Coordinated-joint Translation Fusion Framework with Sentiment-interactive Graph Convolutional Networks for Multimodal Sentiment Analysis

Qiang Lu[a], Xia Sun[a,*], Zhizezhang Gao[a], Yunfei Long[b], Jun Feng[a], Hao Zhang[c]

[a]*School of Information Science and Technology, Northwest University, Xi'an 710127, China*
[b]*School of Computer Science and Electrical Engineering, University of Essex, Colchester CO43SQ, UK*
[c]*Graduate School, Shaanxi University of Chinese Medicine, Xianyang 712083, China*

Abstract

Interactive fusion methods have been successfully applied to multimodal sentiment analysis, due to their ability to achieve data complementarity via interaction of different modalities. However, previous methods treat the information of each modality as a whole and usually treat them equally, failing to distinguish the contribution of different semantic regions in non-textual features towards textual features. It caused that the public regions fail to be captured and private regions are hard to be predicted only with textual. Meanwhile, these methods use sentiment-independent encoder to encode textual features, which may mistakenly identify syntactically irrelevant contextual words as clues for judging sentiment. In this paper, we propose a coordinated-joint translation fusion framework with sentiment-interactive graph to solve these problems. Specifically, we generate a novel sentiment-interactive graph to incorporate sentiment associations between different words into the syntactic adjacency matrix. The relationships between nodes are no longer limited to the sole existence of syntactic associations but fully consider the interaction of emotions between different words. Then, we designed a coordinated-joint translation fusion module. This module utilizes a cross-modal masked attention mechanism to determine whether there is a correlation between the text and non-text inputs, thereby identifying the most relevant public semantic features in the

*Corresponding author

Email addresses: nwulq@stumail.nwu.edu.cn (Qiang Lu), raindy@nwu.edu.cn (Xia Sun), 202210338@stumail.nwu.edu.cn (Zhizezhang Gao), yl20051@essex.ac.uk (Yunfei Long), fengjun@nwu.edu.cn (Jun Feng), 1271009@sntcm.edu.cn (Hao Zhang)

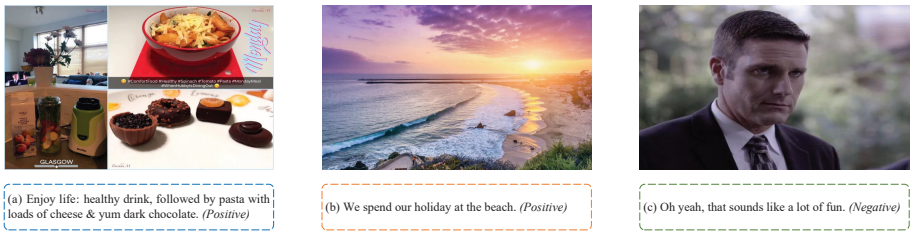
visual and acoustic modalities corresponding to the text modality. Subsequently, a cross-modal translation-aware mechanism is used to calculate the differences between the visual and acoustic modalities features transformed into the text modality and the text modality itself, which allows us to reconstruct the visual and acoustic modalities towards text modality to obtain private semantic features. In addition, we construct a multimodal fusion layer to fuse textual features and non-textual public and private features to improve multimodal interaction effects. Experimental results on publicly available datasets CMU-MOSI and CMU-MOSEI illustrate that our proposed model achieve a best accuracy of 86.5% and 86.1%, and best F1 of 86.4% and 86.1%. A series of further analyses also indicate the proposed framework effectively improve the sentiment identification capability.

Keywords: Multimodal sentiment analysis, Multimodal fusion, Sentiment-interactive graph, Cross-modal masked attention, Cross-modal translation-aware mechanism

1. Introduction

Sentiment analysis aims to predict the sentiment polarities of opinion holders such as positive, negative and neutral [2]. Previous studies focus on the textual sentiment analysis, and it has been extensively applied in daily life [68]. For example, in e-commerce, sentiment analysis can help enterprises improve the quality of goods and services according to feedback of customers by analyzing their reviews [26]. However, with the development of multimedia technology and social networks, people express their views and emotions through more diverse ways in the multimedia scene. Human cognition does not only come from single textual data. In real scenes, text, image and video data often appear simultaneously. Furthermore, it is difficult to accurately judge the sentiment state only by text in some cases, such as irony and sarcasm [22]. Irony and sarcasm often combine neutral or positive textual content and audio expression that does not match the content to complete a negative sentiment expression [23]. Above cases are challenging to be solved fundamentally only by single text data. Therefore, multimodal sentiment analysis that combines multiple modalities has attracted considerable attention in recent years.

Multimodal sentiment analysis (MSA) is an important yet challenging task in natural language processing, and it understands the attitudes and views of opinion holders by combing different modalities such as textual, visual and acoustic data [40]. Multimodal sentiment analysis follows the principle of complementarity which focuses on textual information, and makes more accurate predictions by supplementing the text with visual and acoustic information. For example, Fig. 1 (a) presents a positive sentiment via textual words such as enjoy, healthy and yum, and makes the textual representation more vivid via image data. The text fails to clearly show any sentiment in Fig. 1 (b), and bright colours from image can complement the text data to predict positive sentiment. And text represents a positive sentiment with fun in Fig. 1 (c), but image show a negative sentiment via grim expression. Therefore, multimodal sentiment analysis more adapts to the demand of multimodal scene, and realizes more accurate sentiment prediction via abundant data from multi modalities [19].



(a) Enjoy life: healthy drink, followed by pasta with loads of cheese & yum dark chocolate. *(Positive)*

(b) We spend our holiday at the beach. *(Positive)*

(c) Oh yeah, that sounds like a lot of fun. *(Negative)*

Figure 1: The architecture of the proposed sentiment interaction and multi-graph perception graph convolutional network

This paper aims to achieve more accurate multimodal sentiment prediction by exploring multimodal fusion methods. Early multimodal fusion studies have achieved some progress [4, 30, 31]. However, previous methods have two issues: First, these early multimodal fusions methods usually utilize neural networks such as recurrent neural network (RNN) and convolutional neural network (CNN) to extract the text semantic features. The RNN-based models are only suitable for processing sequential information and cannot handle tree or graph structures, it is insufficient to capture syntactical dependencies within a sentence. And, the CNN-based models can only

perceive multi-word features as consecutive words with the convolution operations over word sequences, but are inadequate to determine sentiments depicted by multiple words that are not next to each other. This may lead them to mistakenly identify syntactically irrelevant contextual words as clues for predicting sentiment [28, 66]. Second, previous studies [43, 63] have demonstrated that the visual and acoustic features can improve the final performance of models. These methods only study the influence of different modal interactions on the performance, and fail to explain why visual or acoustic modalities can assist text semantics and which regions play a complementary role in text semantics. They treat the information of each modality as a whole and fuse textual, visual and acoustic features equally, lacking the ability to distinguish the contribution of different semantic regions in non-textual features towards textual features.

To address the above issues, we propose a coordinated-joint translation fusion network (CJTF) for multimodal sentiment analysis. On the one hand, the consistency principle of multimodal complementarity makes the text information occupy a greater proportion in different modalities. Therefore, the understanding of text semantics is crucial to the final performance. The previous methods usually utilize the RNN and CNN models to extract the semantic information. However, these methods fail to capture textual syntactic features. Graph convolutional networks (GCNs) have been successfully applied to textual sentiment analysis, due to their ability to flexibly capture syntactic information [50, 60, 69]. But these GCN methods have limitations: whether there is dependency between two nodes. If there is a dependency, the node value at the corresponding position of the adjacency graph is set to one; otherwise, it is set to zero. It will lead to the disappearance of the contextual sentiment clues. In our view, the nodes should not only contain syntactic dependencies but also the sentiment interactions between different words should be fully considered. For example, in the sentence "Macbook notebooks quickly die out because of their short battery life.". "Short battery life" leads to "Macbook notebooks quickly die out", there are sentiment correlations and influences between different words. Therefore, we construct a sentiment-interactive graph convolutional network (SIGCN) which fuses sentiment interaction relations to syntactic adjacency matrix to construct sentiment-interactive graph to capture long-distance dependencies of syntactics and sentiment interaction of semantics.

