Information Fusion 99 (2023) 101891

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis

Zuhe Li [a], Qingbing Guo [a], Yushan Pan [b,*], Weiping Ding [c,*], Jun Yu [a], Yazhou Zhang [d], Weihua Liu [e], Haoran Chen [a], Hao Wang [f], Ying Xie [g]

[a] School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
[b] Department of Computing, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
[c] School of Information Science and Technology, Nantong University, Nantong, 226019, China
[d] College of Software Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China
[e] China Mobile Research Institute, Beijing, 100053, China
[f] Xidian University, Xi'an, 710071, China
[g] Putian University, Putian, 351100, China

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Unimodal feature fusion
Linguistic-guided transformer
Self-supervised label generation

ABSTRACT

Fusion and co-learning are major challenges in multimodal sentiment analysis. Most existing methods either ignore the basic relationships among modalities or fail to maximize their potential correlations. They also do not leverage the knowledge from resource-rich modalities in the analysis of resource-poor modalities. To address these challenges, we propose a multimodal sentiment analysis method based on multilevel correlation mining and self-supervised multi-task learning. First, we propose a unimodal feature fusion- and linguistics-guided Transformer-based framework, multi-level correlation mining framework, to overcome the difficulty of multimodal information fusion. The module exploits the correlation information between modalities from low to high levels. Second, we divided the multimodal sentiment analysis task into one multimodal task and three unimodal tasks (linguistic, acoustic, and visual tasks), and designed a self-supervised label generation module (SLGM) to generate sentiment labels for unimodal tasks. SLGM-based multi-task learning overcomes the lack of unimodal labels in co-learning. Through extensive experiments on the CMU-MOSI and CMU-MOSEI datasets, we demonstrated the superiority of the proposed multi-level correlation mining framework to state-of-the-art methods.

1. Introduction

Sentiment analysis is the process of determining opinions and attitudes from various sources systematically. With the increased popularity of multimodal intelligence, sentiment analysis has evolved from conventional unimodal analysis to more complex forms of multimodal analysis [1]. Sentiment cannot be accurately analyzed using only linguistic information; consequently, multimodal sentiment analysis focuses on generalizing text-based sentiment analysis to videos that contain linguistic, acoustic, visual, or other multimodal information [2]. For example, in sarcasm sentiment analysis, satirical information can be recognized more accurately after multimodal features are extracted from both acoustic and visual modalities [1]. Therefore, multimodal sentiment analysis has achieved salient improvements when dealing with complex multimodal data [3] and has received increasing attention [4].

Despite the extensive research on multimodal sentiment analysis, some challenges still need to be overcome. Fusion is one critical challenge [5,6]. Existing multimodal fusion methods can either be model-agnostic or model-based [7,8]. Model-agnostic approaches are mainly basic fusion methods, such as early and late fusion, which directly perform fusion operations on feature vectors or decision results. For example, low-rank multimodal fusion (LMF) [9] was used to obtain the correlations and interactions between low-level features. Although model-agnostic approaches are applicable to most multimodal fusion problems, low-level features may not fully reveal the deep correlations among multimodal data.

Model-based approaches primarily, which are mainly machine and deep learning-based multimodal fusion methods, focus on determining the deep associations between multimodal representations. For example, long short-term memory (LSTM) has been used to introduce time information into multimodal sentiment analysis [10], and the

* Corresponding authors.
E-mail addresses: Yushan.Pan@xjtlu.edu.cn (Y. Pan), ding.wp@ntu.edu.cn (W. Ding).

https://doi.org/10.1016/j.inffus.2023.101891
Received 17 March 2023; Received in revised form 11 June 2023; Accepted 12 June 2023
Available online 16 June 2023

Transformer framework has been employed to fuse multiple unimodal representations [11]. Although model-based approaches are superior, they lack interpretability and typically require large training datasets. Therefore, to address the fusion problem in the field of multimodal sentiment analysis, both the basic relationships among low-level features and the potential correlations among high-level features must be considered during feature fusion.

Another challenge in multimodal sentiment analysis is co-learning, which involves leveraging knowledge from another resource-rich modality in modeling a resource-poor modality [7]. In multimodal sentiment-analysis tasks, text, voice, and video data are obtained simultaneously from the same speaker. When the information from one modality is missing, the challenge of co-learning is to determine and use the correlation between unimodal representations [12]. Some researchers have attempted to solve this problem using multitask learning methods [13] by conducting knowledge transfer through hidden-layer parameter sharing during the joint training process of multiple unimodal tasks. Although these methods can reveal the association among unimodal tasks, independent unimodal human annotations must be introduced to complete the training tasks. Therefore, the key to solving the co-learning problem in multimodal sentiment analysis is to find ways to leverage existing information for joint multi-task training, even in the absence of unimodal independent labels.

In this study, we focused on simultaneously addressing the challenges of fusion and co-learning. Motivated by model-based fusion approaches and multitask learning with independent unimodal annotations, we adopted a multitask learning framework as the backbone of the proposed sentiment analysis model, which comprised one multimodal task and three unimodal tasks. Building on this foundation, we propose a novel multimodal sentiment analysis model that incorporates a multilevel correlation mining framework (MCMF) and a self-supervised label generation module (SLGM). Experiments revealed that the hybrid framework MCMF implements both feature fusion and association mining from different perspectives, and the SLGM module facilitates multi-task learning-based multimodal sentiment analysis in the absence of unimodal labels. The contributions of our study can be summarized as follows:

- We prioritized the relationship between representations and features and utilized unimodal features fusion (UFF) to obtain correlation information from low-level features. This bottom-up approach improved the fusion effect by leveraging the basic relationship between features.
- We prioritized the association among modalities and proposed a linguistics-guided Transformer (LGT) method to extract correlation information from high-level features. This top-down scheme was adopted to overcome the complexity of multimodal information fusion by leveraging the potential correlation among modalities.
- We utilized a multi-task learning framework to jointly train the four tasks for co-learning. To address the lack of unimodal labels for most datasets in the field of multimodal sentiment analysis, we introduced a SLGM that generated unimodal labels based on the relationship between modality representations and labels.

The remainder of this paper is organized as follows. In Section 2, we review related work in the field of multimodal sentiment analysis. In Section 3, we introduce the overall architecture of the proposed model and core concepts of each component. In Section 4, we describe the experiment set-up. In Section 5, we present the performance of the proposed model in experiments. In Section 6, we summarize the study and discuss potential future directions.

2. Related work

Researchers have proposed numerous methods and models for multimodal sentiment analysis. These methods can be tensor-based, deep-learning-based, or multitask learning-based.

Tensor-based methods were among the early approaches. One of the most representative methods is the tensor fusion network (TFN) proposed by Zadeh et al. [14]. It uses a three-fold Cartesian product to fuse multiple unimodal representations as an alternative to simple tensor splicing. Liu et al. [9] proposed an LMF method that uses low-rank tensors to improve the efficiency of multimodal fusion. Sahay et al. [15] proposed a relational tensor network architecture that applies tensor fusion to each video clip at the time level. However, these methods, which utilize low-level features to analyze sentiment at minimal cost, often ignore contextual information [16], resulting in poor performance in dialogue situations.

Deep learning-based methods mainly focus on sentiment analysis in dialogue scenarios, which involve more complex sentiment interactions. Capturing sentiment associations between people is a challenge. Hazarika et al. proposed an interactive conversational memory network [17] that obtains contextual summaries using global memories for multimodal sentiment detection. Zhang et al. proposed a quantum-inspired interactive network [18], a combination of quantum theory and an LSTM network, to learn intra- and inter-utterance interaction dynamics. Ghosal et al. proposed a COSMIC framework [19] that uses elements of commonsense to learn the interactions between speakers.

In addition, new deep learning technologies are becoming popular among researchers owing to their excellent performance. The most representative technology is the Transformer used for machine translation. It is a sequence-to-sequence architecture based solely on attention mechanisms that eliminate recurrent and convolutional structures [20]. The Transformer associates each element in the sequence with other elements when modeling sequential data and mining contexts and has superior accuracy, stability, and speed. Owing to its unique advantages, an increasing number of researchers are attempting to apply it in other fields. Some researchers have used this approach to explore the correlations between unimodal representations. Tsai et al. [21] used a multimodal Transformer (MulT) to obtain the interactions between multimodal sequences across distinct time steps. Transformer can also be used to integrate unimodal features for sentiment analysis. For example, Delbrouck et al. [11] utilized a Transformer framework to fuse multiple unimodal representations for multimodal sentiment analysis. Rahman et al. [22] adopted large pretrained Transformers to integrate multimodal information. Recently, Wang et al. [23] proposed a Transformer-based multimodal encoding–decoding translation network, taking textual content as the primary information through a joint encoding–decoding method. To reduce the impact of personalized speech and visual features, Wang et al. [24] proposed a speaker-independent multimodal representation (SIMR) framework that divides nonverbal inputs into style encoding and content representation, and attempted to locate compatible and incompatible cross-modal interactions simultaneously through a Cross-modal Transformer module. Kim et al. [25] presented all-modalities-in-one Bidirectional Encoder Representations from Transformers (AOBERT), a single-stream Transformer pre-trained on two tasks simultaneously to determine the dependency and relationship between modalities. These studies indicate that Transformers model remains a promising foundational framework in multimodal sentiment analysis.

