



A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis

Tong Zhao¹ · Junjie Peng^{1,2} · Yansong Huang¹ · Lan Wang¹ · Huiran Zhang¹ · Zesu Cai³

Accepted: 1 November 2023 / Published online: 18 November 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Multimodal sentiment analysis leverages various modalities, including text, audio, and video, to determine human sentiment tendencies, which holds significance in fields such as intention understanding and opinion analysis. However, there are two critical challenges in multimodal sentiment analysis: one is how to effectively extract and integrate information from various modalities, which is important for reducing the heterogeneity gap among modalities; the other is how to overcome the problem of information forgetting while modelling long sequences, which leads to significant information loss and adversely affect the fusion performance of modalities. Based on the above issues, this paper proposes a multimodal heterogeneity fusion network based on graph convolutional neural networks (HFNGC). A shared convolutional aggregation mechanism is used to overcome the semantic gap among modalities and reduce the noise effect caused by modality heterogeneity. In addition, the model applies Dynamic Routing to convert modality features into graph structures. By learning semantic information in the graph representation space, our model can improve the capability of remote-dependent learning. Furthermore, the model integrates complementary information among modalities and explores the intra- and inter-modal interactions during the modality fusion stage. To validate the effectiveness of our model, we conduct experiments on two benchmark datasets. The experimental results demonstrate that our method outperforms the existing methods, exhibiting strong generalisation capability and high competitiveness.

Keywords Sentiment analysis · Heterogeneity · Graph convolution · Information fusion

1 Introduction

Social media platforms like TikTok and YouTube have made videos increasingly popular for sharing lifestyles and expressing intentions. Consequently, there has been an exponential growth of multimodal data. Multimodal data is preferable to unimodal data as it provides more extensive information for model training. As a result, sentiment analysis using multimodal data has become a research area of great interest in recent years.

Compared to traditional sentiment analysis tasks, multimodal sentiment analysis aims to integrate multiple sources of information, such as text, vision and acoustics, for task learning. In recent years, multimodal sentiment analysis becomes increasingly important for machine perception and

intention understanding. Many studies in this area have explored various approaches, including tensor operations for multimodal fusion [1, 2] and attention mechanisms for feature learning [3, 4]. Additionally, some researchers have utilized multi-task learning frameworks [5] to improve model robustness and generalization, which has proven effective in achieving good performance.

To efficiently acquire temporal features when modelling contextual information, RNN-based approaches are widely studied. However, RNN and its variants rely on recursion to model sequence data, which can result in long inference duration and issues with gradient explosion and vanishing. Additionally, recursive approaches cannot learn feature dependencies over long distances well. The above phenomenon is particularly problematic for multimodal sequences, which tend to be lengthy and where the impact of long-term dependencies on model performance is significant. As a result, it is crucial to explore alternative approaches for learning long-term dependencies in multimodal sentiment analysis.

✉ Junjie Peng
jjie.peng@shu.edu.cn

Extended author information available on the last page of the article

In addition, existing studies often ignore the differences in information density among modalities when performing multimodal information integration. Specifically, the text modality contains a high intensity of helpful information. In contrast, the visual and acoustic modalities have repetitive and redundant information with lower densities of usable information. As a result, the visual and acoustic modalities cannot achieve the same-level semantic performance as the text modality does in the process of feature extraction, thus affecting the performance of subsequent multimodal fusion. For example, in Fig. 1, both text examples contain important adjectives and adverbs. However, from the sequence of video screenshots, the speaker's facial expression does not change much, and many frames are repetitive. The acoustic waveform map also shows multiple large fluctuations in unimportant positions. In such cases, the visual modality carries plenty of redundant information that cannot provide a useful reference value for model learning. Additionally, the acoustic modality shows large interest in multiple unimportant positions, which can be misleading during model capturing information and make the model mistakenly believe that certain positions contain important information.

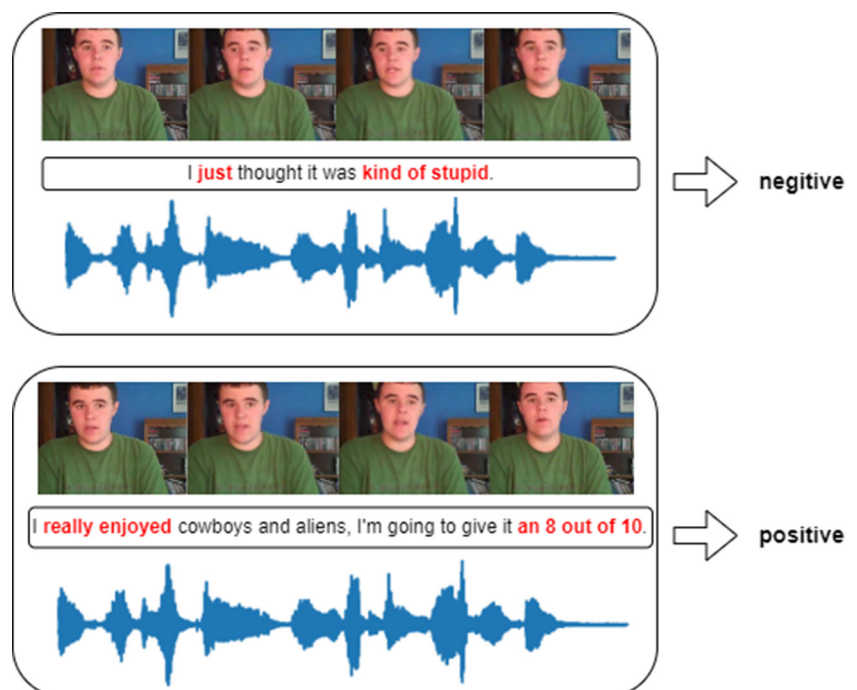
To solve the above problems, we design a graph convolutional-based multimodal heterogeneity fusion network for multimodal sentiment analysis, in which we adopt a convolutional aggregation mechanism to overcome the semantic gap among modalities and reduce the influence of modal heterogeneity. In addition, to address the context propagation problem found in RNN-based approaches, we employ

Dynamic Routing to project feature encoding into the graph space. In this way, we can leverage highly expressive graph structures and graph convolutional neural networks to learn feature dependencies. Furthermore, we utilise a cross-modal attention mechanism that employs the Transformer to enhance the critical semantic features of each modality and learn complementary information between them, resulting in effective multimodal data fusion.

The main contributions of this paper are as follows.

- We propose a multimodal heterogeneity fusion model based on graph convolution (HFNGC). By overcoming the semantic gap between modalities, the model combines the information of different densities and mitigates the effect of length when modelling serialisation using graph structure with better long-term dependency learning capability.
- We adopt a convolutional aggregation mechanism to enhance the semantic representation of visual and acoustic modalities. By enhancing the feature of low information density modalities, the convolutional aggregation mechanism mitigates the effect of noise between modalities due to differences in information density, and improves the performance of subsequent modal fusion.
- By projecting the modal feature representation into the graph space and using Dynamic Routing for graph construction, the model solves the problem of information forgetting when modelling long sequences.
- We conduct extensive experiments on two public datasets, and the results show that our method has significant

Fig. 1 An illustration of the heterogeneity gap among modalities



advantages and better generalisation than existing multimodal sentiment analysis methods.