On the other hand, the previous methods have always emphasized that the complementation between different modalities is benefit for multimodal fusion, but they don't explain which part of visual or acoustic information plays a role and what roles they play in interaction. We consider that there are two features in multimodal interaction. One is public feature that jointly describes entities to enhance text semantics, and the other is private feature that contains unique information to coordinately complement other modalities. Therefore, we design a coordinated-joint translation fusion module to capture public and private features. We first utilize a coordinated module with cross-modal mask attention to extract the non-textual public features that visual and acoustic towards textual to enhance the textual semantics. Specifically, we calculate the text-oriented visual and acoustic cross-modal attention score to identify the most relevant regions in the visual and acoustic modalities that correspond to the text modality. Based on these regions, to capture visual and acoustic public semantics for textual modality, we pay more attention to text features that contain a higher degree of shared semantics. Therefore, we use a cross-modal mask attention mechanism to determine whether there is an association between textual and non-textual modalities to mask irrelevant features, thereby achieving the extraction of non-textual public features. Then, a joint translation fusion module with cross-modal translation-aware mechanism is designed to complement textual semantics by translating non-textual private features. We first generate a query matrix based on textual modality, key matrices based on visual and acoustic modalities, and value matrices based on visual and acoustic modalities. Then, based on query, key and value matrix, we utilize a multi-head self-attention mechanism to calculate the differences between the features of visual and acoustic modalities transformed into textual modality and the text modality itself. Subsequently, we employ an attention-aware mechanism to reconstruct the visual and acoustic modalities towards textual modality by leveraging the differences between the text modality and non-text modalities to obtain private semantic features. Finally, the public and private features are fused into a multimodal fusion layer to predict sentiment polarity. Experiments on two publicly-available datasets show that the textual features which contains semantic and syntactic information as well as public and private feature are contribute to predict multimodal sentiment.

The main contributions can be summarized as follows:

- We propose a sentiment-interactive graph convolutional network to capture the long-distance dependencies of syntactics and sentiment interaction of semantics.
- We design a coordinated module with cross-modal masked attention mechanism to extract the non-textual public features, which enhances the textual semantics representation by calculating text-oriented visual and acoustic public semantic contribution.
- We construct a joint translation module with the cross-modal translation-aware mechanism to capture non-textual private features, which supplements the textual semantics representation by translating text-oriented visual and acoustic private semantic features.
- Experimental results on two public datasets CMU-MOSI and CMU-MOSEI illustrate that our proposed model outperforms advanced baseline methods and verify the effectiveness of our model.

The remainder of this paper is organized as follows. After introducing previous works in Section 2, we propose a coordinated-joint translation fusion network in Section 3. Then we describe the experimental details and analysis in Section 4. Finally, we summarize our work and provide a direction of future work in Section 5.

2. Related work

Previous studies of sentiment analysis have been applied in textual, visual and acoustic filed [14, 15, 39]. With the development of multimedia technologies, multimodal data such as text, image and video are growing exponentially and has gradually become the main form of data. Due to the limited information obtained by unimodal sentiment analysis, achieving accurate analysis in some specific scenarios is difficult. Therefore, multimodal sentiment analysis has attracted considerable attention of sentiment analysis [19]. Different from unimodal sentiment analysis which only contains one modal information, multimodal sentiment analysis combines textual, visual and acoustic

modal information to make expression more vivid and accurate [44]. In this section, we introduce multimodal sentiment analysis in two parts: multimodal feature extraction and multimodal fusion methods.

2.1. Multimodal feature extraction

Previous MSA methods follows the principle of complementarity that focuses on textual information, and predict sentiment polarity via the supplement of non-textual information. They usually apply neural networks to extract multimodal features such as convolutional neural networks, recurrent neural networks and pre-training models (PTMs).

CNNs extract the local features of text well via the local receptive field and weight sharing operation. Poria et al. [36] proposed a multi-kernel learning method which used the text hidden representation extracted by CNNs as the feature of the high-level classifier. On the basis of this work, Poria et al. [37] further discussed the role of the general framework for multi-modal sentiment analysis, and proposed a convolutional MKL method with CNN for multimodal sentiment analysis.

Despite CNN-based methods could capture local semantic information, pooling operations resulted in the loss of overall semantic dependency. RNNs automatically learn the global semantic features and save the sequential information with special gates and cells. Xu et al. [57] proposed a deep semantic network MultiSentiNet for multimodal sentiment analysis. Attention mechanism breaks the limitation of RNNs that the input depends on the output of the previous time, and often used in conjunction with RNNs. Akhtar et al. [1] proposed a contextual inter-modal attention framework based on RNN, which used multimodal and contextual information to simultaneously predict the sentiment and emotion of a discourse in multi-task learning.

Most MSA methods with CNNs and RNNs with attention mechanism are insufficient to capture the complex sentiment dependency in sentences. Simply utilizing the rich knowledge learned as the contextual embedding already achieves a large performance gain such as BERT [12]. Yu et al. [61] introduced a novel weight self-adaption strategy to balance the loss constraints of different tasks. This method extracted text features through BERT and mapped different modalities to a unified space through

ReLU activation function. Ghorbanali et al. [16] proposed a hybrid MSA model based on weighted convolutional neural networks. They utilized BERT model to receive the textual descriptions of the images to extract the text features and imported these features into a weighted convolutional neural network ensemble. The above studies show the excellent performance in capturing semantic information, ignoring an important problem, i.e., syntactic dependency [66].

There are some GCN studies in textual sentiment analysis [5, 24, 58], but most models only regard the relationship between sentences as whether there are syntactic dependencies, i.e., whether they are connected or not, ignoring their internal sentiment interaction relations. Meanwhile, few studies use GCNs in multimodal sentiment analysis. Therefore, we design a sentiment-interactive graphs to address the limitation of most GCNs. The proposed method performs a graph convolution on the top of LSTM to extract the long-distance syntactic dependencies and sentiment interaction of semantics.

2.2. Multimodal fusion methods

Most existing models use multimodal fusion method to map multimodal features to unified semantic space for multimodal sentiment classification. Multimodal fusion predicts the results by integrating information from multiple modalities, focuses on how to integrate multi-modal data with a certain architecture or approach and jointly contribute to solving the target task [67]. Multimodal fusion methods are classified into two categories: Model-agnostic and Model-based. The model-agnostic methods do not directly depend on specific deep learning methods, while the model-based methods apply deep learning models to explicitly solve multimodal fusion problems [3].

Model-agnostic methods are divided into early fusion, late fusion and hybrid fusion. Early fusion, also known as feature-level fusion, completes the fusion of features before inputting the classifier by extracting features from different modal information. Early fusion can better capture the interaction between modalities, and only one model needs to be trained to complete feature fusion of different modalities. Therefore, it is widely used in the early research of multimodal sentiment analysis [10, 34, 52], but it fails to address issues of time asynchronous and redundancy of data. Late fusion is also called

decision-level fusion. Different modal features are modeled separately, and then the output from model is integrated to produce final prediction. The processing of late fusion is irrelevant to features and requires multi-network models for training which can adapt well to the problem of modal missing. Therefore, some studies utilized the late fusion methods to model multimodal sentiment analysis [53, 59]. However, this method lacks low-level interaction of multimodal data and more computationally intensive. Hybrid fusion uses early fusion and late fusion respectively to realize multimodal sentiment analysis [17, 47]. It combines the advantages of early and late fusion, also increases structural complexity and training difficulty of the prediction model.

Model-based methods focus on multimodal data fusion using neural networks such as CNNs, LSTMs and PTMs. Zadeh et al. [62] introduced a Tensor Fusion Network (TFN) to explicitly aggregates multimodal interactions. To combine cues from different modalities, Chen et al. [8] proposed a gated multimodal embedding LSTM (GME-LSTM(A)) with time attention to perform modal fusion at the word level. Each modality has its own representation space and contains some knowledge that other views cannot access. Therefore, Zadeh et al. [63] proposed a memory fusion network (MFN) with multi-view sequential learning to comprehensively and accurately describe multimodal data. Wang et al. [51] believed that integrating the features of different modalities and improving the performance are the main challenges of multimodal sentiment analysis tasks, and proposed an end2end fusion method with transformers for multimodal sentiment analysis. Zhu et al. [70] learned the corresponding relationship between regions and words from text-image pairs, and proposed an image-text interaction network for multi-modal sentiment analysis. Chen et al. [7] introduced a classifier for image-text relevance into the multimodal task, and unified separate fusion strategies into a holistic framework for adaptive sentiment analysis.