Multi-task learning improves the generalization performance of related tasks by jointly training multiple tasks and obtaining additional information [26]. In the context of multimodal sentiment analysis, multitask learning-based methods decompose tasks into unimodal tasks and use a multitask learning framework to train them jointly, leading to more robust models and reduced overfitting for better performance [27]. For example, Yu et al. [13] incorporated independent unimodal human annotations and used multitask learning to learn multimodal and unimodal representations simultaneously. However, the approach requires time-consuming independent and unimodal labeling. To address this issue, Yu et al. [28] proposed a unimodal label generation module (ULGM) based on multimodal labels and modality representations, which can automatically generate unimodal labels for

datasets that lack them. However, automatically generated labels may not always accurately reflect the actual situation.

Most of the above methods focus on only one aspect of the problem of multimodal sentiment analysis without comprehensively considering the representations, features, and associations among different modalities.

(1) Tensor-based methods only deal with the characteristics of the three modalities and ignore the correlation between different modalities.
(2) Deep learning-based methods only focus on high-level modality correlation without fully leveraging the correlation information in low-level features.
(3) Multi-task learning-based methods rely excessively on unimodal labels and cannot adapt to most datasets without unimodal annotations.

To address these problems, we combined the strengths of the aforementioned methods and proposed a novel approach for multimodal sentiment analysis. We used the SLGM to address the problem of unimodal labeling by employing a multitask learning model as the main framework. Additionally, we utilized the MCMF to simultaneously extract the fundamental relationships and potential connections among the modalities. This comprehensive approach enabled us to capture the nuances and complexities of multimodal sentiment analysis and achieve better performance.

3. Methodology

3.1. Overview of the MCMF model

The structure of the proposed model is illustrated in Fig. 1. The input data were fed into the representation learning module, where three unimodal tensors (T^l, T^a, and T^v) containing features and timing information were obtained for each modality. These tensors were then inputted into the MCMF, which was composed of UFF and LGT modules. The UFF module fused the unimodal tensors of the three modalities into a multimodal feature tensor ($F_{(m)}$) using a three-fold Cartesian product that learned the association between low-level features. The LGT module enhanced the correlation between unimodal representations using the linguistic tensor (T^l) as query (Q) vectors to compute the cross-modal multi-head attention score, which was then used to calculate the weighted unimodal features ($F_{(l)}$, $F_{(a)}$, and $F_{(v)}$).

The multimodal sentiment analysis task was divided into one multimodal task (Task m) and three unimodal tasks (Task l, Task a, and Task v), which were jointly trained in the multitask learning framework. The bottom-representation learning network was shared among different tasks using a hard-sharing strategy. To generate unimodal labels, the proposed SLGM module utilized multimodal labels to generate unimodal labels. However, unimodal tasks were only used to assist multimodal tasks, and the output of the multimodal task was taken as the final prediction result.

3.2. Representation learning

In this section, we describe the representation learning module, which comprises two parts, namely unimodal feature extraction and timing information extraction.

Linguistic Features: Traditional linguistic representation methods cannot capture the context representation information to model the phenomenon of polysemy. To solve this problem, we used a pretrained BERT model [29] to extract the linguistic features. The BERT model comprised 12 Transformers. Each layer contained a 768-dimensional hidden layer and a multi-head attention (MHA) with 12 heads. The model provided a better representation by capturing a bidirectional context and finally generating 768-dimensional linguistic features.

Acoustic Features: For acoustic features, we paid more attention to the unique information of acoustic data such as intonation. In line with Li et al. [30], frequency spectrum characteristics, such as mel-frequency cepstral coefficients and constant-Q chromatograms, were taken as acoustic features because of their proven relation to the speaker's sentiment.

Visual Features: For visual data, the facial expression is suffused with sentiment information, and is, therefore, the most important. We used the Openface 2.0 toolkit [31] to recognize facial information in the visual data. The toolkit can extract a series of facial expression features such as facial movement, head direction, and eye direction.

Timing Information: The BERT model can process timing information; therefore, the extracted linguistic features include context information themselves. However, acoustic and visual features do not contain timing information; therefore, they are passed to a bidirectional LSTM (BiLSTM) network [32] to obtain context information. Finally, 16-dimensional acoustic features and 32-dimensional visual features were generated.

3.3. Multi-level correlation mining module for multimodal representations

Following feature extraction, we adopted a multilevel correlation mining module, which comprised a UFF module and an LGT, to obtain multimodal representations.

UFF: For multimodal tasks, we used model-agnostic approaches to fuse features to discover the low-level feature association between multimodal representations. The UFF module overcomes the disadvantages of traditional fusion methods, such as multimodal concatenation, by feeding unimodal features into a higher-dimensional space for fusion. Inspired by the TFN [14], we used a three-fold Cartesian product to fuse multiple unimodal representations and capture bimodal and trimodal interactions by multilevel fusion as follows:

$$\{(T^l, T^a, T^v) | T^l \in [{}^{T^l}_1], T^a \in [{}^{T^a}_1], T^v \in [{}^{T^v}_1]\} \tag{1}$$

where T^l, T^a, T^v denote the three unimodal tensors, and l, a, v indicates linguistic, acoustic, and visual, respectively. A constant dimension of one was added to generate unimodal and bimodal dynamics [14]. The specific calculations can be implemented as follows:

$$F_{(m)} = \left[{}^{T^l}_1\right] \otimes \left[{}^{T^a}_1\right] \otimes \left[{}^{T^v}_1\right] \tag{2}$$

where $\otimes$ denotes the outer product between vectors, and m, l, a, v indicates multimodal, linguistic, acoustic, and visual features, respectively. Thus, the fused features can be obtained by calculating the outer product.

UFF considered different factors in different fusion stages. In the first stage, the one-dimensional tensors, T^l, T^a, and T^v, were used to learn the internal representation of each modality. In the second stage, $T^l \otimes T^a$, $T^l \otimes T^v$, $T^a \otimes T^v$ were used to learn the interaction information between each pair of modalities. In the third stage, the final result was obtained by integrating the results of the previous step. Thus, the UFF fully fused multimodal data step-by-step using a Cartesian product. It first performed primary fusion through a combination of modalities in pairs, following which it performed further fusion by combining the three modalities. This step-by-step fusion method obtained the interaction information embedded in different modalities to the greatest extent from the perspective of multimodal feature fusion.

LGT: For the three unimodal tasks, we used model-based approaches for feature extraction and paid more attention to the deeper correlation between modality representations. We used the LGT to determine the correlation between modalities, and the results were used as input for unimodal tasks. As shown in Fig. 2, we chose linguistic features as the main component and acoustic and visual features as the secondary components. (The reasons for choosing linguistic features as the main components are explained in Section 5.3, and the experiment results are listed in Table 3.)