2 Related work

2.1 Multimodal sentiment analysis

Multimodal sentiment analysis relies on multiple modalities, such as text, visual, and acoustic, to understand human emotions and intentions. One of the essential challenges in this field is effectively fusing multiple modalities and improving information quality. Existing modality fusion methods in this field can be broadly classified into early fusion and late fusion [6]. Early fusion [7] uses simple concatenation, which is easy to operate. Late fusion [8] builds independent training models for each modality and is robust. However, both approaches ignore the complementarity between modalities and are susceptible to losing important information. In recent years, researchers have attempted to establish more efficient modality fusion mechanisms for multimodal sentiment analysis. RNN-based fusion methods [9] have become common approaches in modality fusion. Zadeh et al. [1] proposed Tensor Fusion Network (TFN), which aims to compute feature representations between modalities with the help of the Cartesian product. Liu et al. [2] proposed Low-rank Multimodal Fusion Network (LMF), which improves the computation of tensor fusion networks and reduces memory consumption through a low-rank decomposition factor. Mai et al. [10] proposed Multi-Fusion Residual Memory Network and used residual memory networks to learn the temporal information within the modal. Basiri et al. [11] propose an Attention-based Bidirectional CNN-RNN Deep Model for information extraction by acquiring bi-directional temporal flow information.

In order to capture inter-modal interactions and intra-modal feature dependencies, researchers have paid attention to the superiority of the attention mechanism and applied it widely in the field of multimodal fusion. Wu et al. [12] designed a bimodal information enhancement framework, which uses multi-head attention to explore the interaction between pairs of modalities. Wang et al. [13] proposed a text-enhanced transformer fusion network that uses text-guided cross-modal mapping to capture modal features and preserve differentiated features of the modalities themselves. Xue et al. [14] designed a multi-level attention map network, which utilises multi-level attention graphs for noise reduction and feature enhancement; meanwhile, the network can efficiently extract complex interaction information between attention maps. In order to capture the fine-grained mapping between modalities, Zhu et al. [15] proposed an image-text interaction network based on a cross-modal attention mechanism and designed a cross-modal gating module to achieve deep interaction of multimodal features. Zhang et al. [16] proposed

a long short-term memory network based on the information block multi-head subspace, which utilises the multi-head attention mechanism for parallel multi-space feature extraction.

In recent years, the multi-task learning framework has gained significant attention as it can enhance the model's robustness and reduce the risk of overfitting during training. Akhtar et al. [5] designed a joint learning task framework for sentiment analysis and emotion recognition, claiming that the framework can achieve excellent performance. Peng et al. [17] proposed a multi-stage network based on fine-grained modal labelling, which utilises seven different modality granularities as independent tasks for multi-stage training. Chen et al. [18] designed a weighted attention mechanism based on an auxiliary task of sentiment prediction, which utilises unimodal sub-tasks to assist modality-specific representation learning.

2.2 Graph neural networks

Due to graph structures' high expressiveness and temporal sensitivity, deep learning based on graph structures has attracted widespread attention from researchers. As an effective measure for learning graph structures, the graph neural network has performed well in many recent studies. Wu et al. [19] combined capsule networks [20] and graph data to model multimodal sequences and improved the graph aggregation during training to obtain high-quality feature information. Yang et al. [21] proposed Modal-Temporal Attention Graph (MTAG), which they claim can effectively capture cross-modal and temporal information in the graph. Yang et al. [22] noticed that multi-modality has specific global characteristics in the process of sentiment expression, so they designed multi-channel graph neural networks to capture the global feature representation under multi-channel and introduced multi-head attention mechanisms to achieve multimodal deep interactive learning. Zeng et al. [23] constructed a negative sentiment recognition model based on graph convolutional neural networks and integrated learning, which can effectively identify the negative public opinion in public health events. Zhang et al. [24] argued that sentiment classification affects the effectiveness of traffic event detection tasks; therefore, they combined sentiment knowledge with traffic event detection and constructed a multimodal graph with text and vision as nodes and explored the association between emotions and traffic events through the multimodal graph. Noting the importance of long-range sentiment dependencies and syntactic constraints, Lu et al. [25] proposed a gated graph convolutional network for aspect-level sentiment analysis by combining the gating mechanism as well as the graph convolutional network, which they claimed can guide aspect-level information encoding and can achieve superior performance.

3 Graph convolution-based heterogeneous fusion network

This section introduces the proposed model (HFNGC) in detail. The architecture of HFNGC is shown in Fig. 2, which consists of four main modules: modality encoding layer, graph construction and convolution layer, multimodal fusion layer and prediction layer. The modal encoding layer's primary purpose is to extract each modality's semantic features. We note inconsistencies in the semantic representation level between modalities due to differences in density. Specifically, the textual modality is a human-specific modality which contains dense and highly semantic information. In contrast, the visual and acoustic modalities contain more fine-grained information, which is primarily redundant, resulting in a lower-level representation of information in the visual and acoustic modalities. Therefore, we adopt a convolutional aggregation module in the modal encoding layer to overcome the semantic gap between the modalities. The graph construction and convolution layer employs Capsule Network to project the feature encoding extracted from the modal encoding layer into the graph feature space for graph construction and uses the GCN to learn the semantic information contained in the graph. The task of the multimodal fusion layer is to extract depth features within modalities and to learn information from inter-modal interactions. Finally, the feature representations learned in the multimodal fusion layer are sent to the deep neural network to accomplish the prediction task in the prediction layer.

3.1 Modality encoding layer

Given a multimodal sequence $x = \{x_t, x_a, x_v\}$, we let $x_m \in \mathbb{R}^{l_m \times d_m}$, $m \in \{t, v, a\}$ denote the original unimodal

sequence extracted from video, where t, v, a denotes the textual, visual and acoustic modality, respectively, l_m denotes the sequence length and d_m denotes the feature vector dimension of modality m . Firstly, we perform feature extraction on the original unimodal sequence x . Specifically, for the text modality, we use *Bert* as the text modality encoder considering the superior performance of large pre-trained models of Bert on text [26]. For visual and acoustic modalities, we adopt *BiLSTM* for feature extraction to capture temporal and contextual dependencies within the modality, as shown in (1), (2) and (3).

$$u_t = BERT(x_t; \theta_t) \quad (1)$$

$$u_a = BiLSTM(x_a; \theta_a) \quad (2)$$

$$u_v = BiLSTM(x_v; \theta_v) \quad (3)$$

Where θ_t, θ_a and θ_v are the learnable parameters in feature extractors for text, acoustic and visual modality, respectively. u_t is the output representation of the first-word vector of the final *BERT*. u_a and u_v are the outputs of the final *BiLSTM*.