The above methods utilize different technologies from various angles to attempt multimodal fusion to better realize multimodal emotional analysis. However, they fused textual, visual and acoustic features equally, failing to capture public semantic regions and private regions are hard to be predicted only with textual. It caused the robustness and accuracy of the model to decline. Therefore, we design a coordinated-joint translation modules to explore the semantic contributions of different modal

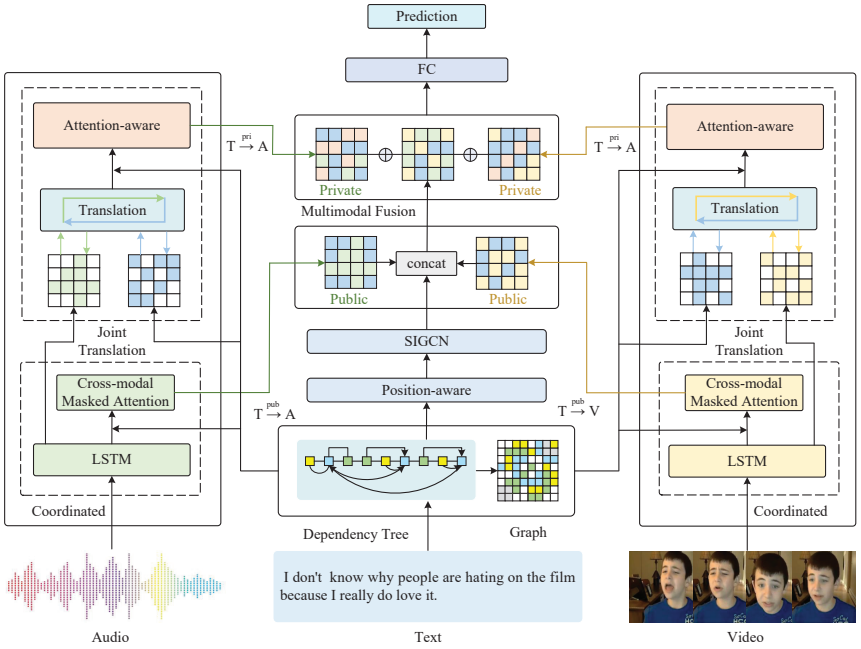


Figure 2: The architecture of proposed CJTF contains three components: sentiment-interactive graph convolutional network, coordinated-joint translation fusion module and multimodal sentiment prediction.

regions.

3. Methodology

In this section, the proposed coordinated-joint translation fusion network (CJTF) is described in detail. As demonstrated in Fig. 2, the architecture of CJTF contains three components: (1) Sentiment-interactive graph convolutional network. (2) coordinated-joint translation fusion modules. (3) Multimodal sentiment prediction. CJTF first design a sentiment-interactive graph convolutional network by deriving sentiment interaction of each node to capture the long-distance dependencies of syntactics and sentiment interaction of semantics. Then, coordinated-joint translation modules are designed to enhance and complement the textual semantics by translating non-textual public and private features. Finally, the public semantic features and private semantic features are fusion to predict multimodal sentiment polarities.

3.1. Notation definition

Given an utterance which includes textual, visual and acoustic modalities, the input of three modalities is denoted as $x^k = \{x_i^k \,|\, 1 \leq i \leq L, k \in \{t, v, a\}\}$, where $x_i^k \in \mathbb{R}^{d_k \times n_k}$, in which $d_k = \{d_k \,|\, k \in \{t, v, a\}\}$ denotes the dimension of unimodal features, and $n_k = \{n_k \,|\, k \in \{t, v, a\}\}$ is the number of utterances. L is the length of the given sequence.

3.2. Sentiment-interactive graph convolutional networks

To extract the basic semantics, we first utilize the LSTMs [18] to encoder the text, video and audio to extract textual, visual and acoustic features h_i^k

$$h_i^k = Bi - LSTM(x_i^k) \tag{1}$$

where $h_i^k = \{h_i^k | 1 \leq i \leq L, k \in \{t, v, a\}$, $h_i^k \in \mathbb{R}^{2d_h}\}$. Then, we perform a graph convolution on the top of h_i^t to capture syntactic information. Different from the previous GCNs that only considered a syntactic dependency between contexts, we design a sentiment-interactive GCN to capture the syntactic dependencies and semantic interaction by deriving the long-range sentiment relations of each node towards itself and contexts.

3.2.1. Original graph

We construct the original graph based on dependency trees to capture the syntactic dependencies of sentences[1]. If node has syntactic relations in dependency tree, the value of node in adjacency graph is set to one, otherwise it is set to zero.

$$D_{i,j}^O = \begin{cases} 1, & \text{if } i = j \text{ or } s_i^t \text{ and } s_j^t \text{ in the dependency tree} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Where $D_{i,j}^O \in \mathbb{R}^{n \times n}$ is the adjacency matrix of original graph, and s_i^t denotes the words of sentence. In the adjacency matrix, each node is set to be adjacent to itself, and the value of diagonal is all set to one.

[1]We use spaCy toolkit to construct the dependency tree: `https://spacy.io/`

3.2.2. Sentiment-interactive graph

The original graph considers the syntactic dependencies of adjacent nodes. To explore the semantic interaction of different nodes, we generate an interactive adjacency matrix by deriving the long-range syntactic relations of each node based on original graph.

$$A_{i,j}^{I} = \begin{cases} 1 + 1/\left(1 + p_e\right), & \text{if } s_i \text{ and } s_j \in T \\ 1 + 1/\left(|j - p_e| + 1\right), & \text{if } s_i \in T \\ 1 + 1/\left(|i - p_e| + 1\right), & \text{if } s_j \in T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $A_{i,j}^{I} \in \mathbb{R}^{n \times n}$ is the adjacency matrix of interactive graph. T is the dependency tree. p_e denotes the end position of sentence, and $|\cdot|$ is an absolute value function. In the sentence, the node-centric word also has a sentiment impact on the current node except sentiment words. To capture this sentiment impact and augment the contextual-aware abilities of nodes, we fuse sentiment interaction relations to interactive adjacency matrix to construct sentiment-interactive graph.

$$G_{i,j}^{SI} = D_{i,j}^{O} + D_{i,j}^{O} * A_{i,j}^{SI} \tag{4}$$

$$A_{i,j}^{SI} = \begin{cases} D_{i,j}^{O} + 1/(p_e + 1) * D_{i,j}^{O} * A_{i,j}^{I}, & \text{if } D_{i,j}^{O} = 1 \\ 1 + 1/(|j - p_e|) * A_{i,j}^{I}, & \text{if } s_i \in T \\ 1 + 1/(|i - p_e|) * A_{i,j}^{I}, & \text{if } s_j \in T \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Where $A_{i,j}^{SI} \in \mathbb{R}^{n \times n}$ is the adjacency matrix of sentiment-interactive graph. $D_{i,j}^{O} \in \mathbb{R}^{n \times n}$ and $A_{i,j}^{SI} \in \mathbb{R}^{n \times n}$ present the adjacency matrices of original graphs and interactive graphs. To further enhance the direction-aware of GCN, we construct the sentiment-interactive graph in unidirectional, i.e., $G_{i,j}^{SI} = G_{j,i}^{SI}$, where $G_{i,j}^{SI}$, $G_{j,i}^{SI} \in \mathbb{R}^{n \times n}$. The process of generating a sentiment-interactive graph is shown in Algorithm 1.

We input the hidden representation h_i^t and graph $G_{i,j}^{SI}$ into GCN after obtaining the sentiment-interactive graph. Significantly, we first utilize the position-aware mechanism

Algorithm 1 The pseudocode of sentiment-interactive graphs

Input:1 A sentence $S = \{s_1^t, s_2^t \cdots, s_n^t\}$; The dependency tree T and matrix $A_{i,j}^I$.
Output:2 The sentiment-interactive graphs $G_{i,j}^{SI}$.

while $T <$maximum number of iterations **do**
  for $i = 1 \rightarrow n; j = 1 \rightarrow n$ **do**
    ▷ The process of generating original graph
    if i = j or s_i and s_j in T **then**
      $D_{i,j}^O \leftarrow 1$
    else
      $D_{i,j}^O \leftarrow 0$
    end if
    ▷ The process of generating sentiment-interactive graph
    if $D_{i,j}^O$ = 1 **then**
      $A_{i,j}^{SI} \leftarrow D_{i,j}^O + 1/(p_e + 1) * D_{i,j}^O * A_{i,j}^I$
    else if s_i in T **then**
      $A_{i,j}^{SI} \leftarrow 1 + 1/(|j - p_e|) * A_{i,j}^I$
    else if s_j in T and A^s **then**
      $A_{i,j}^{SI} \leftarrow 1 + 1/(|i - p_e|) * A_{i,j}^I$
    else
      $A_{i,j}^{SI} \leftarrow 0$
    end if
  end for
  for $i = 1 \rightarrow n; j = 1 \rightarrow n$ **do**
    $G_{i,j}^{SI} = D_{i,j}^O + D_{i,j}^O * A_{i,j}^{SI}$
    $G_{i,j}^{SI} \leftarrow G_{j,i}^{SI}$
    return G^{SI}
  end for
end while

to translate the hidden representation h_i^t to reduce the noise, and then perform the graph convolution in the form of L layers on the top of h_i^t to make each node update via a normalization factor.

