representations, computed after self-attention and subsequent Bi-LSTMs, were directly concatenated with the global multimodal representation, and then fed into the fusion and prediction module. We also compared these results with the standard version of CCDA. The comparative experimental results are shown in Tables 7 and 8.

Table 7. Impact of Correlation Coefficients in Fusion Strategy on CMU-MOSI

	MAE(↓)	Corr(↑)	Acc-7(↑)	Acc-5(↑)	Acc-2(↑)	F1(↑)
Direct Concat	0.713	0.790	46.5	53.8	85.2	85.2
CCDA	**0.696**	**0.807**	**48.0**	**54.8**	**85.7**	**85.6**

Table 8. Impact of Correlation Coefficients in Fusion Strategy on CH-SIMS

	MAE(↓)	Corr(↑)	Acc-5(↑)	Acc-3(↑)	Acc-2(↑)	F1(↑)
Direct Concat	0.408	0.614	41.2	66.4	80.4	80.4
CCDA	**0.393**	**0.628**	**43.3**	**68.3**	**81.1**	**81.0**

According to Tables 7 and 8, it is evident that in multimodal fusion, the model's performance significantly improves when unimodal features are augmented with relevant coefficients compared to direct concatenation. Specifically, there is a 1.5% improvement in Acc-7, indicating that CCDA achieves higher accuracy in multi-class classification.

4.3.2. Cross-Correlation Loss Function

Additionally, this study assumes a certain degree of cross-correlation between self-attention and cross-attention. Thus, we introduced a cross-correlation loss function to facilitate indirect interaction between these two attention mechanisms. To assess the impact of cross-correlation loss on model performance, we conducted ablation experiments on the CMU-MOSI and CH-SIMS datasets, as shown in Tables 9 and 10.

Table 9. Impact of Cross-Correlation Loss in the Objective Function on CMU-MOSI

	MAE(↓)	Corr(↑)	Acc-7(↑)	Acc-5(↑)	Acc-2(↑)	F1(↑)
w/o corr loss	0.708	0.795	47.4	54.2	84.9	84.9
CCDA	**0.696**	**0.807**	**48.0**	**54.8**	**85.7**	**85.6**

Table 10. Impact of Cross-Correlation Loss in the Objective Function on CH-SIMS

	MAE(↓)	Corr(↑)	Acc-5(↑)	Acc-3(↑)	Acc-2(↑)	F1(↑)
w/o corr loss	0.400	0.610	42.0	66.7	80.1	80.1
CCDA	**0.393**	**0.628**	**43.3**	**68.3**	**81.1**	**81.0**

It can be observed that adding cross-correlation loss to the objective function significantly enhances the model's performance. This improvement is particularly pronounced in multi-class tasks, indicating that cross-correlation loss has a substantial impact on model performance in multi-modal sentiment analysis. Further analysis reveals that cross-correlation loss establishes a closer connection between self-attention and cross-attention in the model, enabling better integration of information from multimodal data. This indirect interaction helps the model better understand the relationships between different modalities, thereby improving overall sentiment analysis performance. In multimodal sentiment analysis tasks, such enhanced connectivity is highly beneficial. Moreover, the results on different datasets demonstrate the universality of the improvement brought by cross-correlation loss,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 December 2023 doi:10.20944/preprints202312.0472.v1

indicating that it is not limited to specific datasets. This strengthens the scalability and generality of our approach.

4.3.3. Scaling Factor in Cross-Correlation Loss

When calculating the cross-correlation loss, the model expands the dimensions of the feature sequences. As a result, the values of elements in the correlation matrix become relatively large. To balance the cross-correlation loss in the objective function, we introduced scaling factors. Figure 4 illustrates the impact of scaling factors on the final results. Since we set different feature dimensions for unimodal features from different datasets (128 for CMU-MOSI and CMU-MOSEI, 32 for CH-SIMS), and applied different linear mapping layers for dimension expansion when calculating the cross-correlation loss for different datasets, the optimal scaling factors also vary. Specifically, we used 5e-5 for CMU-MOSI and 1e-3 for CH-SIMS.