Compared to text modality, which is high-level semantic information of human expression, visual and acoustic modalities contain much low-level semantic information, which can be seen as redundant noise. Therefore, rather than directly exploring the correlation between text and either acoustic or visual modalities, we focus on the differences in information density among different modalities. To achieve better inter-modality information integration, we propose a convolutional aggregation module. The module uses convolution to learn information from visual and acoustic modalities and

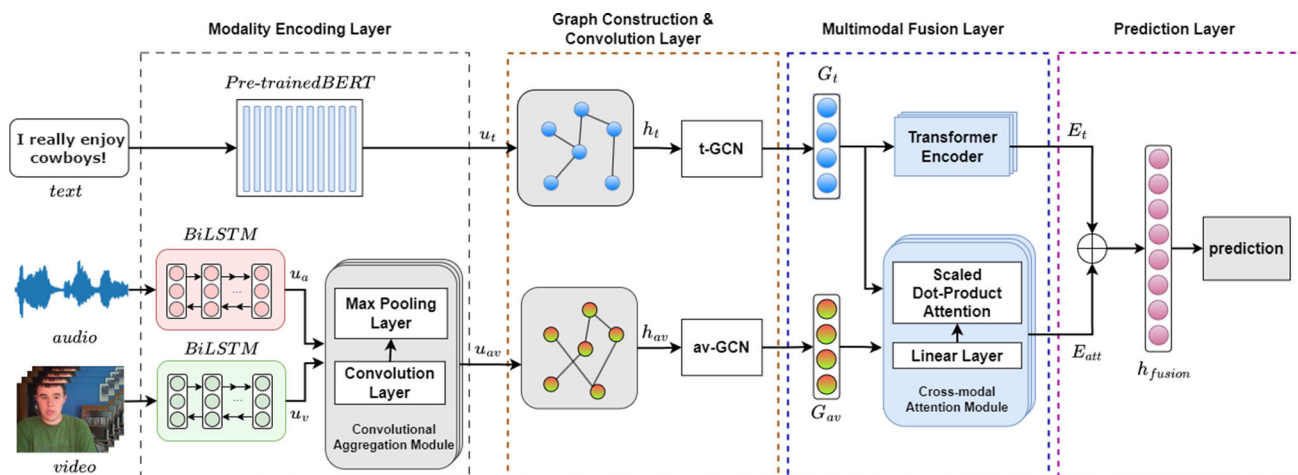


Fig. 2 Overall architecture of HFNGC. \oplus represents concatenation

shares parameters during modal learning to improve the quality of semantic features in feature encoding, as shown in (4)–(5).

$$u_{a'} = \text{maxpooling}(\text{CNN}(u_a; \theta_c)) \quad (4)$$

$$u_{av} = \text{maxpooling}(\text{CNN}(u_v; \theta_c)) \quad (5)$$

Where u_{av} and $u_{a'}$ denote the final output and intermediate results of the convolutional aggregation module, respectively. θ_c denotes the learnable parameter in the convolutional neural network.

3.2 Construction and convolution of the modality feature graph

In previous work, RNN-based structures have been used to investigate sequential semantics in modelling feature embeddings. However, due to the slow training speed and inability to capture long-range dependencies, RNN is unsuitable for modelling long-term sequences. Additionally, the recurrent nature of RNN makes it susceptible to gradient explosion or gradient vanishing, which demands an alternative mechanism to tackle this issue. Fortunately, the graph data structure allows each node to be directly connected to its associated node, making it immune to information forgetting caused by the length of the sequence. Furthermore, the graph convolutional network can simultaneously compute all the information in the graph, which can solve the problem of gradient explosion or gradient vanishing. Given these reasons, we project the modal sequence features into the graph space to explore the feature dependencies within the modalities.

Node generator We note that Dynamic Routing [19, 20] allows each node in the graph to perceive the information

contained in other time steps without information forgetting problems caused by long sequences and can maximise the preservation of valuable feature information. Therefore, in order to adequately learn the inter-node dependencies when constructing the modality feature graph, we employ Dynamic Routing to perform the node computation. Shown as the Node Generator in Fig. 3, the Node Generator takes the modality-encoded vector u_n as the initial input vector, $n \in \{t, av\}$, where t denotes the textual modality and av denotes the combination of visual and acoustic modalities. The Node Generator computes the corresponding shallow capsule vector u_n based on the input vectors. Subsequently, u_n determines the percentage of its own feature output to the high-level capsule vector s_n based on the routing weight c , which generates the Node Generator's output h_n after a nonlinear transformation. In addition, the routing weight c in the Node Generator calculates the similarity between the final output and the shallow capsule vector in each iteration round and updates itself. The specific formulas are shown in (6)–(9).

$$\hat{u}_{j|i}^n = W_{ij}^n u_i^n \quad (6)$$

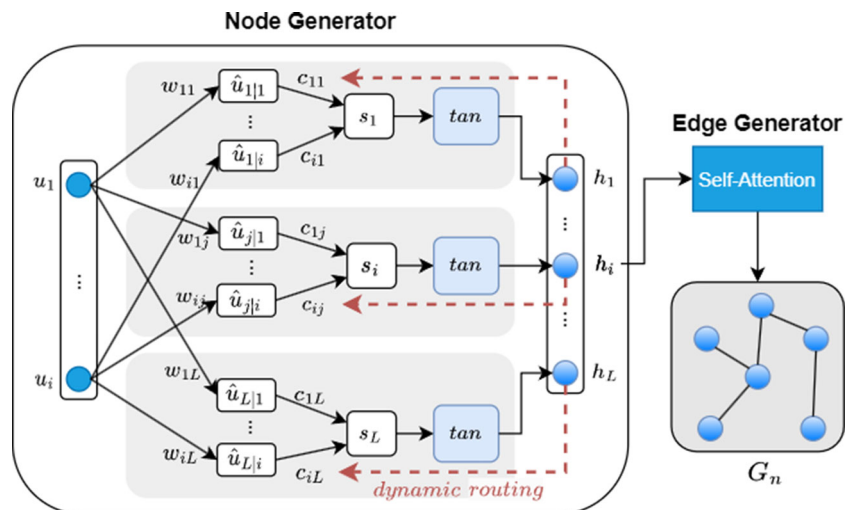
$$s_j^n = \sum_{i=1} c_{ij}^n \hat{u}_{j|i}^n \quad (7)$$

$$c_{ij}^n = \frac{\exp(b_{ij}^n)}{\sum_k \exp(b_{ik}^n)} \quad (8)$$

$$b_{ij}^n \leftarrow b_{ij}^n + b_{j|i}^n \odot h_j^n \quad (9)$$

Where W_{ij}^n , b_{ij}^n and c_{ij}^n are all learnable parameters in the dynamic routing calculation process. With the above node construction process, we can capture the node representation of the modality feature graph $H_n = \{h_1^n, h_2^n, \dots, h_L^n\}$, where $1 \leq i \leq L$ and L denotes the number of nodes.