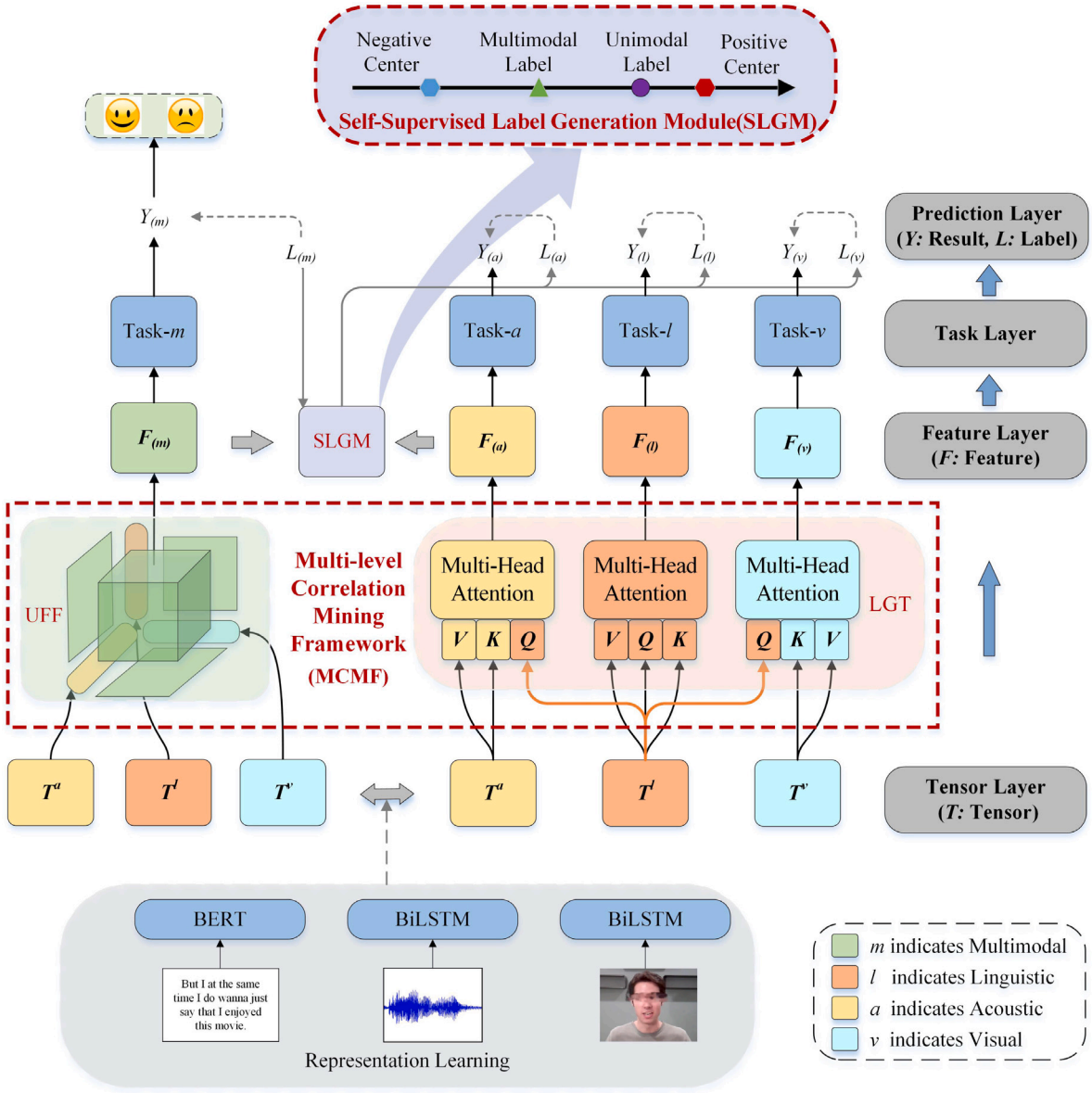

Fig. 1. Structure of the proposed model. Note: MCMF is composed of UFF and LGT. UFF fuses unimodal tensors into the multimodal feature through the three-fold Cartesian product. LGT utilizes unimodal tensors to calculate the cross-modal multi-head attention score. The multi-modal sentiment analysis task is divided into four tasks, which are jointly trained in the multi-task learning framework. SLGM generates unimodal labels based on the relationship among modality representations and labels.

The MHA of Transformer uses different subspace representations of queries, keys, and values to capture data dependencies [20]. The features of each modality were divided into query (Q), key (K), and value (V) vectors to calculate the MHA scores. The correlation information between different modalities was ignored if each modality used its own Q, K, and V vectors to mine the internal data dependencies. To bridge modalities and discover cross-modal correlation information, the LGT utilized linguistic Q for all three modalities to achieve multimodal information fusion from a cross-modal interaction perspective. The traditional Transformer encoder was composed of an MHA layer and a feed-forward neural network (FFN) layer, which we used to process the linguistic feature vectors. The MHA layer was calculated as follows:

$$Attention(Q_l, K_l, V_l) = \mathrm{softmax}((Q_l K_l^T)/\sqrt{d_k})V_l \tag{3}$$

$$head_i = Attention(Q_l W^Q, K_l W^K, V_l W^V) \tag{4}$$

$$F_{(l)} = MHA(Q_l, K_l, V_l) = Concat(head_1, \ldots, head_h)W^O \tag{5}$$

First, the feature vector T^l was divided into three vectors, Q, K, and V, following which a linear transformation was applied to each vector. Next, the Q and K vectors were used in the point product function and softmax function to calculate the similarity weight. Finally, the weighted sum of vector V was obtained from the results of the previous step. This attention calculation was iterated in the MHA layer, with each calculation being treated as a separate "head". The final result was obtained by concatenating the outputs of the multiple heads.

In addition, the output of each layer (MHA and FFN layers) was processed by a residual transformation and layer normalization (A & L):

$$LayerNorm(x + Sublayer(x)) \tag{6}$$

We used a linguistic modality to guide the secondary components in calculating the MHA score. The process was referred to as guided-MHA. In guided-MHA, the Q vector originated from the linguistic modality, and the K and V vectors originated from the acoustic and visual modalities. When learning acoustic and visual representations,

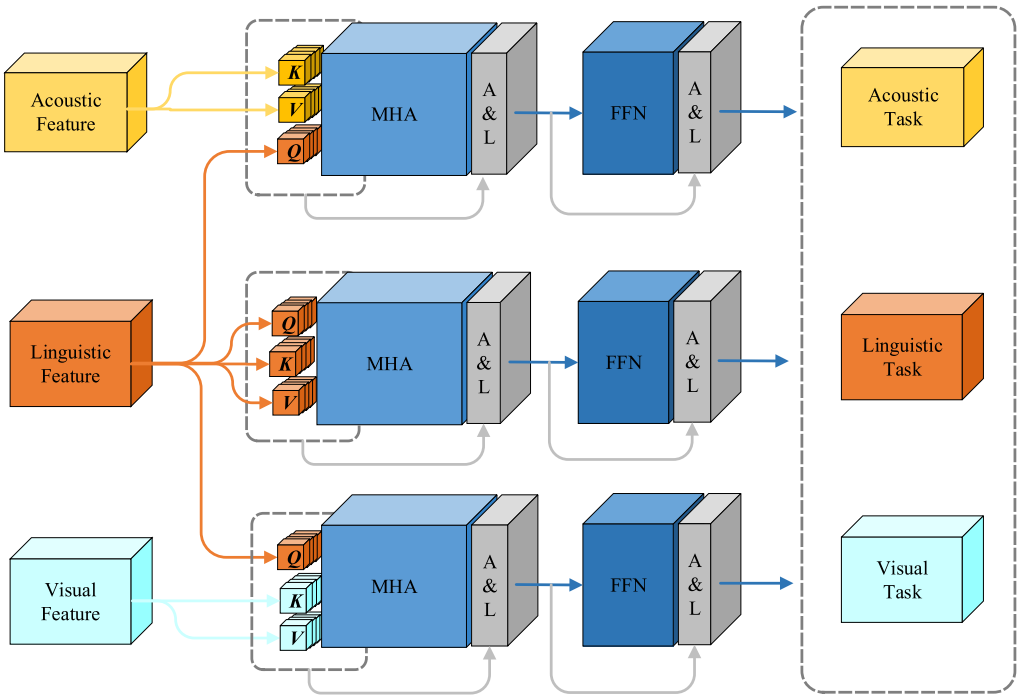

Fig. 2. Structure of the LGT.

linguistic representations were used to introduce information from various spaces.

$$F_{(a)} = Guided\text{-}MHA(Q_l, K_a, V_a) = Concat(head_1, \ldots, head_h)W^O \quad (7)$$

$$F_{(v)} = Guided\text{-}MHA(Q_l, K_v, V_v) = Concat(head_1, \ldots, head_h)W^O \quad (8)$$

3.4. Multi-task learning framework

In this study, we used a hard parameter-sharing mechanism to build a multitask learning framework. The mechanism enabled all tasks to share neurons and weights in the low-level network, while assigning task-specific neurons and weights to each task in the high-level network. As shown in Fig. 3, we utilized the bottom representation learning network as the shared layer and the prediction network as the specific layer for different tasks. Finally, we set up four different tasks in a multi-task learning framework. The input for the multimodal task was denoted as $F_{(m)}$, and the inputs for the three unimodal tasks were represented as $F_{(l)}$, $F_{(a)}$, and $F_{(v)}$. The specific layers for each task are defined by Eqs. (9)~(10), where $s \in m,l,a,v$:

$$F_s^* = ReLU(F_s W_s^{1T} + b_s^1) \quad (9)$$

$$y_s = F_s^* W_s^{2T} + b_s^2 \quad (10)$$

It is worth noting that the unimodal tasks were trained using self-supervised unimodal labels. Therefore, unimodal tasks existed only during the training stage. The prediction result of the multimodal task was used as the final sentiment output.