$$h_i^l = ReLU(\sum_{j=1}^{n} G_{i,j}^{SI} W_l \, g_j^{l-1}/d_i + b_l) \tag{6}$$

$$g_i^l = \frac{n-i}{n} * h_i^t \tag{7}$$

Where $h_i^l \in \mathbb{R}^{2d_h}$ denotes the hidden representation of SIGCN, and $g_i^l \in \mathbb{R}^{2d_h}$ denotes the position-aware hidden representation. $W_l \in \mathbb{R}^{d_h \times 2d_h}$ and $b_l \in \mathbb{R}^{d_h}$ are the weight parameters. $d_i = \sum_{j=1}^{n} G_{i,j}^{si}$ is the degree of dependency tree. $ReLU()$ is a non-linear activation function.

3.3. Coordinated-joint translation fusion module

In the above section, we obtain the textual representation via SIGCN, visual representation h_i^v and acoustic representation h_i^a via LSTMs. Next, we design a coordinated-joint translation fusion module to extract public and private features from three modal representation to enhance and supplement the textual representation. The coordinated-joint translation fusion module contain three components: coordinated module, joint translation module and multimodal fusion module.

3.3.1. Coordinated module

Textual, visual and acoustic information have common attributes when they represent the same object, and these attributes describe the object from different aspects. As shown in Fig.1 (a), drinks and food in the image with mellow colours correspond to text words such as healthy and yum, and they cooperate with each other to enhance semantic representation to predict positive polarity. To enhance the textual representation, we construct a coordinated module to utilize the coordinated interaction of non-modalities. First, we calculate the text-oriented visual and acoustic cross-modal attention score.

$$\beta_t^{v\to l} = W_{vl} tanh(W_{v\to l}[h_t^l; h_t^v] + b_{v\to l}) \tag{8}$$

$$\beta_t^{a\to l} = W_{al} tanh(W_{a\to l}[h_t^l; h_t^a] + b_{a\to l}) \tag{9}$$

Where $\beta_t^{v\to l}$ and $\beta_t^{a\to l}$ denote the attention score of non-textual features towards textual. W_{vl}, $W_{al} \in \mathbb{R}^{1\times 2d_h}$, $W_{v\to l}$, $W_{a\to l} \in \mathbb{R}^{2d_h\times 4d_h}$, and $b_{v\to l}$, $b_{v\to l} \in \mathbb{R}^{2d_h}$ are the trainable weight parameters. h_t^l is the textual representation of SIGCN. $[\,;\,]$ represents the concat operation and $tanh()$ is the non-linear activation function. Then, inspired by previous studies of mask [46, 55], we design a cross-modal mask attention mechanism to mask irrelevant features to capture the public features of non-textual modalities towards textual modality.

$$\alpha_t^{v\to l} = mask_t^{pub} \cdot \frac{\exp(\beta_t^{v\to l})}{\sum_{i=1}^{L} \exp(\beta_j^{v\to l})} \tag{10}$$

$$\alpha_t^{a\to l} = mask_t \cdot \frac{\exp(\beta_t^{a\to l})}{\sum_{i=1}^{L}\exp(\beta_j^{a\to l})} \tag{11}$$

$$mask_t = \begin{cases} 1 + 1/(|t-L|), & \text{if } \{v, a \to l\} \in R^{pub} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Where $\alpha_t^{v\to l}$, $\alpha_t^{a\to l} \in \mathbb{R}^{1\times 2d_h}$ are the attention weight. $exp()$ is the exponential function. $mask_t \in \mathbb{R}^{L\times L}$ denotes the public mask, and $R^{pub} \in \mathbb{R}^{L\times L}$ is a attention graph. Finally, we obtain public feature by calculating h_t^v, $\alpha_t^{v\to l}$ and h_t^a, $\alpha_t^{a\to l}$.

$$h_t^{v\to l} = \sum_{t=1}^{L} \alpha_t^{v\to l} \cdot h_t^v \tag{13}$$

$$h_t^{a\to l} = \sum_{t=1}^{L} \alpha_t^{a\to l} \cdot h_t^a \tag{14}$$

Where $h_t^{v\to l}$, $h_t^{a\to l} \in \mathbb{R}^{L\times 2d_h}$ is the non-textual public features, and h_i^v, h_i^a is the visual and acoustic representation.

3.3.2. Joint translation module

Textual, visual and acoustic information contain a large number of private features in addition to the common attributes that describe objects. As shown in Fig. 1 (b), Private features such as bright sun, blue sea and golden sand beach in the image combine with textual features to complement word sand beach from different aspects. Therefore, we construct a joint translation module with cross-modal translation-aware mechanism to capture non-textual private features to supplement textual semantics. We first design a cross-modal translation mechanism to translate non-textual to capture non-textual private features.

$$\begin{aligned} \eta_t^{v\to l} &= Translation(H_V, H_L) \\ &= softmax(\frac{Q_L {K_V}^{\mathrm{T}}}{\sqrt{d_k}})V_V \\ &= softmax(\frac{H_L W_{Q_L} {W_{K_V}}^{\mathrm{T}} H_V^{\mathrm{T}}}{\sqrt{d_k}})H_V W_{V_V} \end{aligned} \tag{15}$$

$$\begin{aligned}\eta_t^{a\to l} &= Translation(H_A, H_L) \\ &= softmax(\frac{Q_L {K_A}^{\mathrm{T}}}{\sqrt{d_k}})V_A \\ &= softmax(\frac{H_L W_{Q_L} {W_{K_A}}^{\mathrm{T}} H_A^{\mathrm{T}}}{\sqrt{d_k}})H_A W_{V_A}\end{aligned} \tag{16}$$

Where $\eta_t^{v\to l}$, $\eta_t^{v\to l} \in \mathbb{R}^{L\times 2d_h}$ is the non-textual private weight. H_L, H_V and H_A represent the textual hidden representation h_i^l and non-textual representation h_i^v, h_i^a. $Q_L = H_L W_{Q_L}$, $K_V = H_V W_{K_V}$ and $V_V = H_V W_{V_V}$ denote the query matrix, key matrix and value matrix of visual features towards the textual. $Q_L = H_L W_{Q_L}$, $K_A = H_A W_{K_A}$ and $V_A = H_A W_{V_A}$ denote the query matrix, key matrix and value matrix of acoustic features towards the textual. $W_{Q_L} \in \mathbb{R}^{d_t\times d_k}$, $W_{K_V} \in \mathbb{R}^{d_v\times d_k}$ and $W_{V_V} \in \mathbb{R}^{d_v\times d_v}$ are the linear translation weight matrices of visual features towards the textual. $W_{Q_L} \in \mathbb{R}^{d_t\times d_k}$, $W_{K_A} \in \mathbb{R}^{d_a\times d_k}$ and $W_{A_V} \in \mathbb{R}^{d_a\times d_v}$ are the linear translation weight matrices of acoustic features towards textual. Then, we utilize an attention-aware mechanism to compute the weighted sum and represent as the private representations.

$$p_t^{v\to l} = \sum_{t=1}^{L} \eta_t^{v\to l} \cdot h_t^v \tag{17}$$

$$p_t^{a\to l} = \sum_{t=1}^{L} \eta_t^{a\to l} \cdot h_t^a \tag{18}$$

Where $p_t^{v\to l}$, $p_t^{a\to l} \in \mathbb{R}^{L\times 2d_h}$ is the non-textual private representations, and h_i^v, h_i^a is the visual and acoustic representation of LSTMs.