Fig. 3 The process of graph construction. Graph construction has two main parts: node generator and edge generator. The output of Modality Encoding Layer is used as the input for Node Generator. Edge Generator constructs edges using the self-attention mechanism. In addition, before Node Generator outputs the node features, the routes need to undergo r rounds of dynamic updates



Edge generator After capturing the node representations of the modality feature graph, we explore the weight relationships of edges in the modality feature graph using the self-attention mechanism. As shown in (10), we utilize the node representation h_n learned through Dynamic Routing as the input of the edge generator. Subsequently, we focus the edge construction process on the focal information of the modality feature graph using the self-attention mechanism, ultimately generating the edge representation of the modality feature graph.

$$E_n = \text{Relu}\left(\frac{(W_{Q_i}^n H_n)(W_{K_i}^n H_n)^T}{\sqrt{d_n}}\right) \quad (10)$$

Where E_n denotes the adjacency matrix of the modality feature graph, $\text{Relu}(\cdot)$ is a nonlinear activation function, $W_{Q_i}^n \in \mathbb{R}^{d_n \times d_q}$ and $W_{K_i}^n \in \mathbb{R}^{d_n \times d_k}$ are learnable projection matrices, $n \in \{t, av\}$. Through the process described above, we can acquire the modality feature graph $G_n = (H_n, E_n)$, $n \in \{t, av\}$, where E_n is the edge weight representation of the modality feature graph and H_n is the set of node representations of the modality feature graph.

Graph convolutional network (GCN) Graph Convolutional Networks (GCN) are crucial for graph representation learning. GCN allows for simultaneous processing of complex data, capturing essential global information, and has high efficacy in learning about nodes and edges. Therefore, we select the GCN to perform deep learning on the acquired modality feature graph. As shown in Graph Construction & Convolution Layer in Fig. 2, we adopt the modality feature graphs G_t and G_{av} as the inputs of the graph convolutional network, where t-GCN represents the graph convolutional network with input G_t and av-GCN represents the graph convolution network with input G_{av} . The specific equation is shown below.

$$G_n^l = \text{Relu}(\tilde{E}_n G_n^{l-1} W_n^l + b_n^l) \quad (11)$$

Where G_n^l denotes the output of the l -th layer GCN, $\text{Relu}(\cdot)$ denotes the nonlinear activation function, $n \in \{t, av\}$, $\tilde{E}_n = D_n^{-\frac{1}{2}} E_n D_n^{-\frac{1}{2}}$, and D_n is the degree matrix of E_n . In addition, W_n^l and b_n^l are the learnable parameters of the l -th layer GCN, and the initial input to the GCN is the set of node representations of the modality feature graph, $G_n^0 = H_n$.

3.3 Multimodal fusion layer

The primary purpose of this layer is to integrate the feature information learned by the graph convolution network and achieve feature fusion as well as information reinforcement.

The layer has two main parts: cross-modal interaction learning and texture sequence modelling.

As shown in (12)-(17), we define a set of *query*, *key* and *value* to encode features on the input sequence and use a multi-head attention mechanism to achieve cross-modal interaction learning.

$$Q_{av}^i = W_{Q_i}^{av} G_{av} \quad (12)$$

$$K_t^i = W_{K_i}^t G_t \quad (13)$$

$$V_t^i = W_{V_i}^t G_t \quad (14)$$

$$\text{Attention}(Q_{av}, K_t, V_t) = \text{softmax}\left(\frac{Q_{av} K_t^T}{\sqrt{d_t}}\right) V_t \quad (15)$$

$$\text{head}_i = \text{Attention}(Q_{av}^i, K_t^i, V_t^i) \quad (16)$$

$$\begin{aligned} E_{att} &= \text{MultiHeadAttention}(Q_{av}, K_t, V_t) \\ &= \text{Concat}(\text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot W^o, H_{av}) \end{aligned} \quad (17)$$

Where E_{att} denotes the output of the cross-modal attention module, $W_{Q_i}^{av} \in \mathbb{R}^{d_{av} \times d_q}$, $W_{K_i}^t \in \mathbb{R}^{d_t \times d_k}$ and $W_{V_i}^t \in \mathbb{R}^{d_t \times d_v}$ are learnable parameter matrices, $d_q = d_k = d_v = d_t/h$. h is the number of heads, $1 \leq i \leq h$, $\text{Concat}(\cdot)$ is a concatenation operation. Q_{av} , K_t and V_t are a set of *query*, *key* and *value* that we define.

In addition, we consider that textual modality contains rich semantic information and often plays a dominant role in the process of modality fusion, so in order to capture the correlations within the textual modality and obtain richer semantic information, we take G_t which is learned by the graph convolution network as input and introduce the Transformer encoder to perform deep modelling of the textual modality, as shown in (18)-(21).

$$U_t' = \text{MultiHeadAttention}(Q_t, K_t, V_t) \quad (18)$$

$$U_t = \text{LayerNorm}(G_t + U_t') \quad (19)$$

$$E_t' = \text{Relu}(U_t) \quad (20)$$

$$E_t = \text{LayerNorm}(U_t + E_t') \quad (21)$$

Where E_t denotes the final output of the textual sequence modelling module, Q_t , K_t , V_t are a set of *query*, *key* and

value that we define, $Relu(\cdot)$ denotes the nonlinear activation function, and $LayerNorm(\cdot)$ denotes the normalisation operation.

3.4 Prediction layer

We use $Concat(\cdot)$ to combine the output of the multimodal fusion layer as the input of the prediction layer and complete the prediction task in the deep neural network, as shown in (22)-(23).

$$h_{fusion} = Concat(E_{att}, E_t) \quad (22)$$

$$\hat{y}_{fusion} = W_{fusion}h_{fusion} + b_{fusion} \quad (23)$$

Where \hat{y}_{fusion} denotes the final output of the prediction layer, h_{fusion} is the input to the prediction layer, and both W_{fusion} and b_{fusion} are learnable parameters.

4 Experiments and analysis of results

4.1 Experimental setup

4.1.1 Datasets

We evaluate our proposed method on two public datasets typically used in multimodal sentiment analysis, CMU-MOSI and CMU-MOSEI. Specifically, we train the model separately using the CMU-MOSI and CMU-MOSEI datasets and then apply the models obtained from each dataset to their respective test sets. The basic statistical information of the datasets is shown in Table 1.

CMU-MOSI The CMU-MOSI dataset [27] is a multimodal sentiment analysis dataset extracted from 93 monologue videos on YouTube. The dataset contains 2199 clips, of which the training, validation and test sets contain 1284, 229 and 686 video clips, respectively, and all of these clips contain sentiment ranging from [-3, +3] in intensity ratings, where +3 is the strongest positive emotion, and -3 is the strongest negative emotion.

CMU-MOSEI The CMU-MOSEI dataset [28] is a multimodal sentiment analysis dataset extracted from 5,000 videos on YouTube, which is a further extension of CMU-MOSI. The dataset contains 22,856 video clips, all of which are

annotated, and the training set, validation set, and test set contain 16,326, 1,871 and 4,659 video clips, respectively.