3.5. Self-supervised label generation module

Most multimodal sentiment analysis datasets lack the independent unimodal annotations required for the proposed model. Inspired by the ULGM [28], we designed an SLGM to address this issue. The SLGM uses multimodal annotations to generate unimodal annotations. To leverage the modality representation information and multimodal annotations to

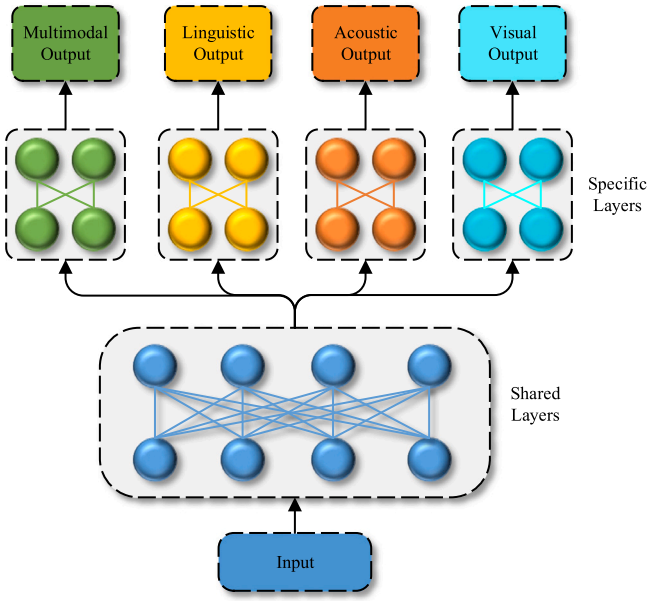

Fig. 3. Multi-task Learning Framework.

generate a unimodal supervision value, we considered two correlations when designing the SLGM. First, there is a mapping between modality representation and modality supervision value. Second, mapping using different modalities is directly proportional.

$$(F_m \circledast L_m) \propto (F_u \circledast L_u) \quad (11)$$

where $m \in \{multimodal\}$, $u \in \{linguistic, acoustic, visual\}$, F is the modality representation, L is the modality supervision value, $\circledast$ is the mapping relationship, $\propto$ is the proportional relationship.

It is evident that the unimodal and multimodal labels in the above equation are highly correlated and their sentiment polarities remain consistent. However, in real-world scenarios, the sentiment polarities

of three unimodal and multimodal labels may differ. “Crying joy” is a good example. There is no doubt that the multimodal label of this sample is “positive”. However, the visual modality might be labeled as ”negative” owing to the “crying” facial expression, if evaluated separately. To address this issue, we divided modality representations into two categories based on sentiment polarity. We then obtained the centers of both categories separately and derived a positive and negative representation center from each modality.

$$C_p = \frac{\sum_{i=1}^{N} I(y(i) > 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) > 0)} \tag{12}$$

$$C_n = \frac{\sum_{i=1}^{N} I(y(i) < 0) \cdot F_i}{\sum_{i=1}^{N} I(y(i) < 0)} \tag{13}$$

where $I(\cdot)$ is an indicator function, N is the number of samples, F_i is the representation of the i_{th} samples.

In the next step, we used the Bhattacharyya coefficient to calculate the degree of deviation between the sample and the corresponding class center:

$$S_p = \sum_{j=1}^{K} \sqrt{F(j)C_p(j)} \tag{14}$$

$$S_n = \sum_{j=1}^{K} \sqrt{F(j)C_n(j)} \tag{15}$$

where K represents the number of elements in the modality representation.

The sentiment polarity of a sample could be determined by the degree of deviation in these three situations.

(1) If $S_p > S_n$, the sample was closer to the positive center, and the sentiment polarity was positive. We chose S_p to calculate the sample supervision value.
(2) If $S_p < S_n$, the sample was closer to the negative center, and the sentiment polarity was negative. It was more appropriate to use S_n to calculate the sample supervision value.
(3) If $S_p = S_n$, the sample was at the boundary between the positive center and the negative center, and the sentiment polarity was neutral. The supervision value of this sample was 0.

Fig. 4 shows an example of sentiment polarity determination. The unimodal representation of the sample is closer to the positive center, which satisfies Situation (1). Therefore, unimodal labels should be positive.

Next, we used the ratio and difference to represent the mapping $\circledast$; thus, (11) can be represented as

$$S_m / L_m = S_u / L_u \tag{16}$$

$$S_m - L_m = S_u - L_u \tag{17}$$

$$L_u = L_m + \frac{S_u - S_m}{2} * \frac{L_m + S_m}{S_m} \tag{18}$$

where S_m and S_u indicate the multimodal and unimodal deviation degrees, respectively, and L_m and L_u indicate the multimodal label and unimodal label, respectively.

To solve the problem in which the result is unstable in each epoch of label generation, we dynamically updated the unimodal label as follows:

$$y_u^{(i)} = \frac{1}{2} * \frac{i-1}{i+1} * y_u^{(i-2)} + \frac{1}{2} * \frac{i-1}{i+1} * y_u^{(i-1)} + \frac{2}{i+1} * y_u^{i}, (i \geq 3) \tag{19}$$

From the third epoch, the results of the i_{th} epoch are related to those of the $(i-1)_{th}$ and $(i-2)_{th}$ epochs. After several iterations, the labels generated for the n_{th} epoch will be stable. The self-supervised label generation policy is presented in Algorithm 1.

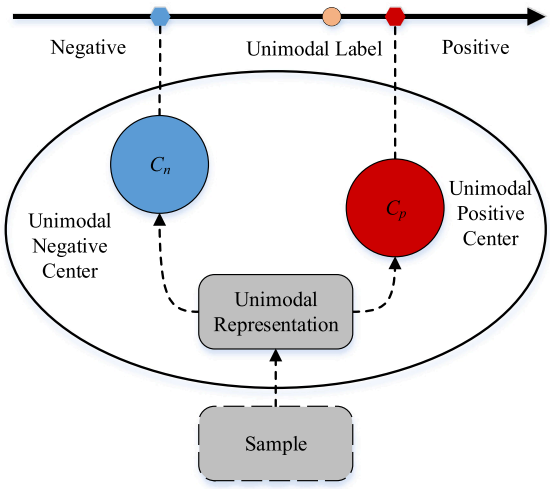

Fig. 4. Example of sentiment polarity determination in SLGM.

4. Experiment settings

In this section, the datasets, baselines, and parameter settings used in the study are introduced. During the experiment, we used the Adam optimizer to train the model with an initial learning rate and adopted L1 Loss as the optimization objective.

4.1. Datasets

In this study, we used two classical datasets from the field of multimodal sentiment analysis to verify the effectiveness of our model. Figs. 5 and 6 show the sentiment distributions for both datasets.

The **CMU-MOSI** dataset [33] is the first opinion-level annotated corpus for sentiment and subjectivity analyses of online videos. The dataset contains 2199 short video clips, and the samples are labeled using sentiment annotation within the range of [−3, 3]. The number of the positive, negative and neutral segments are 1080, 1022 and 97. The samples in the CMU-MOSI dataset were split. The training, test, and validation sets contain 1284, 686, and 229 samples, respectively.

The **CMU-MOSEI** dataset [34] contains 22856 utterance videos from more than 1000 online YouTube speakers. It contains sentiment annotations within the range of [−3, 3] and emotion annotations, including happiness, sadness, anger, fear, disgust, and surprise, according to Ekman emotions theory. The number of the positive, negative and neutral clips are 11264, 6594 and 4998. The CMU-MOSEI dataset was also split into training, validation, and test sets. There are 16326 samples in the training set, 4659 samples in the test set, and 1871 samples in the validation set, respectively.

4.2. Baselines

MulT. The MulT [21] uses directional pairwise cross-modal attention to improve its framework. It utilizes the interactions between multimodal sequences across distinct time steps to achieve semantic blending.

B2+B4. The gated mechanism for attention (B2+B4) utilizes self-attention to capture long-term context and uses a gating mechanism to selectively learn cross-attended features [35].

MISA. The modality-invariant and -specific representations (MISA) [36] learns the commonness and independence of modal representations by projecting each modality onto two distinct subspaces.