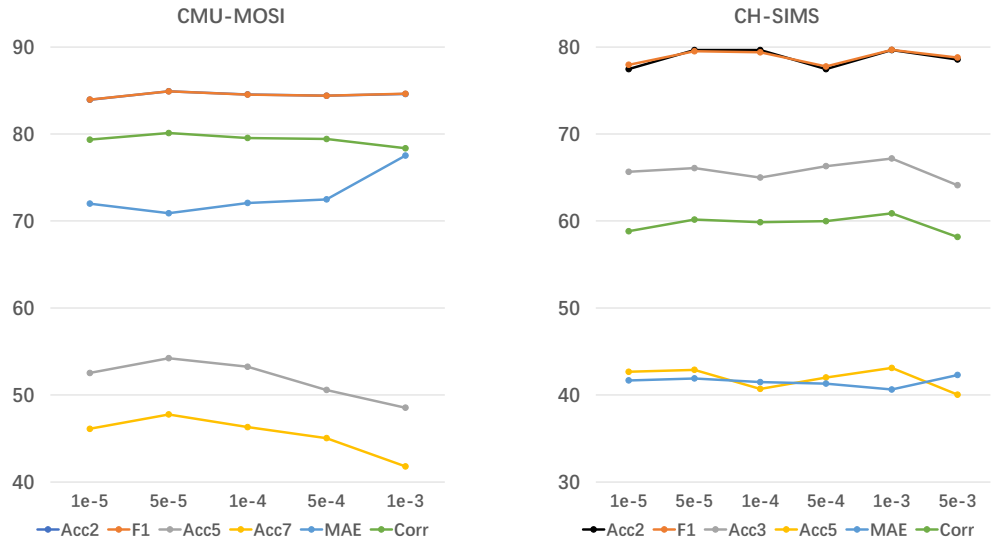

Figure 4. Impact of Scaling Weights in Cross-Correlation Loss

5. Conclusion

In this paper, we introduced the Cross-Correlation In Dual Attention (CCDA) model aimed at fusing multimodal features and perceiving human sentiment analysis. We enriched the self-attention mechanism by incorporating relevant coefficients on unimodal features, enabling them to partially attend to the information from the initial modality features during the self-attention level. Additionally, to effectively handle potential inter-modal correlations between dual-attention mechanisms, We innovatively propose the cross-correlation loss function. By adding the interrelation loss to the objective function and minimizing the objective function, we accomplish indirect interaction between dual attention.

We conducted comprehensive experiments on three commonly used public datasets in the multi-modal sentiment analysis domain, including CMU-MOSI, CMU-MOSEI, and CH-SIMS, which is the most comprehensive multi-modal sentiment analysis dataset in the Chinese community. We compared the CCDA model with baseline models and found that our model demonstrated a significant advantage on all three datasets. Through experimentation, we demonstrated the strong performance of the CCDA model in multi-modal sentiment analysis tasks, offering new insights for further research and applications in this field.

Given the challenges faced in real-world multi-modal sentiment analysis, especially in scenarios involving missing modal information or noisy data, future research could focus on enhancing the model's robustness and accuracy in handling modal information gaps, noise filtering, and correction.

This would ensure the effectiveness and reliability of the model in a wider range of practical applications.

Author Contributions: Conceptualization, P.W.; methodology, P.W.; software, P.W.; validation, P.W., S.L. and J.C.; formal analysis, P.W.; investigation, P.W.; resources, P.W.; data curation, P.W.; writing—original draft preparation, P.W.; writing—review and editing, P.W.; visualization, P.W.; supervision, P.W.; project administration, P.W.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61762085), And the Natural Science Foundation of Xinjiang Uygur Autonomous Region Project (2019D01C081).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in this study is available on request from the corresponding author.

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

References

1. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* **2010**, *16*, 345–379.
2. 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. *Information Fusion* **2023**, *91*, 424–444.
3. Somandepalli, K.; Guha, T.; Martinez, V.R.; Kumar, N.; Adam, H.; Narayanan, S. Computational media intelligence: Human-centered machine analysis of media. *Proceedings of the IEEE* **2021**, *109*, 891–910.
4. Stappen, L.; Baird, A.; Schumann, L.; Bjorn, S. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing* **2021**.
5. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A survey of multimodal sentiment analysis. *Image and Vision Computing* **2017**, *65*, 3–14. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.
6. Poria, S.; Hazarika, D.; Majumder, N.; Mihalcea, R. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing* **2020**.
7. Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A. Affective computing and sentiment analysis. *A practical guide to sentiment analysis* **2017**, pp. 1–10.
8. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* **2017**, *37*, 98–125.
9. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169–176.
10. Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; Morency, L.P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* **2013**, *28*, 46–53.
11. Poria, S.; Cambria, E.; Hussain, A.; Huang, G.B. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks* **2015**, *63*, 104–116.
12. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016, pp. 439–448.
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 Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 163–171.
14. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2018, Vol. 32.
15. Glodek, M.; Reuter, S.; Schels, M.; Dietmayer, K.; Schwenker, F. Kalman filter based classifier fusion for affective state recognition. In Proceedings of the Multiple Classifier Systems: 11th International Workshop, MCS 2013, Nanjing, China, May 15-17, 2013. Proceedings 11. Springer, 2013, pp. 85–94.