4.1.2 Baselines

In order to evaluate the performance of the proposed model, we compare the results with several advanced baselines:

TFN: TFN [1] uses the Cartesian product to explore the interactions between multiple modalities.

LMF: LMF [2] is a further refinement of TFN, which uses low-rank variants to improve the quality of model learning.

MuT: MuT [29] focuses on the interactions between multimodal sequences across different time steps and uses an improved Transformer architecture to integrate features on unaligned multimodal sequences.

MTAG: MTAG [21] designs a novel graph fusion operation which captures vital information in the modality-temporal graph using techniques such as dynamic pruning.

BIMHA: BIMHA [12] explores the critical value of bimodal interactions and uses a multi-head attention mechanism to enhance modal information.

GraphCAGE: GraphCAGE [19] models un-aligned multimodal sequences based on graph neural models and Capsule Network to explicitly learn long-range dependency.

MISA: MISA [30] notes the heterogeneity and invariance in the modalities and projects each modality into two different subspaces to achieve effective integration of modal information.

FmlMSN: FmlMSN [17] annotates sentiment labels for unimodal modalities and bimodal modalities and tri-modal modalities, and trains the modalities at different granularities as separate tasks in multiple stages.

TETFN: TETFN [13] utilises text as the dominant modality to learn cross-modal interactions and uses single-peak prediction to preserve differentiated information among modalities.

4.1.3 Experimental parameter settings

We use pre-trained Bert to convert textual modality into 768-dimensional embedding representations. For the CMU-MOSI dataset, we set the batch size to 32, the learning rate to $3e-5$, the number of heads of multi-head attention to 4, and the dropout of the prediction layer to 0.3. In addition, we set the number of hidden units in the single modal encoding layer to 128, 64 and 64 for text, acoustic and visual modalities, respectively. For the CMU-MOSEI dataset, we set the batch size to 128, the learning rate to $5e-5$, the number of heads for multi-head attention to 8, and the dropout of the prediction layer to 0.3. In addition, we set the number of hidden units in the unimodal encoding layer for text, acoustic and visual modalities are 256, 128 and 128, respectively.

Table 1 Datasets statistics in MOSI and MOSEI

Dataset	Train	Valid	Test	Total
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856

We use the Adam optimiser for training and apply the early stop method with patience set to 8 and use the strategy of grid search to find suitable parameters. Moreover, in order to maintain the consistency of evaluation metrics on different models, we follow previous works [19], and report: 1) 2-class accuracy (Acc-2), 7-class accuracy (Acc-7) and F1-score (F1) for classification, 2) Mean Absolute Error (MAE) and Pearson correlation (Corr) for regression.

In addition, our model requires an average training time of 11.31s per epoch on CMU-MOSI and 42.25s per epoch on CMU-MOSEI. The model parameter sizes are 549M for CMU-MOSI and 667M for CMU-MOSEI.

4.2 Results analysis

Table 2 shows the results of the experiments on CMU-MOSI, where the result in bold in each column indicates that it is the best indicator in the corresponding column. As can be seen from the table, our method yields a significant improvement in each evaluation metric. For the regression task, compared to the best method in the table, our method improves by 0.011 in MAE. For the regression task, our method still shows excellent performance compared to the best-performing TETFN, with improvements of 0.44% and 0.35% on the F1 score and the binary classification task, respectively. For the seven classification task, our method also achieves optimal performance, with a 3.91% improvement in seven classification accuracy over the best method in the table, demonstrating our method's superiority.

In addition, as can be seen from the experimental results in Table 2, although TFN and LMF use the outer product to explore the effective information within and between modalities, they both neglect the modalities' heterogeneity, making their performance unsatisfactory. BIMHA takes note of the variance in the contribution of different modalities and improves the performance compared to previous work; however, it still does not achieve optimal results

as BIMHA ignores the problem of information forgetting when modelling long sequences. MTAG and GraphCAGE project serialised information into the graph representation space, which alleviates the dependency problem presented in lengthy sequence modelling and shows better performance. However, the above methods are still inferior to our method; this is because our method not only uses a graph structure to deal with the sequence modelling problem but also designs a novel convolutional aggregation module to overcome the noise problem caused by differences in modal information density, which allows our method to capture higher quality feature representations.

Table 3 shows the experimental results on the CMU-MOSEI dataset, where the result in bold in each column indicates that it is the best indicator in the corresponding column, and the result in † in each column indicates that it is the second-best indicator in the corresponding column. In addition, the results for F1, Acc-2, and Acc-7 are retained to two decimal places to maintain data consistency between the methods in the table. Based on the results in the table, it can be observed that our method demonstrates extremely high superiority and performance. Compared to the best model in the table, our method shows excellent performance on the classification task, achieving the best results for its F1 score and second place for Acc-2. For the regression task, our method shows highly competitive performance, with MAE improving by 0.008 compared to the best-performing model in the table. It is worth noting that the results of MISA on seven classification accuracy and corr are slightly higher than those of our method, and we speculate that this may be since MISA uses different training loss functions for different modal subspaces respectively, which leads to a more accurate training direction for the task and makes it easier to obtain superior classification results. The results of TETFN in the classification task are closer to ours, which may be because TETFN focuses on the differentiation information of the single peaks themselves and improves the model's sensitivity to the dif-

Table 2 Experimental results on CMU-MOSI

Model	MAE↓	Corr↑	F1(%)↑	Acc-7(%)↑	Acc-2(%)↑
TFN	0.947	0.673	77.95	34.46	77.99
LMF	0.950	0.651	77.80	33.82	77.90
MulT	0.889	0.686	81.00	39.10	81.10
MTAG	0.941	0.692	80.40	31.90	80.50
BIMHA	0.925	0.671	78.50	36.44	78.57
GraphCAGE	0.933	0.684	82.10	35.40	82.10
MISA	0.783	0.761	81.7	42.30	81.8
FmlMSN	0.977	0.669	79.78	31.44	80.09
TETFN	0.717	0.800	83.83	-	84.05
HFNGC(ours)	0.706	0.800	84.27	46.21	84.40

↑ this is indicates the bigger the indicator, the better the performance

↓ this is indicates the lower the indicator, the better the performance

Table 3 Experimental results on CMU-MOSEI

Model	MAE↓	Corr↑	F1(%)↑	Acc-7(%)↑	Acc-2(%)↑
TFN	0.573	0.714	78.96	51.60	78.50
LMF	0.591	0.694	81.60	51.59	81.60
MuT	0.559	0.733	81.56	50.70	81.15
MTAG	0.645	0.614	75.90	48.20	79.10
BIMHA	0.559	0.731	83.35	52.11	84.07
GraphCAGE	0.609	0.670	81.80	48.90	81.70
MISA	0.555	0.756	83.80	52.20†	83.60
FmlMSN	0.569	0.719	83.56	52.69	83.45
TETFN	0.551†	0.748	84.18†	-	84.25
HFNGC(ours)	0.543	0.755†	84.19	51.45	84.10†

↑ this is indicates the bigger the indicator, the better the performance

↓ this is indicates the lower the indicator, the better the performance

† this is the second-best result in this column

bolded is the best result in this column

ferentiation information. In addition, it is worth noting that BIMHA provides slightly higher results than our method does for seven classification accuracy. However, BIMHA still needs to be improved to that of our method in other evaluation metrics, which indicates that our method has a more powerful generalization capability and can perform better modality integration and information extraction.