CM-BERT. Cross-modal BERT (CM-BERT) [37] uses the interaction of linguistic and acoustic modalities to improve a pretrained BERT

Algorithm 1: Self-Supervised Label Generation Module (SLGM)

Input: Modality representations F_m, F_u and multimodal labels L_m
Output: Unimodal supervision values L_u

for $n \in [1, End]$ **do**
  for *batch in dataLoader* **do**
    for $m \in [fusion]$ **do**
      Compute the batch multimodal class center C_p^m, C_n^m using Equation (12-13)
      for $t \in [1, N]$ **do**
        Compute the sample similarity $S_p^m(t)$ and $S_n^m(t)$ using Equation (14-15)
      end
    end
    for $u \in [linguistic, acoustic, visual]$ **do**
      Compute the batch unimodal class center C_p^u and C_n^u using Equation (12-13)
      for $t \in [1, N]$ **do**
        Compute the sample similarities $S_p^u(t)$ and $S_n^u(t)$using Equation (14-15)
        if $S_p > S_n$ **then**
          $S_u = S_n^u, S_m = S_n^m$
        else if $S_p < S_n$ **then**
          $S_u = S_p^u, S_m = S_p^m$
        else if $S_p = S_n$ **then**
          $L^u(t) = 0$
        Compute the unimodal supervision values $L^u(t)$ using Equation (18).
      end
      Save $L_u = \{L^u(1), L^u(2), ..., L^u(N)\}$
    end
    Update the unimodal label using Equation (19).
  end
end

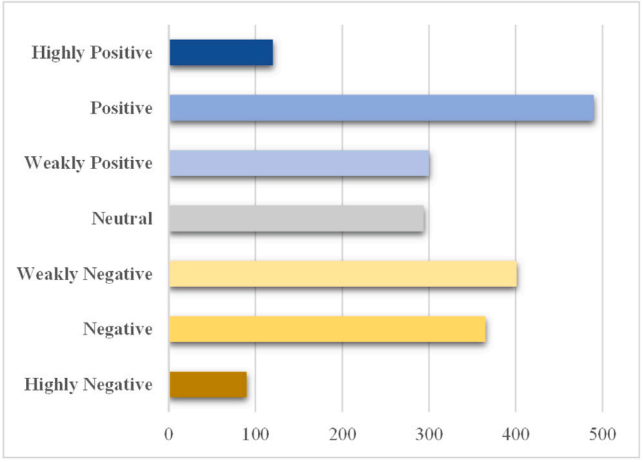

Fig. 5. Distribution of sentiment over the CMU-MOSI dataset [33].

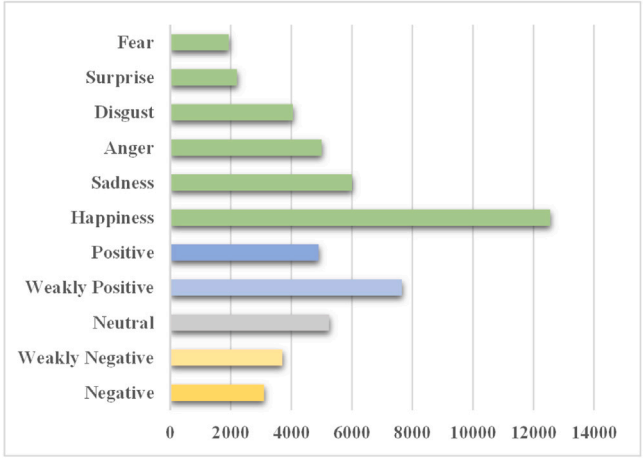

Fig. 6. Distribution of emotions and sentiment over the CMU-MOSEI dataset [34].

model. Its core concept is masked multimodal attention, which adjusts the weights of words through cross-modal interactions.

MAG-BERT. The multimodal adaptation gate (MAG) [22] improves sentiment analysis by introducing multimodal nonverbal data into the pre-training model BERT.

Self-MM. Self-supervised multitask learning for multimodal sentiment analysis (self-MM) [28] aids multimodal tasks by jointly learning multimodal and unimodal tasks using a multi-task learning framework.

TEDT. The Transformer-based encoding–decoding translation (TEDT) [23] network takes natural language as the main modality and the nonnatural language data as the auxiliary modality, and improves the effect of multimodal feature integration through a modality reinforcement cross-attention module.

SIMR. The SIMR [24] framework separates the nonverbal data into style encoding and content representation to reduce the impact of the personalized acoustic and visual features, and simultaneously discovers compatible and incompatible cross-modal interactions using an enhanced Transformer module.

AOBERT. AOBERT [25] is a single-stream Transformer that determines dependency and relationship between modalities using pre-training tasks.

4.3. Experiment details

The proposed model was implemented based on Python3.6 and Pytorch1.2.0. All the experiments were performed using a TESLA-V100 GPU. Considering that CMU-MOSI and CMU-MOSEI have already been

divided into training, validation, and test sets, and that most of the SOTA models followed this setting in their experiments, we also conducted the experiment according to this setting. For a fair comparison, the average of the results of five repeated experiments was taken as the final result. In addition, the range of the batch size was $\{16, 32\}$, the range of the learning rate was $\{1e-4, 1e-3, 5e-3\}$, the range of learning rate decay period was $\{5, 10, 20\}$, the range of hidden layer size was $\{16, 32, 64, 128\}$, the range of gradient clipping threshold was $\{-1.0, 0.8, 1.0\}$, and the range of training period was $\{10, 20, 30, 40\}$.

4.4. Evaluation metrics

The CMU-MOSI and CMU-MOSEI datasets provide seven levels of sentiment labels in the interval of $\{-3, +3\}$ because the multimodal sentiment analysis task can be evaluated both as a classification problem and a regression problem. The evaluation indices for the regression task were the mean absolute error (MAE) and Pearson's correlation (Corr), and the binary accuracy (Acc-2) and F1-Score for the classification task. Except MAE, the model performance was directly proportional to the value of the evaluation metric.

5. Results and analysis

5.1. Quantitative results

Table 1 lists the results for the CMU-MOSI dataset. We compared the performances of the different models on both the classification and regression tasks based on specific evaluation metrics. For the classification tasks, the left of the "/" represents "negative/non-negative", while the right represents "negative/positive". In our five replicate experiments, the maximum standard deviation of all the indicators was 0.69.

The best performance results are highlighted in bold. In general, our model achieved excellent results, ranking among the top three for all the metrics. Our model showed clear advantages in terms of Acc-2 and F1-score when calculated as "negative/positive", which was only slightly inferior to the newly proposed model TEDT. When calculated as "negative/non-negative", our method still achieved competitive results compared to AOBERT. TEDT performed best in the "negative/positive" classification tasks because a modality reinforcement cross-attention module and a noise-filtering gate module were employed to reduce the negative impact of nonnatural language data. However, the "negative/non-negative" classification results of TEDT were not reported. On the contrary, AOBERT performed poorly in the "negative/positive" classification tasks. Our model shows good adaptability, as it performed well in both "negative/non-negative" and "negative/positive" classification tasks. Our model also achieved significant improvements in terms of the MAE and Corr for the regression tasks. These experiment results demonstrate that multitask learning provides a new approach for multimodal sentiment analysis and that joint training using multiple tasks is superior to training using a single task. Furthermore, multilevel association mining yielded more useful information than single-level mining.

We also tested the performance of the proposed model on another multimodal sentiment-analysis dataset. Table 2 presents the results for the CMU-MOSEI dataset. Some comparisons are missing because the studies did not provide the results for this specific dataset. The "negative/non-negative" classification results of TEDT were still not presented, but AOBERT did achieve good results in almost all the classification tasks. This is because AOBERT was trained using a bert-large-uncased pre-trained model with 340M parameters. Nevertheless, our method achieved competitive results on all the indices compared to AOBERT without using such a large model. From this perspective, our method still has its advantages. In the five replicate experiments conducted, the maximum standard deviation of all the indicators was 0.74.

Table 1
Result on CMU-MOSI dataset.

Model	MAE	Corr	Acc-2	F1-Score
MulT [21]	0.871	0.698	–/83.0	–/82.8
MISA [36]	0.783	0.761	81.8/83.4	81.7/83.6
CM-BERT [37]	0.729	0.791	84.5/-	84.5/-
MAG-BERT [22]	0.712	0.796	84.2/86.1	84.1/86.0
Self-MM [28]	0.713	0.798	84.00/85.98	84.42/85.95
TEDT [23]	0.709	**0.812**	–/**89.3**	–/**89.2**
SIMR [24]	0.706	0.798	84.2/86.1	84.0/86.1
AOBERT [25]	0.856	0.700	**85.2**/85.6	**85.4**/86.4
Ours	**0.69**	0.81	85.15/88.43	85.25/88.43

Table 2
Result on CMU-MOSEI dataset.

Model	MAE	Corr	Acc-2	F1-Score
B2+B4 [35]	–	–	81.14	78.53
MulT [21]	0.580	0.703	–/82.5	–/82.3
MISA [36]	0.555	0.756	83.6/85.5	83.8/85.3
Self-MM [28]	0.530	**0.765**	82.81/85.17	82.53/85.30
TEDT [23]	0.524	0.749	–/**86.2**	–/**86.1**
STMR [24]	0.580	0.696	82.5/82.9	81.9/82.9
AOBERT [25]	0.515	0.763	**84.9**/**86.2**	**85.0**/85.9
Ours	**0.51**	0.74	84.66/86.16	84.72/85.88

In addition, we present the sentiment analysis results for a video clip in the CMU-MOSI dataset. As shown in Fig. 7, three consecutive sentences were selected from the video for testing. The figure shows the entire sentence and key frame pictures. The red line connects each image to its corresponding word. The left side of the figure shows the label and prediction results. The prediction results of Segments 2 and 3 were consistent with the labels. However, the prediction result of Segment 1 was different from that of the label. This is because the first few frames of the clip misled the model. It can be clearly observed that the first picture in Segment 1 expresses a positive sentiment because the man is smiling. The corresponding text is neutral. Under these circumstances, the final analysis results were affected and the sentiment was predicted to be weakly positive.