3.3.3. Multimodal fusion module

In this section, we first compute the non-textual public features and textual representation as the public representations.

$$h_t^{pub} = Multimodal_{fusion}[h_t^{v\to l}; h_t^{v\to l}; h_t^l] \tag{19}$$

Where $h_t^{pub} \in \mathbb{R}^{L\times 6d_h}$ denotes the final public representations. $Multimodal_{fusion}$ is a fusion layer which contains a Bi-LSTMs and a self-attention module. Then, we utilize

a ReLU activation function to fuse the public representations and private representations to predict multimodal sentiment.

$$\hat{y} = W_f(ReLu(W_h[h_t^{pub}; p_t^{v\to l}; p_t^{a\to l}] + b_h)) + b_f \tag{20}$$

Where $\hat{y}$ is the predict label. $W_f \in \mathbb{R}^{1\times d_h}$, $W_h \in \mathbb{R}^{d_h\times 10d_h}$, $b_f \in \mathbb{R}$ and $b_h \in \mathbb{R}^{d_h}$ denote the trainable weight parameters. The overall learning of the model is to optimize all the parameters, and minimize the loss function as far as possible.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} ||y_i - \hat{y}_i||_2^2 \tag{21}$$

4. Experiments

In this section, we first describe the experimental datasets in Section 4.1. Then, the implementation details and baseline models are described in Sections 4.2 and 4.3. To evaluate the performance of our proposed model, we compare it with advanced baselines on CMU-MOSI and CMU-MOSEI, and utilize the Acc, F1-score, MAE and Corr as the evaluation metrics in Section 4.4. Next, we conduct an ablation study to evaluate the contribution of the SIGCN module, the coordinated module and the joint translation module in Section 4.5 and explore the influence of the number of SIGCN layers to model performance in Sections 4.6. To prove the long-distance syntactic dependencies and non-textual public and private features is benefit to improve the performance of multimodal sentiment analysis, we conduct a case study with baselines in Section 4.7. Finally, we construct a visualization of the CJTF to demonstrate the interaction between different modalities of the CJTF in Section 4.8.

4.1. Experimental datasets

We conduct experiments on two publicly available datasets, CMU-MOSI datasets: CMU Multimodal Opinion level Sentiment Intensity datasets [2] [64] and CMU-MOSEI datasets: CMU Multimodal Opinion Sentiment and Emotion Intensity datasets [3] [65].

[2] http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/
[3] http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/

CMU-MOSI obtains 93 video clips from YouTube, and generates 2199 subjective video clips and 1503 objective video clips through subjective annotation. The sentiment label range of each utterance is annotated as $[-3, 3]$. CMU-MOSEI collects 3228 video clips and 23453 sentences from YouTube. The sentiment annotation is similar to that of CMU-MOSI which adopts a sentiment score of $[-3, 3]$, and adopts six emotions of joy, sadness, surprise, anger, disgust and fear. In accordance with most previous studies, we adopt binary classification accuracy (ACC), F1-Score (F1), Mean Absolute Error (MAE) and the correlation coefficient (Corr) to evaluate on CMU-MOSI and CMU-MOSEI. We use 1284, 229 and 686 utterances as training, validation and testing set on CMU-MOSI, and use 16216, 1871 and 4625 utterances as training, validation and testing set on CMU-MOSEI.

4.2. Implementation details

In our experiments, we adopt the GloVe vectors [33] with 300 dimensions to initialize the word embedding, and the origin visual features and acoustic features are obtained by Facet [13] and COVAREP [11]. The batch size is set to 32 and the number of SIGCN layers is set to 2. The dimension of hidden state is set to 100. The max length of CMU-MOSI and CMU-MOSEI are set to 50 and 128. We use the Adam optimizer to optimize all models and the learning rate is set to 0.001. To optimize the model training, we average the experimental results of 40 runs with random initialization.

Table 1: Hyperparameters of our model.

Hyperparameters	MOSI	MOSEI
Batch Size	32	32
Max Length L	50	128
Hidden Size d_h	100	100
Hidden Size of BERT	768	768
Learning Rate	1e-3	1e-3
Learning Rate of BERT	5e-5	5e-5
Optimizer	Adam	Adam
Dropout	0.5	0.5
GCN layer	2	2
Epoch	40	40

4.3. Baseline models

We compare CJTF with 18 state-of-the-art baselines, 6 attention-based, 2 graph-based, 5 interaction-based, and 5 pre-trained-based:

MFN [63] simulated the dynamics in a specific view and utilized a memory attention network to fuse specific and cross view for predicting the sentiment category.

LMF [27] proposed a Low-rank Multimodal Fusion method that performed multimodal fusion using low-rank tensors to improve efficiency.

RAVEN [49] designed a recurrent attended variation embedding network (RAVEN) to model expressive nonverbal representations.

MulT [45] proposed an end-to-end model which adopted the cross-modal attention to learn representations directly from unaligned multimodal streams.

CIA [6] utilized inter-modal interactive modules and context-aware attention module to increase the confidence of individual task in prediction.

MCTN [35] proposed a method to learn robust joint representations by translating between modalities to ensure that joint representations retain maximal information from all modalities.

MMGraph [29] devised a graph pooling fusion network to automatically learn the associations between various nodes from different modalities.

GATE [21] introduced the conditional gating mechanism to learn better cross modal information and applied a self attention layer on unimodal contextual representations to capture long term dependencies.

MAG-BERT [38] fine-tuned BERT and attached a carefully designed Multimodal Adaptation Gate (MAG) to the models.

GraphCAGE [54] adopted the graph construction and graph aggregation to compute in parallel in the time dimension.

TCSP [55] designed the cross-modal prediction task to explore the shared and private semantics via training two cross-modal prediction models.

MPT [9] applied a sampling function to generate sparse attention matrices and compressed a long sequence to a shorter sequence of hidden states.

MMLATCH [32] used the high-level representations extracted by the network and low-level input features to model the interaction relations.

CubeMLP [41] explored multimodal approaches with a feature-mixing perspective, and introduced a multimodal feature processing framework based entirely on MLP.

PS-Mixer [25] designed a Polar-Vector (PV) to determine the polarity of the sentiment and devised the MLP-Communication module to reduce the interference of noise and facilitate multimodal interactions.

EMT [42] proposed a generic and unified framework to employ utterance-level representations from each modality as the global multimodal context to interact with local unimodal features and mutually promote each other.

AOBERT [20] introduced a single-stream transformer which was pre-trained on two tasks simultaneously to address traditional fusion methods have some loss of intramodality and inter-modality.

TETFN [48] proposed a novel Text Enhanced Transformer Fusion Network which learned text-oriented pairwise cross-modal mappings for obtaining effective unified multimodal representations.

4.4. Results and analysis

To evaluate the performance of our proposed model, we utilize Acc, F1-score, MAE and Corr as the evaluation metrics on CMU-MOSI and CMU-MOSEI, as shown in Table 2 and 3. The compared baseline models are divided into attention-based, graph-based, interaction-based and bert-based. First, compared with attention-based baselines, we can find the performance of them is lower than CJTF. Attention-based models often utilize LSTMs to encode textual features, such as MFN and CIA. LSTMs are a variant of the RNN model that processes sequential information to extract semantic features of textual modalities and integrates modalities through attention mechanisms. However, these models are unable to handle tree structures, resulting in their inability to capture crucial syntactic dependency information within textual modalities and the sentiment interaction relations between different words. Additionally, because the current input of LSTM depends on the output of the previous time step, they may mistakenly identify syntactically irrelevant contextual words as clues for judging sentiment. On the basis of using LSTM to capture semantic information, CJTF performs graph convolution operations on the top of LSTM to capture syntactic dependency

Table 2: Performance of CJTF compared to 18 baselines on CMU-MOSI with the evaluation metrics. The upward arrow indicates that the higher this indicator is, the better it is, and the downward arrow is the opposite.

Model		CMU-MOSI			
		Acc ↑	F1 ↑	MAE ↓	Corr ↑
Attention-based	MFN	77.4	77.3	0.965	0.632
	RAVEN	78.0	76.6	0.915	0.691
	MulT	81.5	80.6	0.861	0.711
	CIA	79.9	79.5	0.914	0.689
	GATE	82.9	80.6	-	-
	MPT	82.8	82.9	-	-
Graph-based	MMGraph	80.6	80.5	0.933	0.684
	GraphCAGE	82.1	82.1	0.933	0.684
Interaction-based	LMF	76.4	75.7	0.912	0.668
	MCTN	79.3	79.1	0.909	0.676
	TCSP	80.9	81.0	0.908	0.710
	MMLATCH	-	-	-	-
	PS-Mixer	82.1	82.1	0.794	0.748
Pre-Trained-based	MAG-BERT	82.5	82.6	0.731	0.789
	CubeMLP	85.6	85.5	0.770	0.767
	EMT	85.0	85.0	0.710	0.798
	AOBERT	85.2	85.4	0.856	0.700
	TETFN	86.1	86.1	0.717	0.800
CJTF(our)	LSTM+GCN	83.6	83.5	0.905	0.721
	BERT+GCN	**86.5**	**86.4**	**0.704**	**0.810**

information. This enables CJTF to extract both semantic and syntactic information simultaneously, resulting in improved extraction of textual modal features, thereby improving the performance of multimodal sentiment analysis.