In conclusion, the above experimental results demonstrate our proposed method's remarkably superior performance and powerful generalisation capability.

4.3 Ablation analysis

In order to verify the correctness and effectiveness of our proposed method, we conduct ablation studies based on CMU-MOSEI, which is experimentally analysed from two perspectives: the contribution of different module combinations and the contribution of different modality combinations, as shown in Tables 4 and 5.

Table 4 shows the experimental results of our method with different combinations of modules. We denote C for the convolutional aggregation module, G for the graph construction and convolution layer, and F for the multimodal fusion layer. The experiments show that the model performs much better

using all modules than other module combinations. It can be noted that the most significant degradation in model performance is observed when the convolutional aggregation module is deprecated. The above experiments demonstrate that the variability in modality information density affects model performance and prove that the convolutional aggregation module we designed can effectively improve the information quality of the modalities. When the graph construction is abolished, the model performance in terms of F1 score and accuracy drops by 1.87% and 2.07%, respectively, which is also a non-negligible drop. The above results are excellent evidence for the importance and usefulness of graph construction for sequential modelling. After abolishing the multimodal fusion layer, the experimental results are inferior to that of the original model in all evaluation metrics, which proves that modality fusion is a crucial component for the multimodal sentiment analysis and shows that the measure of modality fusion employed in our approach is effective.

Table 5 shows the experimental results for different modality combinations. It can be seen that the textual modality achieves the best results in the unimodal prediction task, which is consistent with our expectations. The results of the unimodal prediction task illustrate that, owing to the presence of the large pre-trained model and the richness of informa-

Table 4 Experimental results for different combinations of modules on CMU-MOSEI

Model	MAE↓	Corr↑	F1(%)↑	Acc-7(%)↑	Acc-2(%)↑
HFNGC(ours)	0.543	0.755	84.19	51.45	84.10
HFNGC w/o C	0.559	0.746	78.32	50.96	77.39
HFNGC w/o G	0.554	0.741	82.32	51.26	82.03
HFNGC w/o F	0.581	0.708	81.92	51.08	81.75

↑ this is indicates the bigger the indicator, the better the performance

↓ this is indicates the lower the indicator, the better the performance

"w/o C" indicates that the convolutional aggregation module is not used

"w/o G" indicates that the graph construction and convolution layer is abrogated

"w/o F" indicates that the multimodal fusion layer is not used

bolded is the best result in this column

Table 5 Experimental results for different modality combinations on CMU-MOSEI

Model	MAE↓	Corr↑	F1(%)↑	Acc-7(%)↑	Acc-2(%)↑
A	0.814	0.254	65.66	41.23	69.18
V	0.812	0.214	60.02	42.00	70.68
T	0.574	0.727	78.05	49.71	77.14
A+V	0.802	0.214	60.02	42.01	70.68
A+T	0.564	0.736	81.89	51.21	81.47
V+T	0.557	0.739	79.24	51.27	78.54
All	0.543	0.755	84.19	51.45	84.10

↑ this is indicates the bigger the indicator, the better the performance

↓ this is indicates the lower the indicator, the better the performance
bolded denotes the best result in each group

tion contained in the textual modality, the textual modality can provide a superior representation after the initial feature extraction, which indicates that the textual modality contains more useful information and is very helpful for the training of the sentiment prediction model. However, the vast amount of superfluous information in audio and video, combined with the absence of high-quality representations during the initial feature extraction phase, results in inability to precisely follow useful information. Hence, they are unable to achieve satisfactory training outcomes. In the bimodal prediction task, A+T and V+T achieve excellent results, while A+V obtains an inferior performance. The above results indicate that the involvement of the textual modality can guide model learning effectively, which is consistent with our expectations. It is worth noting that the combination of audio and visual modalities (A+V) shows improvement compared to using either audio (A) or visual (V) modality alone, which shows that the mutual enrichment of non-textual modalities contributes to the improvement of semantic quality. However, when comparing the results of A+V with those of text modality (T), we can observe that there is still a gap between the text modality and the joint representation of A+V. The phenomenon above occurs because of the non-text modalities' poor information quality. Although the convolutional aggregation module enhances the semantic gap to some extent, it

cannot eliminate the gap. In addition, it is worth noting that the A+V combination does not perform as well in the bimodal prediction task, but it still achieves a good improvement compared to the single-peak prediction task. This suggests that inter-modal interactions can lead to better information extraction and information supplementation. Furthermore, it can be noted that the best results are obtained for the multimodal prediction task with all modalities are involved, which indicates that multiple modalities can bring rich and useful information, and that our method can effectively integrate and learn information from multiple modalities.

4.4 Noise reduction capability of the convolutional aggregation module

To prove the effectiveness of the convolutional aggregation module in reducing noise caused by information density differences among modalities, we conduct experiments on CMU-MOSEI.

As shown in Table 6, A+V denotes that the visual and acoustic modalities are directly fed into the subsequent module without convolutional aggregation; convA+convV denotes that the visual and acoustic modalities perform convolution and graph construction respectively; HFNGC denotes the original model.

As can be seen from Table 6, A+V presents the worst performance in both classification and regression tasks. However, the visual modality and acoustic modality, when convolved separately, outperform A+V in all evaluation metrics, and their detailed performance is shown as convA+convV in Table 6. The above results illustrate that performing feature enhancement for non-textual modalities is essential, which can reduce the ratio of irrelevant information in non-textual features and reduce noise interference in model training.

In addition, we note that the performance of convA+convV is inferior to that of HFNGC in most of the evaluation metrics, which proves that separate convolution cannot achieve the maximum benefit for modal gain, whereas the convolutional aggregation module in HFNGC can improve the feature quality and reduce the impact of irrelevant information by mutual gain through non-textual modalities.

Table 6 Effectiveness of HFNGC in solving noise on CMU-MOSEI

Model	MAE↓	Corr↑	F1(%)↑	Acc-7(%)↑	Acc-2(%)↑
A+V	0.544	0.725	78.21	50.39	77.22
convA+convV	0.537	0.751	82.24	51.03	81.75
HFNGC	0.543	0.755	84.19	51.45	84.10

↑ this is indicates the bigger the indicator, the better the performance

↓ this is indicates the lower the indicator, the better the performance

A+V denotes that the visual and acoustic modalities are not convolutionally aggregated

convA+convV denotes that the visual and acoustic modalities are convolved separately

bolded is the best result in this column

The above experiments can prove that the presence of the convolution aggregation module can reduce the influence of noise and increase the proportion of useful information in modalities.