5.2. Visualization of attention in LGT

To illustrate the function of LGT, we visualized the effect of MHA in the traditional Transformer and guided-MHA in LGT, respectively. We chose a sentence ("I thought it was fun") from the CMU-MOSI dataset as an example. The intra- and inter-modal attention relationship is illustrated in Fig. 8. MHA only calculates the self-attention scores within each modality. By contrast, guided-MHA calculates both intra-modal self-attention scores and cross-modal attention scores.

Considering it is easier to analyze the textual content directly, we visualized the self-attention score matrix of the linguistic modality using MHA and guided-MHA, as shown in Fig. 9, to explore the impact of guided-MHA. The color gradients represent the attention scores between words. It can be observed that the attention weight between words in this sentence changed when guided-MHA was adopted. We used blue and black boxes to emphasize the most significant change after using guided-MHA. In the MHA matrix, the word "thought" obtained a high attention score on the word "was". The word "fun" obtained a high attention score on the words "thought" and "was". However, in the guided-MHA matrix, the scores changed. The attention scores between "I", "thought" and "fun" increased significantly, indicating that guided-MHA captured the general meaning of the entire sentence more accurately. This is because acoustic and visual modalities provide useful reference information through interactions between the modalities. This also shows that LGT can reveal the correlation information between modalities, which might be beneficial for sentiment analysis.

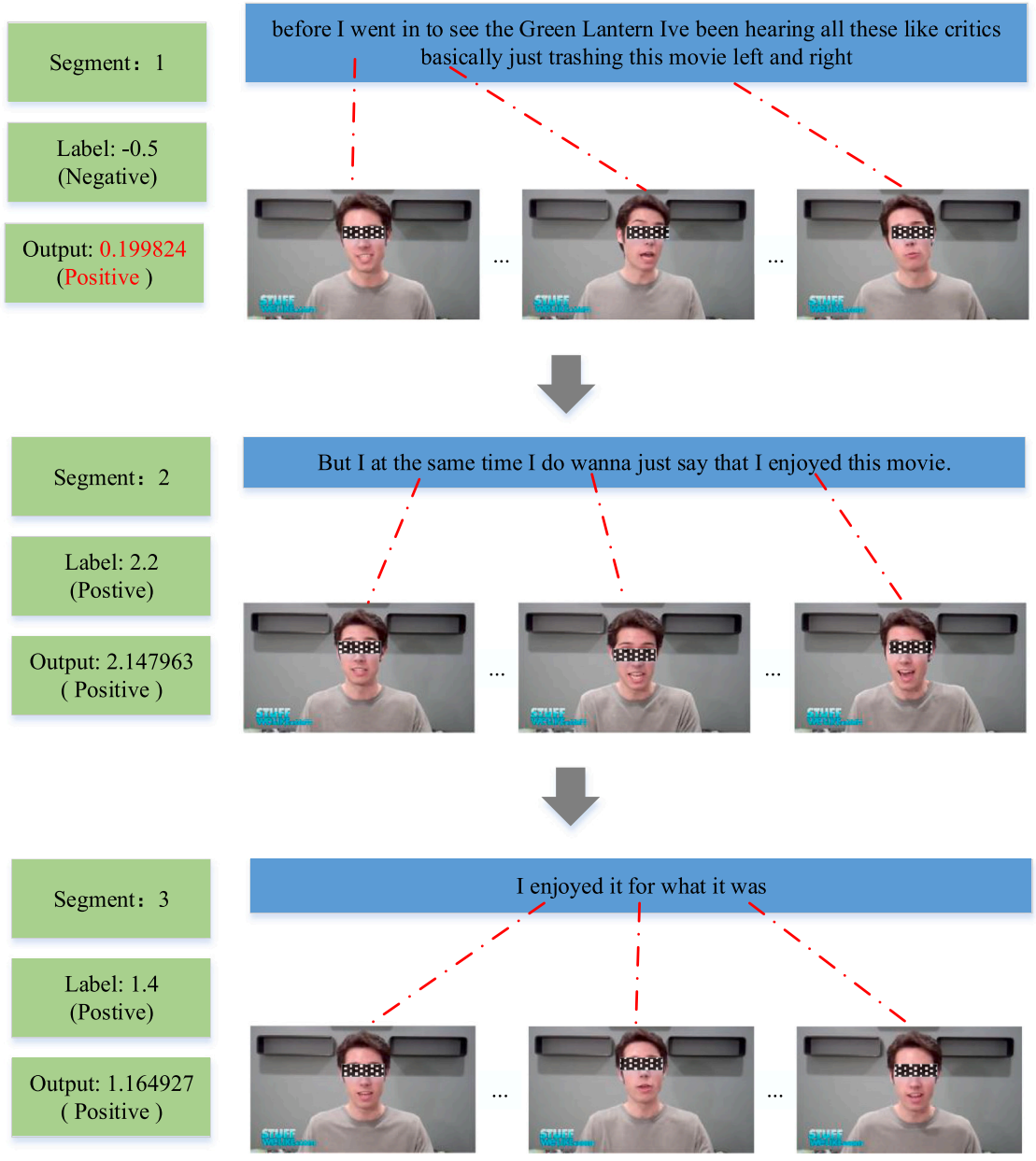

Fig. 7. Several examples of video sentiment analysis.

5.3. Ablation study

To explore the contribution of each component to the proposed model, ablation experiments were conducted on the CMU-MOSI dataset.

First, we compared the effectiveness of each component in the model. Fig. 10 shows the results. Additionally, we adjusted for the missing components in the ablation study to ensure the structural integrity of the model. When the UFF module was removed, we fused unimodal features through feature stitching. When the LGT module was missing, we directly used unimodal representations as the input for unimodal tasks. In the absence of the SLGM, we used multimodal labels instead of unimodal labels.

The performance comparison results for a single component showed that UFF had the greatest impact on the performance of the model. When only the UFF module was used, the model achieved acceptable sentiment prediction results. This indicates that the UFF module uncovers the basic relationship among modalities, which plays the most critical role in multimodal sentiment analysis tasks. However, problems were encountered when both components were combined. The best combination was UFF+SLGM, which was consistent with our expectations. However, the other combinations performed unsatisfactorily. UFF+LGT lacked independent unimodal labels, and LGT+SLGM could not uncover the basic relationship among modalities. Nevertheless, when LGT was removed, the sentiment prediction accuracy was significantly worse than the best result with all the three components. This demonstrated that the LGT effectively mined the potential correlations among modalities and improved the sentiment prediction accuracy.

Next, we compared the effectiveness of the different tasks. As the three unimodal representations may not appear simultaneously, we made some changes to the UFF during the comparison. When only one unimodal representation was used, it became the main component of the UFF and guided-MHA was no longer used. When there were two unimodal representations, the linguistic modality was still used as the main component, and the other modality was used as the secondary component with guided-MHA. When neither representation contained a