16. Cai, G.; Xia, B. Convolutional neural networks for multimedia sentiment analysis. In Proceedings of the Natural Language Processing and Chinese Computing: 4th CCF Conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings 4. Springer, 2015, pp. 159–167.
17. Alam, F.; Riccardi, G. Predicting personality traits using multimodal information. In Proceedings of the Proceedings of the 2014 ACM multi media on workshop on computational personality recognition, 2014, pp. 15–18.
18. Wang, H.; Meghawat, A.; Morency, L.P.; Xing, E.P. Select-additive learning: Improving generalization in multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017, pp. 949–954.
19. Poria, S.; Cambria, E.; Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2539–2544.
20. 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). IEEE, 2020, pp. 4477–4481.
21. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35, pp. 10790–10797.
22. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
23. 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 Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
24. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.P. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920* **2018**.
25. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, pp. 7216–7223.
26. 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 Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 873–883.
27. 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 Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 163–171.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
29. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, Vol. 2019, p. 6558.
30. Arjmand, M.; Dousti, M.J.; Moradi, H. Teasel: a transformer-based speech-prefixed language model. *arXiv preprint arXiv:2109.05522* **2021**.
31. Cheng, H.; Yang, Z.; Zhang, X.; Yang, Y. Multimodal Sentiment Analysis Based on Attentional Temporal Convolutional Network and Multi-layer Feature Fusion. *IEEE Transactions on Affective Computing* **2023**.
32. Sun, L.; Lian, Z.; Liu, B.; Tao, J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing* **2023**.
33. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* **2017**.
34. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* **2018**.
35. Barezi, E.J.; Fung, P. Modality-based factorization for multimodal fusion. *arXiv preprint arXiv:1811.12624* **2018**.

36. Luo, H.; Ji, L.; Huang, Y.; Wang, B.; Ji, S.; Li, T. Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. *arXiv preprint arXiv:2112.01368* **2021**.
37. Lin, Z.; Liang, B.; Long, Y.; Dang, Y.; Yang, M.; Zhang, M.; Xu, R. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In Proceedings of the Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 7124–7135.
38. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256* **2022**.
39. Pérez-Rosas, V.; Mihalcea, R.; Morency, L.P. Utterance-level multimodal sentiment analysis. In Proceedings of the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 973–982.
40. Han, W.; Chen, H.; Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412* **2021**.
41. Zhang, S.; Li, B.; Yin, C. Cross-Modal Sentiment Sensing with Visual-Augmented Representation and Diverse Decision Fusion. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22010074.
42. Liang, P.P.; Liu, Z.; Tsai, Y.H.H.; Zhao, Q.; Salakhutdinov, R.; Morency, L.P. Learning representations from imperfect time series data via tensor rank regularization. *arXiv preprint arXiv:1907.01011* **2019**.
43. Mai, S.; Hu, H.; Xing, S. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In Proceedings of the Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 481–492.
44. 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. International Joint Conferences on Artificial Intelligence Organization, 2019.
45. Jin, T.; Huang, S.; Li, Y.; Zhang, Z. Dual low-rank multimodal fusion. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 377–387.
46. Verma, S.; Wang, J.; Ge, Z.; Shen, R.; Jin, F.; Wang, Y.; Chen, F.; Liu, W. Deep-HOSeq: Deep higher order sequence fusion for multimodal sentiment analysis. In Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM). IEEE, 2020, pp. 561–570.
47. Lian, Z.; Tao, J.; Liu, B.; Huang, J. Conversational emotion analysis via attention mechanisms. *arXiv preprint arXiv:1910.11263* **2019**.
48. Xiao, L.; Wu, X.; Wu, W.; Yang, J.; He, L. Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 4578–4582.
49. Chen, Q.; Huang, G.; Wang, Y. The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2022**, *30*, 2689–2695.
50. Fu, Z.; Liu, F.; Xu, Q.; Qi, J.; Fu, X.; Zhou, A.; Li, Z. NHFNET: A non-homogeneous fusion network for multimodal sentiment analysis. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022, pp. 1–6.
51. Wang, H.; Li, X.; Ren, Z.; Wang, M.; Ma, C. Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23052679.
52. Hou, S.; Tuerhong, G.; Wushouer, M. UsbVisdaNet: User Behavior Visual Distillation and Attention Network for Multimodal Sentiment Classification. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23104829.
53. Wang, F.; Tian, S.; Yu, L.; Liu, J.; Wang, J.; Li, K.; Wang, Y. TEDT: Transformer-Based Encoding–Decoding Translation Network for Multimodal Sentiment Analysis. *Cognitive Computation* **2023**, *15*, 289–303.
54. Tang, J.; Liu, D.; Jin, X.; Peng, Y.; Zhao, Q.; Ding, Y.; Kong, W. Bafn: Bi-direction attention based fusion network for multimodal sentiment analysis. *IEEE Transactions on Circuits and Systems for Video Technology* **2022**, *33*, 1966–1978.
55. Wu, Y.; Zhao, Y.; Yang, H.; Chen, S.; Qin, B.; Cao, X.; Zhao, W. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. *arXiv preprint arXiv:2203.00257* **2022**.
56. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* **2016**.

57. Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; Yang, K. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Proceedings of the Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 3718–3727.
58. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1122–1131.
59. Sun, Y.; Mai, S.; Hu, H. Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Transactions on Affective Computing* **2022**.