4.5 Justification for the number of dynamic routing iterations

In order to justify the number of Dynamic Routing iterations, we conduct experiments on CMU-MOSEI, as shown in Fig. 4. The horizontal coordinates in the figure indicate the number of Dynamic Routing iterations in each round of training, and the vertical coordinates indicate the accuracy and F1 score of our model. The solid line indicates the trend of the model's performance in terms of accuracy and the dashed line indicates the trend of the model's performance in terms of F1 scores.

From Fig. 4, it can be observed that as the iteration increases, the model's performance exhibits local fluctuations; however, the overall performance shows a downward trend. We believe that the above phenomenon is due to the fact that as the dynamic routing iterations increase, the model can continuously optimize its parameters and gradually approach the training data, so the training set error gradually decreases. However, if we continue to increase the iterations, it starts to capture the noise and outliers in the training set, which can lead to overfitting. The above phenomenon proves that the choice of the dynamic routing iteration is significant for the performance of the model.

In addition, it is worth noting that the model can obtain the best results when the iteration is set to 3, which justifies our parameter selection.

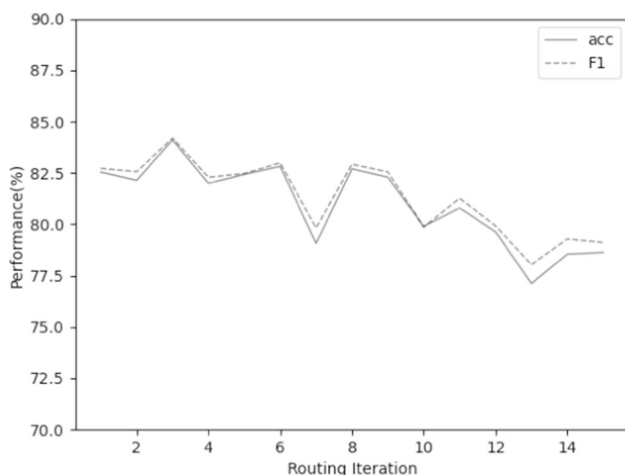


Fig. 4 Experimental results for different dynamic routing iterations on the CMU-MOSEI dataset. The solid line indicates the trend of the model's performance in terms of accuracy and the dashed line indicates the trend of the model's performance in terms of F1 scores

4.6 Data visualization

As shown in Fig. 5, we use different modality combinations for the sentiment prediction under LSTM.

With A, V and T indicating the performance of the single-peaked modality task, it can be seen that the information quality of the text modality is superior to that of the visual and acoustic modalities at the same level of modality extraction measures. We can notice from the above results that there is a difference in information density between different modalities and that enhancing the information quality of the visual and acoustic modalities is necessary.

In order to investigate the effect of different fusion methods on visual and acoustic modalities, we conduct experiments using concatenation (AV-Concat) and convolutional aggregation (AV-Fusion) respectively, as shown in Fig. 5. We can see that convolutional aggregation for modal aggregation performs better than others do. The results show that using convolutional aggregation can effectively improve the quality of the modality information and reduce the noise information resulting from the variability of the modality density.

5 Conclusion

In this paper, we propose a multimodal heterogeneous fusion network based on graph convolution. The model can effectively reduce the noise resulting from different information densities among modalities in multimodal sentiment analysis and ensure that the information representation of modalities is at the same semantic level. In addition, we employ graph

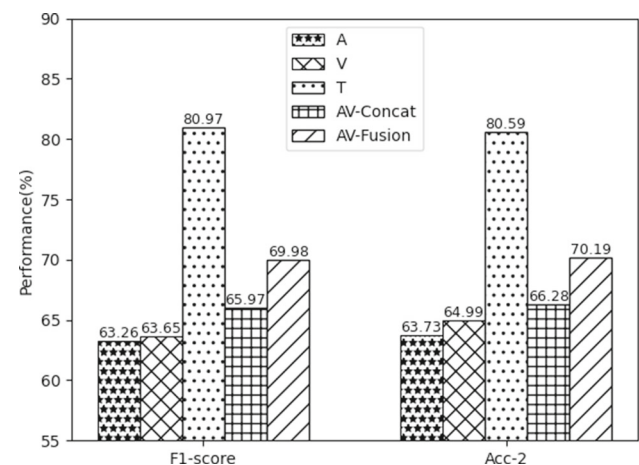


Fig. 5 Experimental results for different modality combinations under LSTM. T denotes the performance of text modality, A indicates the performance of acoustic modality, V denotes the performance of visual modality. AV-Concat indicated that the visual and acoustic modalities are concatenated as input to the LSTM. AV-Fusion indicated that the visual and acoustic modalities are convolved as input to the LSTM

structures to solve the gradient issues in RNN-based structures and the information loss problem when modelling long sequences.

Extensive experiments are carried out on the CMU-MOSI and CMU-MOSEI datasets. The experimental results show that the proposed method can extract and integrate modality information effectively, which outperforms the existing models with strong competitiveness.

In future work, we will further explore more approaches for modality enhancement, such as focusing on the impact of more dimensions on information representation. Furthermore, our future work will focus on enhancing modality heterogeneity for the classification task and explore better ways to improve the model's performance.

Acknowledgements The authors would like to thank the funding from the Open Project Program of Shanghai Key Laboratory of Data Science (No. 2020090600004) and the resources and technical support from the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600).

Author Contributions [Tong Zhao]: Conceptualization of this study, Methodology, Software, Writing-Original Draft. [Junjie Peng]: Conceptualization of this study, Writing-Review & Editing, Supervision. [Yansong Huang]: Formal analysis, Visualization. [Lan Wang]: Validation, Investigation. [Huiran Zhang]: Conceptualization of this study, Resources. [Zesu Cai]: Conceptualization of this study, Writing-Review & Editing.

Funding This study was supported by the Open Project Program of Shanghai Key Laboratory of Data Science (No. 2020090600004) and the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600).

Availability of data and materials In this work, we have used two publicly available datasets, CMU-MOSI dataset and CMU-MOSEI dataset, both of which can be available at <https://github.com/A2Zadeh/CMU-MultimodalSDK>.

Declarations

Ethics approval This article has never been submitted to more than one journal for simultaneous consideration. This article is original.

Consent to participate The authors have approved this article before submission, including the names and order of authors.

Consent for publication The authors agreed with the content and gave explicit consent to submit.

Competing interests The authors declared that they have no conflict of interest to this article.