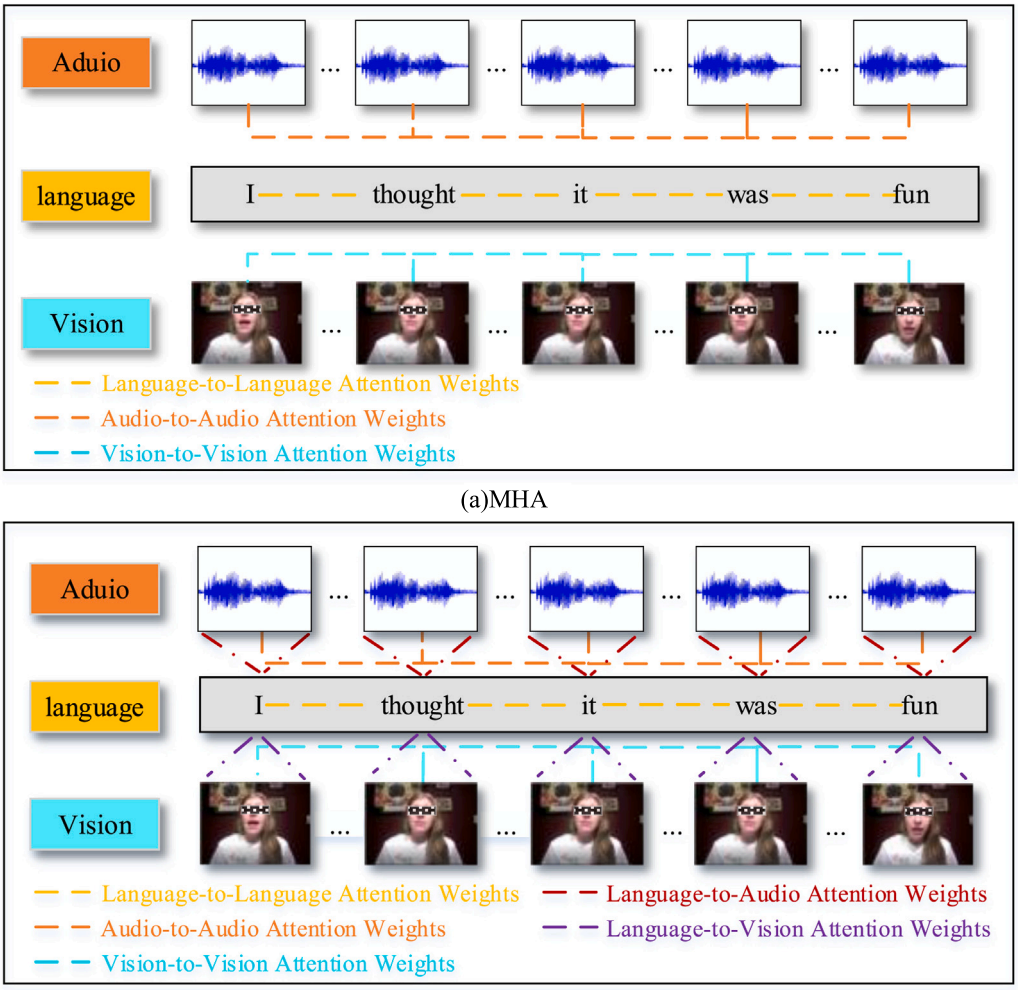

Fig. 8. Visualization of the attention relationship using MHA and guided-MHA.

linguistic modality, guided-MHA was not used. Fig. 11 shows the results of multimodal sentiment analysis with different tasks, where M, L, A, and V represent the multimodal, linguistic, acoustic, and visual tasks, respectively. The results show that using a combination of multiple modalities is effective for sentiment analysis. Satisfactory results were achieved when two or three modalities were combined.

Furthermore, we made linguistic modality the main component in the design of the LGT because, as shown in Fig. 11, the language subtask had the most significant impact of all the subtasks. Additionally, we attempted to design an acoustic-guided Transformer (AGT) with acoustic modality as the main component and a visual-guided Transformer (VGT) with visual modality as the main component. As shown in Table 3, the experiment results of the performance comparison on the CMU-MOSI dataset confirmed that the contribution of the linguistic modality was the most prominent. The LGT achieved significant advantages in classification tasks. These results justify the assumption of modality contribution in sentiment analysis, and, by extension, the LGT design in this study.

Table 3
Horizontal comparison results of LGT, AGT, and VGT.

Model	MAE	Corr	Acc-2	F1-Score
AGT	**0.64**	**0.81**	83.41/86.57	83.52/86.58
VGT	0.69	0.80	84.28/87.04	84.33/87.00
LGT	0.69	**0.81**	**85.15/88.43**	**85.25/88.43**

5.4. SLGM experiments

To evaluate the reasonability and robustness of the SLGM, we extracted several SLGM-generated labels during the training process and compared them with human annotations. As shown in Table 4, the unimodal labels of Samples 1–3 were consistent with the sentiment polarity of human annotation, indicating that the unimodal labels generated by the SLGM were valuable. However, the unimodal labels of Samples 4–5 exhibited a negative offset, unlike human annotation, which showed that the SLGM adapted to special situations such as irony. Overall, these results demonstrate the reasonability and robustness of the SLGM for generating unimodal labels for multimodal sentiment analysis.

5.5. Discussion

Based on previous experiments, we obtained the following findings:

(1) Multi-task learning with both multimodal labels and independent unimodal labels yields more sentiment information and achieves higher performance. This is because relying solely on multimodal annotations may cause the model to overlook unique attributes in the unimodal data.
(2) Among the existing multimodal data, linguistic modalities provide the richest sentiment information. Therefore, the proposed

Table 4
Several examples of generated labels.

	Original Information	MOSI-M	SLGM-L	SLGM-A	SLGM-V
1	But I at the same time I do want to just say that I enjoyed this movie.	2.2	1.1925	0.3874	0.8557
2	Anyhow it was really good	2.4	1.7874	0.0639	1.1984
3	But you know for this one I just didn't care	−1.8	−1.5281	−0.9052	−1.3683
4	Um however it is this loyalty to the original source that gives it its flaws	−0.2	−0.0327	0.0874	0.0001
5	And there were times that I thought it was funny	0.6	0.8281	0.0052	−0.0856

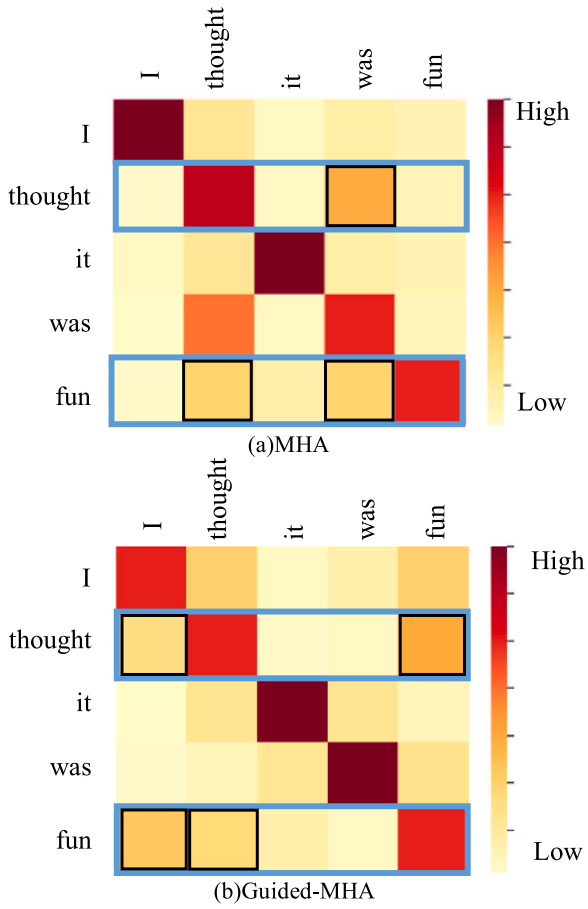

Fig. 9. Visualization of the self-attention score matrix of the linguistic modality.

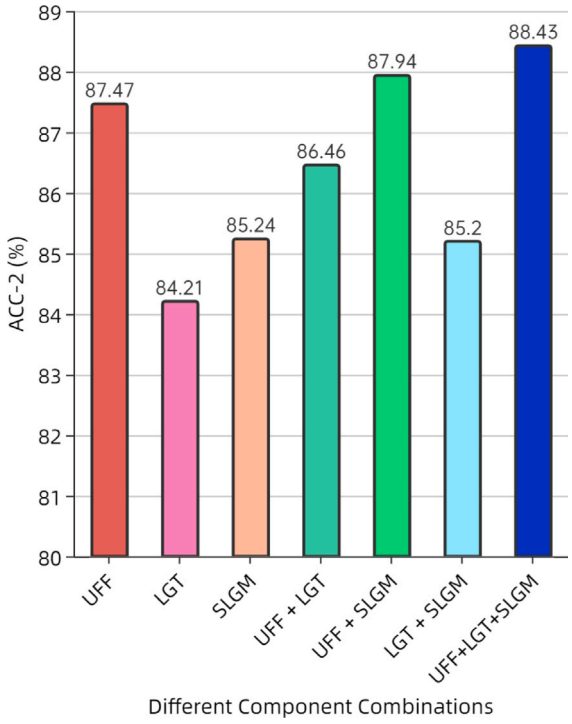

Fig. 10. Ablation experiment results for each component.

LGT model can reasonably adjust the weight of words in sentences during the sentiment analysis process, leveraging the rich information available from linguistic modalities.

(3) The experiment results reveal a mapping relationship between the representation space of modalities and their corresponding label space. Based on this concept, the proposed SLGM was designed to generate independent unimodal labels with a certain degree of reliability.

However, the use of hard-sharing mechanism in multi-task learning requires a strong correlation between the different subtasks. Noise in the automatically generated unimodal labels affects the effectiveness of correlation mining between unimodal and multimodal tasks. Therefore, further exploration is required to determine how to preserve rich correlation information between tasks and achieve better generalization effects.