Secondly, we compare our model with graph-based baselines. From table 2, we observed that CJTF still have superior performance. Although MMGraph and GraphCAGE utilize the method of constructing graphs for multimodal sentiment analysis, they first create nodes from sequence, then define edges based on these created nodes. All the nodes and edges compare the graph which contains sufficient information about long-range dependencies. This method of generating graphs has two drawbacks: firstly, the

Table 3: Performance of CJTF compared to 18 baselines on CMU-MOSEI with the evaluation metrics. The upward arrow indicates that the higher this indicator is, the better it is, and the downward arrow is the opposite.

Model		CMU-MOSEI			
		Acc ↑	F1 ↑	MAE ↓	Corr ↑
Attention-based	MFN	80.6	80.0	0.612	0.687
	RAVEN	79.1	79.5	0.614	0.662
	MulT	80.1	80.9	0.630	0.664
	CIA	80.4	78.2	0.683	0.594
	GATE	81.1	78.5	-	-
	MPT	82.6	82.8	-	-
Graph-based	MMGraph	81.4	81.7	0.608	0.675
	GraphCAGE	81.7	81.8	0.609	0.670
Interaction-based	LMF	82.0	82.2	0.623	0.677
	MCTN	80.8	80.6	0.611	0.670
	TCSP	82.8	82.7	0.576	0.715
	MMLATCH	82.8	82.9	0.582	0.704
	PS-Mixer	83.1	83.1	0.537	0.765
Pre-Trained-based	MAG-BERT	83.8	83.7	0.539	0.753
	CubeMLP	85.1	84.5	0.529	0.760
	EMT	86.0	86.0	0.527	0.774
	AOBERT	84.9	85.0	0.515	0.763
	TETFN	85.2	85.3	0.551	0.748
CJTF(our)	LSTM+GCN	84.3	84.1	0.536	0.757
	BERT+GCN	**86.1**	**86.1**	**0.513**	**0.788**

process of constructing dependency graphs fails to introduce syntactic relationships within the sentence, only considering the semantic relevance of the context. The second is the inability to capture the sentiment interaction between different words and determine the sentiment impact of different words. CJTF generates original syntactic trees to capture the syntactic dependency information and constructs the sentiment-interactive graphs by integrating the sentiment relations into the syntactic dependencies to fully utilize such sentiment associations.

Interaction-based baseline belong to compound models which contain different fusion components and strategies. These models often use different encoders to extract

features of different modalities, and then use interactive modeling to fuse modalities. This method can capture the interactive features of different modalities to improve the performance of the model to a certain extent. However, these models only focus on global information of all modalities and treat them equally, ignoring the influence of public features and private features with each modality to multimodal interaction. The public features enhance the textual semantic representation to make the model more robust and the private features complement the textual semantic representation to correctly predict sentiment. CJTF considers the cooperativity of public features and complementarity of private features simultaneously. It first utilizes a coordinated module with cross-modal mask attention to extract the non-textual public features that visual and acoustic towards textual to enhance the textual semantics. Then, a joint translation fusion module with cross-modal translation-aware mechanism is designed to complement textual semantics by translating non-textual private features. CTJF fuses the public and private features into a multimodal fusion layer to predict sentiment polarity. Therefore, CJTF achieves better accuracy by enhancement of public semantics and supplement of private semantics.

In addition, considering the excellent performance of the transformer based Pre-Trained Language Model in textual semantic extraction, we applied both LSTM+GCN and BERT+GCN for text encoding. We found that when we use LSTM+GCN to extract textual modal features, the model performs better than MAG-BERT on Acc and F1, but performs poorly on MAE and Corr. Compared with LSTM, BERT uses the multi-head self-attention mechanism which divides the vector space into multiple blocks to learn more abundant semantic representation in different spaces. And, residual connection better solves the problems of gradient vanishing and exploding, and layer normalization improves the generalization ability which makes the predicted value closer to the true value. Therefore, the MAG-BERT model is superior to our model. We believe that the BERT pre-trained model can extract dynamic semantics based on contextual context compared to traditional LSTM, making it easier to discover sentiment clues. When we use the BERT+GCN to encode textual modality, we found that the performance of the model is superior to all BERT-based baseline models. This indicates that fully considering syntactic information and sentiment interaction while obtaining context

Table 4: Ablation study of CJTF model on different components. For component part, we remove SIGCN module, coordinated module and joint translation module on CMU-MOSI respectively.

Model	CMU-MOSI			
	Acc ↑	F1 ↑	MAE ↓	Corr ↑
CJTF w/o SIGCN module	81.3	81.4	0.943	0.685
CJTF w/o coordinated module	81.8	81.7	0.919	0.693
CJTF w/o joint translation module	81.6	81.6	0.925	0.689
CJTF	**83.6**	**83.5**	**0.905**	**0.721**

Table 5: Ablation study of CJTF model on different components. For component part, we remove SIGCN module, coordinated module and joint translation module on CMU-MOSEI respectively.

Model	CMU-MOSEI			
	Acc ↑	F1 ↑	MAE ↓	Corr ↑
CJTF w/o SIGCN module	82.4	82.1	0.579	0.714
CJTF w/o coordinated module	83.1	83.2	0.5510	0.733
CJTF w/o joint translation module	82.9	83.0	0.566	0.729
CJTF	**84.3**	**84.1**	**0.536**	**0.757**

based dynamic semantics can help CJTF better achieve multimodal sentiment analysis.

4.5. Ablation study

We conduct an ablation study to evaluate the contribution of the SIGCN module, the coordinated module and the joint translation module, as shown in Table 4 and Table 5. CJTF w/o SIGCN indicates that we remove the SIGCN module, and the performance on four evaluation indicators has deteriorated. SIGCN module is used to capture the syntactic and semantic information of textual features, it will fail to capture the long-range dependencies of syntactics and sentiment interaction of semantics when we remove the SIGCN module. CJTF w/o coordinated module and CJTF w/o joint translation module indicates that we remove the coordinated module and the joint translation module respectively. We observe the performance of CJTF is further degraded, so that we believe the coordinated-joint translation component can capture non-textual public and private features to enhance and complement textual representation.

When we remove these two modules, CJTF will fuse textual, visual and acoustic features equally that ignores the enhancement and supplement of public and private features in different modalities to modal interaction. Therefore, we think the proposed modules can significantly improve the model performance.

4.6. Influence of the number of SIGCN layer

To explore the influence of SIGCN layers on performance, we evaluate the number of GCN layers ranging from 1 to 10 on CMU-MOSI and CMU-MOSEI. As shown in Fig. 3 and Fig. 4, With 1-layer SIGCN, the CJTF performs poor performance that the model cannot propagate far enough to capture the long-range syntactic dependencies and sentiment interactions. CJTF achieves the best performance in terms of accuracy, F1, MAE and Corr when the GCN layer is set to 2. In contrast, the performance of CJTF

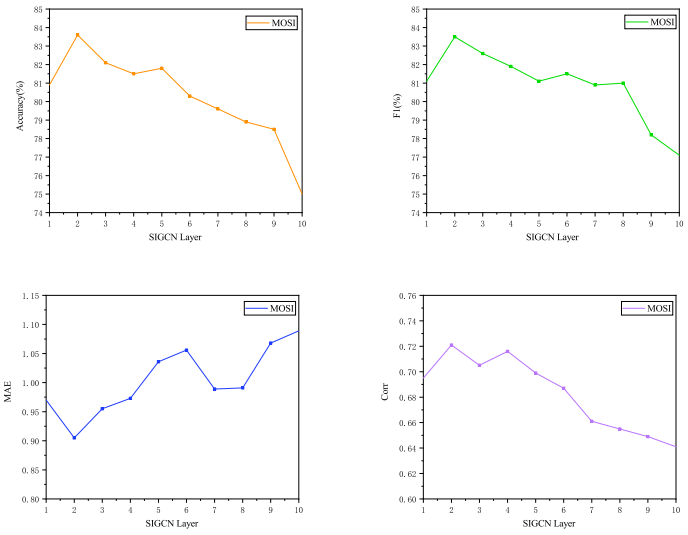


Figure 3: Influence of the numbers of SPGCN layers on CMU-MOSI with the evaluation metrics.

decreases with the increase of the layers because the phenomenon of over-smoothing [56] that makes the features of all nodes increasingly similar.