References


- Zadeh A, Chen M, Poria S, Cambria E, Morency L (2017) Tensor fusion network for multimodal sentiment analysis. In: Palmer M, Hwa R, Riedel S (eds) Proceedings of the 2017 conference on empirical methods in natural language processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017, pp 1103–1114. <https://doi.org/10.18653/v1/d17-1115>
- Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency L (2018) Efficient low-rank multimodal fusion with modality-specific factors. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, vol 1: Long papers, pp 2247–2256. <https://doi.org/10.18653/v1/P18-1209>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency L (2019) Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019. AAAI Press, pp 7216–7223. <https://doi.org/10.1609/aaai.v33i01.33017216>
- Akhtar MS, Chauhan DS, Ghosal D, Poria S, Ekbal A, Bhattacharyya P (2019) Multi-task learning for multi-modal emotion recognition and sentiment analysis. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, vol 1 (Long and Short Papers). Association for computational linguistics, pp 370–379. <https://doi.org/10.18653/v1/n19-1034>
- Baltrusaitis T, Ahuja C, Morency L (2019) Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 41(2):423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>
- Gkoumas D, Li Q, Lioma C, Yu Y, Song D (2021) What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Inf Fusion* 66:184–197. <https://doi.org/10.1016/j.inffus.2020.09.005>
- Abdu SA, Yousef AH, Salem A (2021) Multimodal video sentiment analysis using deep learning approaches, a survey. *Inf Fusion* 76:204–226. <https://doi.org/10.1016/j.inffus.2021.06.003>
- Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P (2018) Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 32. <https://doi.org/10.1609/aaai.v32i1.12021>
- Mai S, Hu H, Xu J, Xing S (2020) Multi-fusion residual memory network for multimodal human sentiment comprehension. *IEEE Trans Affect Comput* 13(1):320–334. <https://doi.org/10.1109/TAFFC.2020.3000510>
- Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR (2021) Abcdm: an attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Futur Gener Comput Syst* 115:279–294. <https://doi.org/10.1016/j.future.2020.08.005>
- Wu T, Peng J, Zhang W, Zhang H, Tan S, Yi F, Ma C, Huang Y (2022) Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl Based Syst* 235:107676. <https://doi.org/10.1016/j.knosys.2021.107676>
- 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 Recognit* 136:109259. <https://doi.org/10.1016/j.patcog.2022.109259>

14. Xue X, Zhang C, Niu Z, Wu X (2023) Multi-level attention map network for multimodal sentiment analysis. *IEEE Trans Knowl Data Eng* 35(5):5105–5118. <https://doi.org/10.1109/TKDE.2022.3155290>
15. Zhu T, Li L, Yang J, Zhao S, Liu H, Qian J (2023) Multimodal sentiment analysis with image-text interaction network. *IEEE Trans Multimed* 25:3375–3385. <https://doi.org/10.1109/TMM.2022.3160060>
16. Zhang X, Chen Y, He L (2023) Information block multi-head subspace based long short-term memory networks for sentiment analysis. *Appl Intell* 53(10):12179–12197. <https://doi.org/10.1007/s10489-022-03998-z>
17. Peng J, Wu T, Zhang W, Cheng F, Tan S, Yi F, Huang Y (2023) A fine-grained modal label-based multi-stage network for multimodal sentiment analysis. *Expert Syst Appl* 221:119721. <https://doi.org/10.1016/j.eswa.2023.119721>
18. Chen Q, Huang G, Wang Y (2022) The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. *IEEE ACM Trans Audio Speech Lang Process* 30:2689–2695. <https://doi.org/10.1109/TASLP.2022.3192728>
19. Wu J, Mai S, Hu H (2021) Graph capsule aggregation for unaligned multimodal sequences. In: *Proceedings of the 2021 international conference on multimodal interaction*, pp 521–529. <https://doi.org/10.1145/3462244.3479931>
20. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: *Advances in neural information processing systems 30: annual conference on neural information processing systems 2017*, December 4–9, 2017, Long Beach, CA, USA, pp 3856–3866. <https://proceedings.neurips.cc/paper/2017/hash/2cad8fa47bbef282badbb8de5374b894-Abstract.html>
21. Yang J, Wang Y, Yi R, Zhu Y, Rehman A, Zadeh A, Poria S, Morency L-P (2021) Mtag: modal-temporal attention graph for unaligned human multimodal language sequences. In: *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp 1009–1021. <https://doi.org/10.18653/v1/2021.naacl-main.79>
22. Yang X, Feng S, Zhang Y, Wang D (2021) Multimodal sentiment detection based on multi-channel graph neural networks. In: *Zong C, Xia F, Li W, Navigli R (eds) Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, ACL/IJCNLP 2021, (vol 1: Long Papers), Virtual Event, August 1–6, 2021. Association for computational linguistics*, pp 328–339. <https://doi.org/10.18653/v1/2021.acl-long.28>
23. Zeng Z, Sun S, Li Q (2023) Multimodal negative sentiment recognition of online public opinion on public health emergencies based on graph convolutional networks and ensemble learning. *Inf Process Manag* 60(4):103378. <https://doi.org/10.1016/j.ipm.2023.103378>
24. Zhang Y, Tiwari P, Zheng Q, El-Saddik A, Hossain MS (2023) A multimodal coupled graph attention network for joint traffic event detection and sentiment classification. *IEEE Trans Intell Transp Syst* 24(8):8542–8554. <https://doi.org/10.1109/TITS.2022.3205477>
25. Lu Q, Zhu Z, Zhang G, Kang S, Liu P (2021) Aspect-gated graph convolutional networks for aspect-based sentiment analysis. *Appl Intell* 51(7):4408–4419. <https://doi.org/10.1007/s10489-020-02095-3>
26. Xu Q, Peng J, Zheng C, Tan S, Yi F, Cheng F (2023) Short text classification of chinese with label information assisting. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 1–18. <https://doi.org/10.1145/3582301>
27. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. *IEEE Intell Syst* 31(6):82–88. <https://doi.org/10.1109/MIS.2016.94>
28. Zadeh AB, Liang PP, Poria S, Cambria E, Morency L-P (2018) 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 (vol 1: Long Papers)*, pp 2236–2246. <https://doi.org/10.18653/v1/P18-1208>
29. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R (2019) Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the conference. Association for computational linguistics. Meeting*, vol 2019, p 6558. <https://doi.org/10.18653/v1/p19-1656>
30. Hazarika D, Zimmermann R, Poria S (2020) MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In: *Chen CW, Cucchiara R, Hua X, Qi G, Ricci E, Zhang Z, Zimmermann R (eds) MM '20: the 28th ACM international conference on multimedia, Virtual Event / Seattle, WA, USA, October 12–16, 2020. ACM*, pp 1122–1131. <https://doi.org/10.1145/3394171.3413678>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Tong Zhao¹ · Junjie Peng^{1,2}  · Yansong Huang¹ · Lan Wang¹ · Huiran Zhang¹ · Zesu Cai³

Tong Zhao
zhaotong@shu.edu.cn

Yansong Huang
huangyansong@shu.edu.cn

Lan Wang
wanglan1997@shu.edu.cn

Huiran Zhang
hrzhangsh@shu.edu.cn

Zesu Cai
caizesu@hit.edu.cn

¹ School of Computer Engineering and Science, Shanghai University, Shanghai, China

² Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China

³ School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China