6. Conclusion

We proposed a multimodal sentiment analysis model that mines associations among modalities and correlations among multimodal and unimodal labels. The model employs three unimodal sentiment analysis tasks with independent labels to aid multimodal sentiment tasks. To fully leverage multimodal information, we used UFF to fuse low-level features and LGT to determine the correlation between multimodal representations. To address the lack of reliable independent annotations in the training process for unimodal tasks, we proposed the SLGM to generate unimodal labels based on multimodal labels. The proposed model achieved competitive results on the CMU-MOSI and CMU-MOSEI datasets. The MCMF achieved better performance in multimodal data fusion and multimodal association mining by combining bottom-up and top-down methods. The multitask learning framework for multimodal sentiment analysis was more robust than that of the single-task for

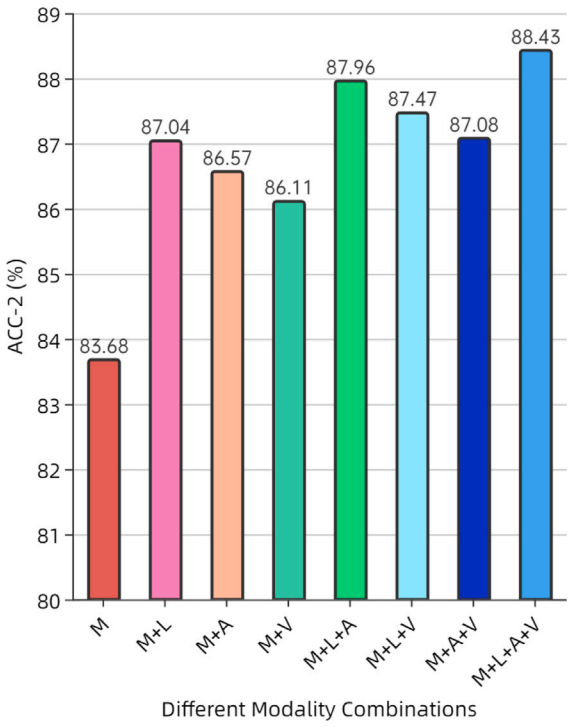

Fig. 11. Ablation experiment results on different tasks.

co-learning multimodal information. In unimodal label generation, the SLGM, based on the association between modality representations and labels, produced more reliable labels. In future work, we plan to improve the robustness and applicability of our model by using a more flexible multitask learning framework in the presence of label noise.

CRediT authorship contribution statement

Zuhe Li: Methodology design and development; creation of models, Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation). **Qingbing Guo:** Conducting research and investigation process, specifically performing the experiments and data/evidence collection, Programming, software development, designing computer programs, implementation of computer code and supporting algorithms, and testing existing code components. **Yushan Pan:** Ideas: Formulation or evolution of overarching research goals and aims, Development or design of methodology and creation of models. Oversight and leadership responsibility for research activity planning and execution, including mentorship external to the core team. **Weiping Ding:** Verification, whether as part of the activity or separately, of the overall replication/reproducibility of the results/experiments and other research outputs. **Jun Yu:** Preparation, creation, and/or presentation of the published work, specifically writing the initial draft (including substantive translations). **Yazhou Zhang:** Preparation, creation, and/or presentation of the published work, specifically, Writing the initial draft (including substantive translation). **Weihua Liu:** Management activities to annotate (produce metadata), scrub data, and maintain research data (including software code where it is necessary to interpret the data) for initial use and later reuse. **Haoran Chen:** Provision of instrumentation, computing resources and other analytical tools. **Hao Wang:** Preparation, creation, and/or presentation of published work by those from the original research group, specifically critical review, commentary, or revision, including the pre-or postpublication stages. **Ying Xie:** Critical review, commentary, or revision, including the pre-or postpublication stages.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this study are openly available in [CMU-MOSI] and [CMU-MOSEI], reference number [33] and [34].

Acknowledgments

This study was supported by the National Natural Science Foundation of China under Grant 61702462, 62276146 and 61906175, the XJTLU RDF-21-02-008, the Henan Provincial Science and Technology Research Project under Grant 222102210010, 222102210064, 232102211006, 232102210044, the Research and Practice Project of Higher Education Teaching Reform in Henan Province under Grant 2019SJGLX320 and 2019SJGLX020, the Undergraduate Universities Smart Teaching Special Research Project of Henan Province under Grant Jiao Gao [2021] No. 489-29, the Academic Degrees & Graduate Education Reform Project of Henan Province under Grant 2021SJGLX 115Y.

References

[1] Sarah A. Abdu, Ahmed H. Yousef, Ashraf Salem, Multimodal video sentiment analysis using deep learning approaches, a survey, Inf. Fusion 76 (2021) 204–226.
[2] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, Kurt Keutzer, Emotion recognition from multiple modalities: Fundamentals and methodologies, IEEE Signal Process. Mag. 38 (6) (2021) 59–73.
[3] Yingying Jiang, Wei Li, M. Shamim Hossain, Min Chen, Abdulhameed Alelaiwi, Muneer Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, Inf. Fusion 53 (2020) 209–221.
[4] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al., A systematic review on affective computing: Emotion models, databases, and recent advances, Inf. Fusion (2022).
[5] Chao Zhang, Zichao Yang, Xiaodong He, Li Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Top. Sign. Proces. 14 (3) (2020) 478–493.
[6] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, Xiangjie Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion (2023).
[7] Tadas Baltrušaitis, Chaitanya Ahuja, Louis-Philippe Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[8] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, Amir Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion (2022).
[9] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf. (Long Pap.), Volume 1, 2018, pp. 2247–2256.
[10] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, Gerhard Rigoll, LSTM-modeling of continuous emotions in an audiovisual affect recognition framework, Image Vis. Comput. 31 (2) (2013) 153–163.
[11] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, Stéphane Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, in: Proc. Annu. Meet. Assoc. Comput Linguist., 2020, pp. 1–7.
[12] Sijie Mai, Ya Sun, Ying Zeng, Haifeng Hu, Excavating multimodal correlation for representation learning, Inf. Fusion 91 (2023) 542–555.
[13] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, Kaicheng Yang, Ch-sims: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3718–3727.
[14] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, Louis-Philippe Morency, Tensor fusion network for multimodal sentiment analysis, in: EMNLP - Conf. Empir. Methods Nat. Lang. Process., Proc., 2017, pp. 1103–1114.
[15] Saurav Sahay, Shachi H. Kumar, Rui Xia, Jonathan Huang, Lama Nachman, Multimodal relational tensor network for sentiment and emotion classification, 2018, arXiv preprint arXiv:1806.02923.

[16] Yazhou Zhang, Lu Rong, Dawei Song, Peng Zhang, A survey on multimodal sentiment analysis, Pattern Recognit. Artif. Intell. 33 (2020) 426–438.
[17] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, Roger Zimmermann, Icon: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.
[18] Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, Panpan Wang, Quantum-inspired interactive networks for conversational sentiment analysis, IJCAI Int. Joint Conf. Artif. Intell. 2019-August (2019) 5436–5442.
[19] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, Soujanya Poria, Cosmic: Commonsense knowledge for emotion identification in conversations, 2020, arXiv preprint arXiv:2010.02795.
[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[21] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, Ruslan Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Volume 2019, NIH Public Access, 2019, p. 6558.
[22] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, Ehsan Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Volume 2020, NIH Public Access, 2020, p. 2359.
[23] Fan Wang, Shengwei Tian, Long Yu, Jing Liu, Junwen Wang, Kun Li, Yongtao Wang, TEDT: Transformer-based encoding-decoding translation network for multimodal sentiment analysis, Cogn. Comput. 15 (2023) 289–303.
[24] Jianwen Wang, Shiping Wang, Mingwei Lin, Zeshui Xu, Wenzhong Guo, Learning speaker-independent multimodal representation for sentiment analysis, Inform. Sci. 628 (2023) 208–225.
[25] Kyeonghun Kim, Sanghyun Park, AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.
[26] Yu Zhang, Qiang Yang, A survey on multi-task learning, IEEE Trans. Knowl. Data Eng. 34 (12) (2021) 5586–5609.
[27] Yazhou Zhang, Jinglin Wang, Yaochen Liu, Lu Rong, Qian Zheng, Dawei Song, Prayag Tiwari, Jing Qin, A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations, Inf. Fusion (2023).
[28] Wenmeng Yu, Hua Xu, Ziqi Yuan, Jiele Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Volume 35, Number 12, 2021, pp. 10790–10797.
[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, Volume 1, 2019, pp. 4171–4186.
[30] Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, Helen Meng, Inferring user emotive state changes in realistic human-computer conversational dialogs, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 136–144.
[31] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, Louis-Philippe Morency, Openface 2.0: Facial behavior analysis toolkit, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, IEEE, 2018, pp. 59–66.
[32] Weijiang Li, Fang Qi, Ming Tang, Zhengtao Yu, Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification, Neurocomputing 387 (2020) 63–77.
[33] Amir Zadeh, Rowan Zellers, Eli Pincus, Louis-Philippe Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016, arXiv preprint arXiv:1606.06259.
[34] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, Louis-Philippe Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.
[35] Ayush Kumar, Jithendra Vepa, Gated mechanism for attention based multi modal sentiment analysis, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 4477–4481.
[36] Devamanyu Hazarika, Roger Zimmermann, Soujanya Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.
[37] Kaicheng Yang, Hua Xu, Kai Gao, Cm-bert: Cross-modal bert for text-audio sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 521–528.