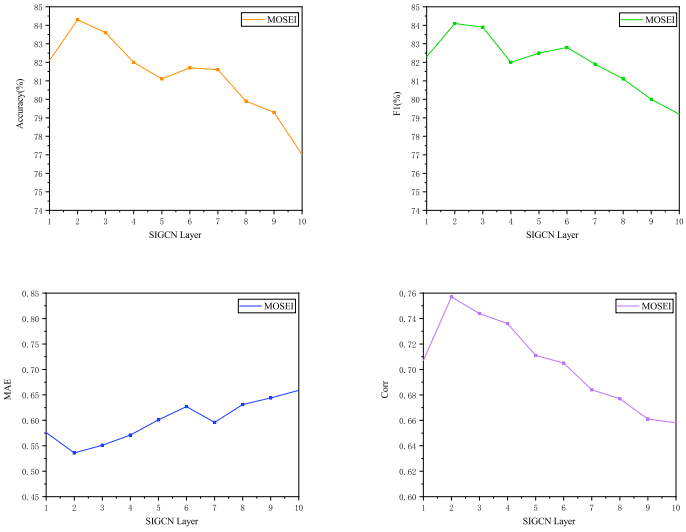


Figure 4: Influence of the numbers of SPGCN layers on CMU-MOSEI with the evaluation metrics.

4.7. Case study

To prove the sentiment interaction and non-textual public and private features are benefit to improve the performance of multimodal sentiment analysis, we conduct a case study with baselines as shown in Table 6. There are four utterances that the No.1 and 4 are positive and the No. 2 and 3 are negative. We observe that all models accurately predicted the sentiment polarity on No.1 and 2 due to these two utterances explicitly expressing the sentiment of opinion holders. In the third utterance, the TCSP and CJTF correctly predict the negative polarity but MCTN predict incorrectly. We think that the reason for this is that the MCTN fail to achieve the information interaction of different modalities. In the fourth utterance, the MCTN and TCSP all predict the negative polarity incorrectly. We believe that the reason is that these models pay too much attention to the negative word “hating”, and non-text visual and acoustic information fails to help the model focus on the positive word “love”. Compared with MCTN based on LSTM encoder, we think the MCTN pays more attention to word “love” and ignores the influence of word “but” and “bored”. We find CJTF focuses on “love”, “but” and “bored”, so that it can better capture context semantics and sentiment interaction relations

Table 6: Case study of proposed CJTF model compared with baselines, and different degrees of orange indicate the attention of the proposed CJTF to utterances.

No	Utterances	Sentiment			
		Actual	MCTN	TCSP	CJTF
1	The thing is its a very great translation of the book .	$Pos_{\checkmark}$	$Pos_{\checkmark}$	$Pos_{\checkmark}$	$Pos_{\checkmark}$
2	I thought that two hours was way too long for this movie .	$Neg_{\checkmark}$	$Neg_{\checkmark}$	$Neg_{\checkmark}$	$Neg_{\checkmark}$
3	Kids are gonna love the film , but for me i just a little bored .	$Neg_{\checkmark}$	$Pos_{\times}$	$Neg_{\checkmark}$	$Neg_{\checkmark}$
4	I don't know why people are hating on the film because I really do love it .	$Pos_{\checkmark}$	$Neg_{\times}$	$Neg_{\times}$	$Pos_{\checkmark}$

to make the model more excellent in semantic representation construction; Meanwhile, compared with the fusion methods of TCSP, TCSP thinks the word "don't" and "hating" are more important while CJTF focuses on word "love". We believe that CJTF with the cross-modal masked attention mechanism and cross-modal translation-aware mechanism can capture the semantic contribution of different regions of non-textual modalities towards text modality, thus achieving more accurate multimodal sentiment analysis.

4.8. Visualization

To demonstrate the interaction between different modalities of the CJTF, we visualize an utterance with textual and visual features from the CMU-MOSI dataset as shown in Fig. 5. We observe that the positive word "interesting" have a stronger correlation with visual features that corners of the mouth raised, and it is weakly related to these visual features that represent normal faces. We believe the reason for this is that the CJTF can enhance and supplement the textual information with different visual images via the different contributions of the public and private features to model.

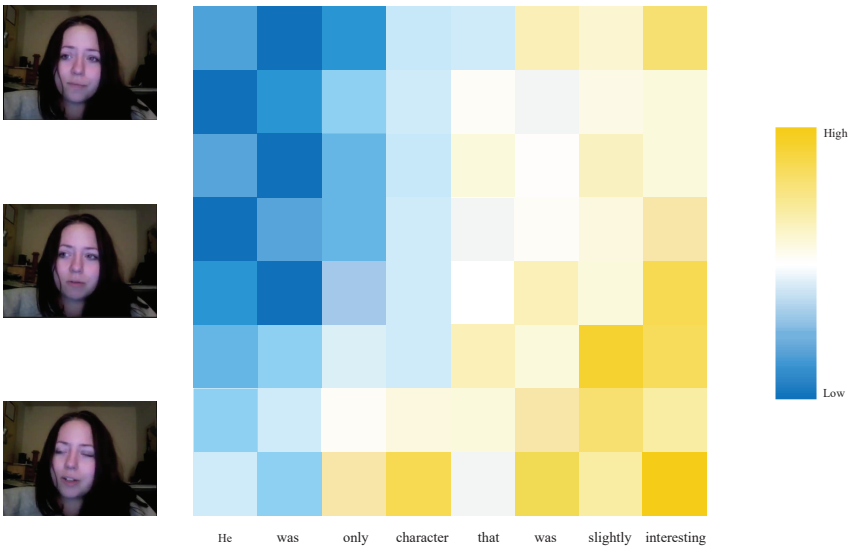


Figure 5: Visualization of the CJTF

5. Conclusion

This paper aims to explore new multimodal fusion methods to break the limitation of existing multimodal sentiment analysis methods. Based on the principles of consistency and complementarity, non-textual modalities have public regions that jointly express semantics and private regions that enjoy semantics individually. The public regions enhance the textual semantic representation to make the model more robust and the private regions complement the textual semantic representation to correctly predict sentiment. However, these methods only study the influence of different modal interactions on the performance, and fail to explain why visual or acoustic modalities can assist text semantics and which regions play a complementary role in text semantics. Meanwhile, there is a lack of exploration of sentiment interaction in multimodal interactive modeling. Therefore, compared with previous works, our method explores the sentiment interaction in textual semantics via graph structure, and clarifies the reason that visual and acoustic regions can enhance and supplement multimodal semantics. It is determined that there are two semantic features in the multimodal fusion processing:

the non-textual public semantic features and private semantic features. The non-textual public features enhance the textual semantics representation by calculating the visual and acoustic public semantic contribution towards textual features, and the non-textual private features supplement the textual semantics representation by translating the visual and acoustic private semantic features towards textual features.

In this paper, we propose a coordinated-joint translation fusion (CJTF) framework for multimodal sentiment analysis, which integrates public and private features of non-textual information to enhance and complement textual features containing semantics and syntactics from sentiment-interactive GCN. Specially, we design a sentiment-interactive GCN to extract the long-distance dependencies of syntactics and sentiment interaction of semantics to address the issue that models mistakenly identify syntactically irrelevant contextual words as clues for judging sentiment. Furthermore, the coordinated-joint translation module is designed to enhance and complement textual features by calculating the public semantic contribution and translating visual and acoustic private semantics features towards textual features. In addition, the proposed model is suitable for semantic extraction using multiple models, including LSTM and BERT. Experimental results on two public datasets CMU-MOSI and CMU-MOSEI illustrate that our proposed model outperforms all advanced baselines and verify the effectiveness of our model. However, our model may fail to capture the complex relations of internal and cross different modalities. Therefore, we would like to utilize graph contrastive learning method to generate the graph strategy to capture sentiment relations in the future work.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No.61877050 and 2022 Yulin Science and Technology Plan Project "Research on Learning Behavior Analysis and Emotion Recognition Technology in Large-scale Online Education" under Grant No.CXY-2022-177

References

[1] Akhtar, M.S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., Bhattacharyya, P., 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 370–379.

[2] Al-Ayyoub, M., Khamaiseh, A.A., Jararweh, Y., Al-Kabi, M.N., 2019. A comprehensive survey of arabic sentiment analysis. Information processing & management 56, 320–342.

[3] Baltrušaitis, T., Ahuja, C., Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41, 423–443.

[4] Cai, C., He, Y., Sun, L., Lian, Z., Liu, B., Tao, J., Xu, M., Wang, K., 2021. Multimodal sentiment analysis based on recurrent neural network and multimodal attention, in: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, pp. 61–67.

[5] Cai, H., Tu, Y., Zhou, X., Yu, J., Xia, R., 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network, in: Proceedings of the 28th international conference on computational linguistics, pp. 833–843.

[6] Chauhan, D.S., Akhtar, M.S., Ekbal, A., Bhattacharyya, P., 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5647–5657.

[7] Chen, D., Su, W., Wu, P., Hua, B., 2023. Joint multimodal sentiment analysis based on information relevance. Information Processing & Management 60, 103193.

[8] Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P., 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM international conference on multimodal interaction, pp. 163–171.

[9] Cheng, J., Fostiropoulos, I., Boehm, B., Soleymani, M., 2021. Multimodal phased transformer for sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2447–2458.

[10] Datcu, D., Rothkrantz, L.J., 2011. Emotion recognition using bimodal data fusion, in: Proceedings of the 12th International Conference on Computer Systems and Technologies, pp. 122–128.

[11] Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S., 2014. Covarep—a collaborative voice analysis repository for speech technologies, in: 2014 ieee international conference on acoustics, speech and signal processing (icassp), IEEE. pp. 960–964.

[12] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

[13] Ekman, P., 1992. An argument for basic emotions. Cognition & emotion 6, 169–200.

[14] El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern recognition 44, 572–587.

[15] Feldman, R., 2013. Techniques and applications for sentiment analysis. Communications of the ACM 56, 82–89.

[16] Ghorbanali, A., Sohrabi, M.K., Yaghmaee, F., 2022. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. Information Processing & Management 59, 102929.

[17] Hazarika, D., Zimmermann, R., Poria, S., 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM international conference on multimedia, pp. 1122–1131.

[18] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

[19] Kaur, R., Kautish, S., 2022. Multimodal sentiment analysis: A survey and comparison. Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines , 1846–1870.

[20] Kim, K., Park, S., 2023. Aobert: All-modalities-in-one bert for multimodal sentiment analysis. Information Fusion 92, 37–45.

[21] Kumar, A., Vepa, J., 2020. Gated mechanism for attention based multi modal sentiment analysis, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4477–4481.

[22] Liang, B., Lou, C., Li, X., Gui, L., Yang, M., Xu, R., 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4707–4715.

[23] Liang, B., Lou, C., Li, X., Yang, M., Gui, L., He, Y., Pei, W., Xu, R., 2022a. Multi-modal sarcasm detection via cross-modal graph convolutional network, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1767–1777.

[24] Liang, B., Su, H., Gui, L., Cambria, E., Xu, R., 2022b. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. Knowledge-Based Systems 235, 107643.

[25] Lin, H., Zhang, P., Ling, J., Yang, Z., Lee, L.K., Liu, W., 2023. Ps-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. Information Processing & Management 60, 103229.

[26] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5, 1–167.

[27] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A.B., Morency, L.P., 2018. Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2247–2256.

[28] Lu, G., Li, J., Wei, J., 2022. Aspect sentiment analysis with heterogeneous graph neural networks. Information Processing & Management 59, 102953.

[29] Mai, S., Xing, S., He, J., Zeng, Y., Hu, H., 2020. Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. arXiv preprint arXiv:2011.13572 .

[30] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S., 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based systems 161, 124–133.

[31] Morency, L.P., Mihalcea, R., Doshi, P., 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web, in: Proceedings of the 13th international conference on multimodal interfaces, pp. 169–176.

[32] Paraskevopoulos, G., Georgiou, E., Potamianos, A., 2022. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4573–4577.

[33] Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

[34] Pérez-Rosas, V., Mihalcea, R., Morency, L.P., 2013. Utterance-level multimodal sentiment analysis, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 973–982.

[35] Pham, H., Liang, P.P., Manzini, T., Morency, L.P., Póczos, B., 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6892–6899.

[36] Poria, S., Cambria, E., Gelbukh, A., 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 2539–2544.

[37] Poria, S., Chaturvedi, I., Cambria, E., Hussain, A., 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th international conference on data mining (ICDM), IEEE. pp. 439–448.

[38] Rahman, W., Hasan, M.K., Lee, S., Zadeh, A.B., Mao, C., Morency, L.P., Hoque, E., 2020. Integrating multimodal information in large pretrained transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359–2369.

[39] Revina, I.M., Emmanuel, W.S., 2021. A survey on human face expression recognition techniques. Journal of King Saud University-Computer and Information Sciences 33, 619–628.

[40] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.F., Pantic, M., 2017. A survey of multimodal sentiment analysis. Image and Vision Computing 65, 3–14.

[41] Sun, H., Wang, H., Liu, J., Chen, Y.W., Lin, L., 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 3722–3729.

[42] Sun, L., Lian, Z., Liu, B., Tao, J., 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Transactions on Affective Computing .

[43] Sun, Z., Sarma, P., Sethares, W., Liang, Y., 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8992–8999.

[44] Tang, J., Li, K., Jin, X., Cichocki, A., Zhao, Q., Kong, W., 2021. Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5301–5311.

[45] Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6558–6569.

[46] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. stat 1050, 20.

[47] Verma, S., Wang, J., Ge, Z., Shen, R., Jin, F., Wang, Y., Chen, F., Liu, W., 2020. Deep-hoseq: Deep higher order sequence fusion for multimodal sentiment analysis, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE. pp. 561–570.

[48] Wang, D., Guo, X., Tian, Y., Liu, J., He, L., Luo, X., 2023. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition 136, 109259.

[49] Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.P., 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7216–7223.

[50] Wang, Y., Yang, N., Miao, D., Chen, Q., 2022. Dual-channel and multi-granularity gated graph attention network for aspect-based sentiment analysis. Applied Intelligence , 1–13.

[51] Wang, Z., Wan, Z., Wan, X., 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis, in: Proceedings of The Web Conference 2020, pp. 2514–2520.

[52] Wimmer, M., Schuller, B., Arsic, D., Radig, B., Rigoll, G., 2008. Low-level fusion of audio and video feature for multi-modal emotion recognition, in: Proc. 3rd Int. Conf. on Computer Vision Theory and Applications VISAPP, Funchal, Madeira, Portugal, pp. 145–151.

[53] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.P., 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems 28, 46–53.

[54] Wu, J., Mai, S., Hu, H., 2021a. Graph capsule aggregation for unaligned multimodal sequences, in: Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 521–529.

[55] Wu, Y., Lin, Z., Zhao, Y., Qin, B., Zhu, L.N., 2021b. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4730–4738.

[56] Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.i., Jegelka, S., 2018. Representation learning on graphs with jumping knowledge networks, in: International conference on machine learning, PMLR. pp. 5453–5462.

[57] Xu, N., Mao, W., 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2399–2402.

[58] Yang, Y., Sun, X., Lu, Q., Sutcliffe, R., Feng, J., 2023. A sentiment and syntactic-aware graph convolutional network for aspect-level sentiment classification, in:

ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

[59] You, Q., Luo, J., Jin, H., Yang, J., 2015. Joint visual-textual sentiment analysis with deep neural networks, in: Proceedings of the 23rd ACM international conference on Multimedia, pp. 1071–1074.

[60] Yu, B., Zhang, S., 2023. A novel weight-oriented graph convolutional network for aspect-based sentiment analysis. The Journal of Supercomputing 79, 947–972.

[61] Yu, W., Xu, H., Yuan, Z., Wu, J., 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10790–10797.

[62] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P., 2017. Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1103–1114.

[63] Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P., 2018a. Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI conference on artificial intelligence.

[64] Zadeh, A., Zellers, R., Pincus, E., Morency, L.P., 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems 31, 82–88.

[65] Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P., 2018b. 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), pp. 2236–2246.

[66] Zhang, C., Li, Q., Song, D., 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4568–4578.

[67] Zhang, C., Yang, Z., He, X., Deng, L., 2020. Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing 14, 478–493.

[68] Zhao, L., Liu, Y., Zhang, M., Guo, T., Chen, L., 2021. Modeling label-wise syntax for fine-grained sentiment analysis of reviews via memory-based neural model. Information Processing & Management 58, 102641.

[69] Zhou, J., Huang, J.X., Hu, Q.V., He, L., 2020. Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. Knowledge-Based Systems 205, 106292.

[70] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J., 2022. Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia .