Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Fusing pairwise modalities for emotion recognition in conversations

Chunxiao Fan [a,b], Jie Lin [a], Rui Mao [c,*], Erik Cambria [c]

[a] Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei, 230009, Anhui, China
[b] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, 230088, Anhui, China
[c] School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798, Singapore

ARTICLE INFO

Keywords:
Multimodal
Feature fusion
Graph convolution network
Emotion recognition in conversations

ABSTRACT

Multimodal fusion has the potential to significantly enhance model performance in the domain of Emotion Recognition in Conversations (ERC) by efficiently integrating information from diverse modalities. However, existing methods face challenges as they directly integrate information from different modalities, making it difficult to assess the individual impact of each modality during training and to capture nuanced fusion. To deal with it, we propose a novel framework named Fusing Pairwise Modalities for ERC. In this proposed method, the pairwise fusion technique is incorporated into multimodal fusion to enhance model performance, which enables each modality to contribute unique information, thereby facilitating a more comprehensive understanding of the emotional context. Additionally, a designed density loss is applied to characterise fused feature density, with a specific focus on mitigating redundancy in pairwise fusion methods. The density loss penalises feature density during training, contributing to a more efficient and effective fusion process. To validate the proposed framework, we conduct comprehensive experiments on two benchmark datasets, namely IEMOCAP and MELD. The results demonstrate the superior performance of our approach compared to state-of-the-art methods, indicating its effectiveness in addressing challenges related to multimodal fusion in the context of ERC.

1. Introduction

Emotion Recognition in Conversations (ERC) represents a specialised subfield within emotion recognition, specifically dedicated to the discernment and interpretation of emotions expressed during verbal exchanges. ERC emphasises on the intricate interplay of various modalities within the conversational context, which including spoken language, facial expressions, body language, and potentially exchanged textual information during dialogues. The utilisation of multimodal fusion techniques can serve to enhance the effectiveness for model performance, which is proved across diverse applications [1–6].

Consequently, multimodal fusion emerges as a critical component in the ERC domain, involving the amalgamation of information from diverse sources or modalities, such as facial expressions, spoken language, and textual information, which can utilise varied cues to attain a more nuanced understanding of emotional expression in conversational interactions. Generally, existing multimodal fusion methods can be categorised into two types: model-independent methods and model-based methods. Model-independent methods include early fusion, late fusion, intermediate fusion, and hybrid fusion.

Early fusion integrates raw data from different modalities for subsequent feature extraction and classification [7,8]. Late fusion trains each modal data separately, obtaining prediction results, and fuses the multiple models in the later stage using decision-making or ensemble methods [9]. Intermediate fusion transforms different modal data into high-dimensional features, and fuse them at the intermediate layer of the model [9,10]. Hybrid fusion performs early fusion on modalities with weak correlation and data synchronisation, and also takes late fusion on modalities with strong correlation and different data updates [11,12]. However, model-independent methods often struggle to effectively capture interactions among different modalities, leading to the development of model-based multimodal fusion methods that leverage relationships between modalities to capitalise on model advantages.

Model-based methods, in turn, consist of traditional methods and deep learning-based methods. Traditional methods include multiple kernel learning and graphical model-based methods. Multiple kernel learning methods take a combination of basic kernels to replace a single kernel, transforming the kernel selection problem into one of selecting combination coefficients.

* Corresponding author.
E-mail addresses: fanchunxiao@hfut.edu.cn (C. Fan), 2021171245@mail.hfut.edu.cn (J. Lin), rui.mao@ntu.edu.sg (R. Mao), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.inffus.2024.102306
Received 22 December 2023; Received in revised form 12 February 2024; Accepted 13 February 2024
Available online 15 February 2024

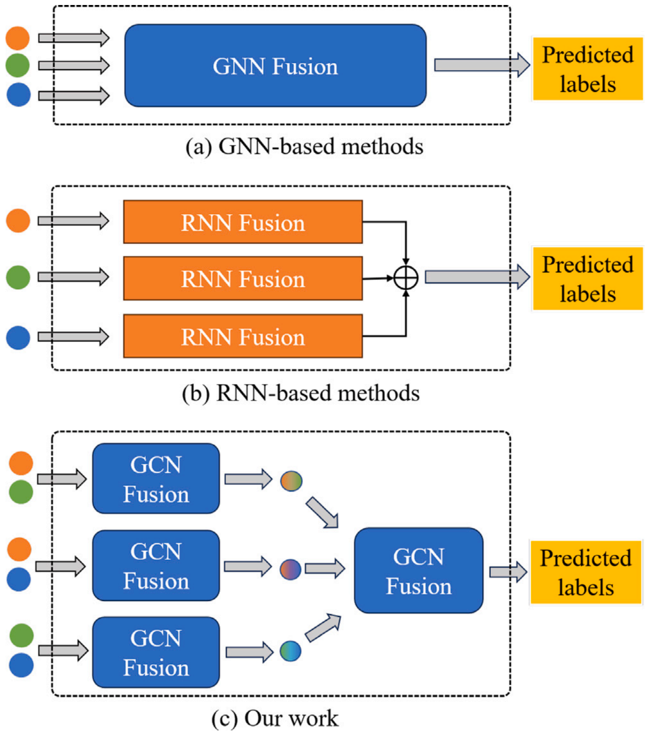


Fig. 1. The difference between the existing deep learning-based and our proposed approaches. (a) The GNNs represent different modalities directly as a graph, which proficiently capture distant contextual information. (b) The RNNs autonomously process each modality before subsequently combining them, which play a crucial role in tasks necessitating the integration of temporal multimodal information. (c) Our proposed work pairs modalities for comprehensive information fusion and then model the fused information as a graph to obtain the predicted labels.

Graphical model-based methods utilise graph structures to represent conditional dependencies between random variables, performing multimodal fusion [13–17]. Deep learning-based approaches facilitate end-to-end training for multimodal representation and fusion components. These methods demonstrate superior performance when contrasted with non-neural network-based systems [18,19]. Specifically, the employed architectures (see Fig. 1) predominantly involve Graph Neural Networks (GNNs) and Recurrent Neural Networks (RNNs). The GNNs demonstrate robust capabilities in modelling relationships, exemplified by methods such as MMGCN and GraphCFC, which proficiently capture distant contextual information [20,21]. Notably, these GNNs represent different modalities directly as a graph. Simultaneously, the RNNs play a crucial role in tasks necessitating the integration of temporal multimodal information. This involves the utilisation of architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs). In this context, RNNs autonomously process each modality before subsequently combination [22–24].

The existing approaches for multimodal fusion directly integrate information from different modalities, posing challenges in evaluating the individual impact of each modality during training and capturing the nuanced fusion of various modalities. In multimodal fusion, challenges arise from structural heterogeneity and disparities in the significance of various modalities, presenting impediments to cohesive integration. The principal challenges inherent in multimodal fusion pertain to the direct fusion for different modalities, resulting in information loss and confusion attributable to variations in characteristics and disparities in data distribution among modalities. Distinct modalities (including text, images, or audio) exhibit unique characteristics, thereby complicating the process of reconciling these disparities through direct fusion. Additionally, direct fusion impedes the capturing of nuanced relationships or alignments between semantic levels across diverse modalities, consequently leading to the loss of valuable information.

To address this issue, we propose a novel multimodal fusion method which facilitates comprehensive information integration through the paired fusion of modalities. Nevertheless, the existence of similar information across distinct modalities poses a significant obstacle, as the strategy of pairing modalities for fusion introduces the issue of modal redundancy. The pairwise fusion approach may result in features characterised by elevated correlation, thereby introducing duplicated information and potential redundancies. In instances where modalities exhibit similarity or correlation in information content, the heightened correlation among features imparts redundant information pertaining to the same conceptual or contextual content. The redundancy has the potential to adversely affect model performance, particularly in tasks demanding efficient information utilisation.

To achieve comprehensive multimodal fusion and address the challenges for pairwise information fusion, we undertake the following initiatives:

(1) The pairwise fusion method is integrated into multimodal fusion to enhance model performance. It discerningly manages relationships within each modality pair, thereby improving overall fusion performance. Information fusion from each modality pair heightens feature diversity, resulting in more enriched and varied feature representations, consequently enhancing the model's data representation capacity. The integration of data from each modality pair ensures comprehensive information utilisation, affording each modality the opportunity to contribute unique information, which facilitates the model's comprehensive understanding of the task.

(2) The designed density loss aims to characterise the density of fused features, focusing on addressing the challenge of feature redundancy inherent in pairwise fusion methods. The designed density loss is intended to penalise the density of fused features, prompting the model to systematically alleviate redundancy during training. Additionally, the incorporation of the density loss incentivises the model to prioritise the learning of distinctive features and facilitates the integration of information from diverse modalities, in contrast to a simplistic fusion of all available features.

The proposed method undergoes evaluation on the task of ERC, utilising two benchmark datasets: IEMOCAP [25] and MELD [26]. Extensive experiments demonstrate that our approach achieves an average F1 score approximately 4.01% higher (IEMOCAP) and 6.47% higher (MELD) than state-of-the-art methods.

The contributions of this work can be summarised as follows:

- An novel pairwise modalities fusion approach for ERC is proposed to optimise the efficiency of multimodal fusion, thereby enhancing overall model performance. The integration of data from each modality pair ensures comprehensive information utilisation, enabling each modality to contribute unique information and facilitating a more comprehensive understanding. Information fusion from each modality pair increases feature diversity, resulting in enriched and varied feature representations, thereby enhancing the model's data representation capacity.
- The density loss is designed and introduced in multimodal fusion to address concerns related to modality redundancy by constraining the size of model parameters during training. The formulated density loss characterises the density of fused features, with a specific focus on mitigating feature redundancy inherent in pairwise fusion methods. This designed density loss is intended to penalise the density of fused features, which can effectively minimise feature redundancy systematically reducing redundancy during model training. Consequently, the proposed framework can elevate the efficiency and robustness of multimodal fusion models.

- Comprehensive experiments are conducted on the IEMOCAP and MELD datasets, attest to the superior performance exhibited by the proposed framework relative to incumbent methods. These results affirm the efficacy of the framework in effectively addressing the inherent challenges associated with multimodal fusion.

The remainder of this paper is structured as follows: Section 2 discusses relevant prior work. In Section 3, we offer a comprehensive description of the proposed model. Section 4 details the dataset configurations and baseline models utilised in our experiments. Section 5 presents the experimental results and their analysis. Our work is concluded in Section 6.

2. Related work

The evolution of affective computing has progressed from unimodal processing [27–29] to multimodal processing [21,30–32]. This is because multimodal fusion can improve performance using the joint recognition of information from multiple modalities, fusing complementary information among different modalities, thus making recognition more stable. The existing multimodal fusion methods can be divided into two main categories: model-independent fusion methods and model-based fusion methods. Model-independent fusion methods can be further categorised into early fusion, late fusion, hybrid fusion. On the other hand, model-based methods can be segmented into traditional methods and deep learning-based methods.

2.1. Model-independent fusion methods

The model-independent fusion methods do not explicitly rely on the underlying model architecture for integrating information from different modalities. The model-independent fusion methods consist of early fusion, late fusion, and hybrid fusion.

2.1.1. Early fusion

Early fusion involves the extraction of features from different modalities, followed by concatenation of these features into a unified representation, which is subsequently utilised for emotion prediction. [1,23] utilise early fusion by concatenating features from diverse modalities for emotion prediction. Such a method may face challenges in capturing nuanced contextual relationships between modalities. [12] takes early fusion to combine features from audio, video, and text modalities, generating fused features. Subsequently, a hierarchical fusion Graph Convolutional Networks (GCN) is applied to perform feature fusion and facilitate emotion recognition. Thus, the early fusion methods combine information from different modalities, and emphasise its role in various studies for emotion prediction. It is important to note that while early fusion is a straightforward approach, capturing intricate inter-modal relationships can be challenging.

2.1.2. Late fusion

Late fusion involves integrating and combining the decisions made by each modality to produce the final result, which is a form of decision fusion. [33,34] employ late fusion to combine multimodal features, which is executed at the classification result level. The decisions made by each modality are integrated to derive the ultimate result. [23] employs three separate LSTM-based models, each dedicated to processing audio, video, and text features independently to capture contextual information. The late fusion is implemented to consolidate the decisions made by these models. However, the limitation of assuming modality independence is acknowledged as it may hinder the capture of intricate inter-modal interactions.

2.1.3. Hybrid fusion

Hybrid fusion combines elements of both early and late fusion methods. It utilises early fusion for modalities characterised by weak data synchronisation and correlation, while opting for late fusion in scenarios exhibiting strong correlation and disparate data update patterns. [35] adopts a hybrid fusion approach, initially fusing each of the two modalities through early fusion. Subsequently, the intermediate outputs are also fused to give the final results. [36–39] employ hybrid fusion methods to explore interactions between modalities in isolated or temporal discourse. However, there remains a challenge in effectively capturing interactions between different modalities, resulting in limitations in fully exploiting contextual information and handling complex interactions within dialogues.

Model-independent multimodal fusion methods are versatile approaches applicable across diverse data types and models. However, their generic nature also limits their capacity to fully exploit intricate relationships between different modalities or harness specific features inherent in the data and models. Despite the broad applicability, the kind of methods can hardly leverage the full potential of multimodal information, and the performance varies depending on the complexity and characteristics of the underlying data and models.

2.2. Model-based fusion methods

The model-based fusion methods leverage specific features inherent in the data and models, aiming to achieve enhanced performance. These approaches typically outperform model-independent methods by effectively utilising the intricate relationships between different modalities and capitalising on the strengths inherent in the chosen models. The focus on exploiting modality-specific characteristics and modelling intricacies contributes to improved overall performance compared to more generic model-independent fusion techniques.

2.2.1. Traditional methods

Traditional multimodal fusion methods are commonly categorised into multi-kernel learning and graphical models. Multi-kernel learning enhances the fusion of heterogeneous data by incorporating modality-specific kernels, facilitating a more nuanced representation of diverse modalities. Graphical models exploit temporal relationships within data and offer adaptive and interpretable interpretability. [40] uses a multiple kernel learning approach by combining feature vectors from textual, visual, and audio modalities to train a classifier.

2.2.2. Deep learning-based methods

Deep learning has offered a significant advantage in multimodal fusion, leveraging their robust information learning capabilities. Effective fusion of features in neural networks allows models to capture dependencies and complementarities between modalities, thereby improving overall performance. These methods can generally be categorised into GNN-based and RNN-based approaches:

GNN-based methods: GNNs have gained widespread application for processing non-Euclidean data in various fields, such as computer vision, recommendation systems, and natural language processing [41,42]. There are widely applied GNN techniques, such as GAT [43], FastGCN [44], and Graph-SAGE [45], to solve emotion recognition problems. Some recent works adopt graph structures to fuse unimodal and multimodal features, which can obtain the contextual relationships of features and the complementary relationships between modalities.

DialogueGCN [30] proposed to apply a GCN to capture long-distance contextual information in dialogue. DialogueGCN treats each utterance as a node and connects any nodes within the same window in the dialogue. DialogueCRN [46] designed an inference module to obtain emotional information and tried to understand the context of the conversation from a cognitive perspective. DAG-ERC [47] treats each session as a directed acyclic graph (DAG). The DAG can collect information about query utterances from adjacent and remote utterances, which is similar to a combination of graph structure and recursive structure, achieving better performance.

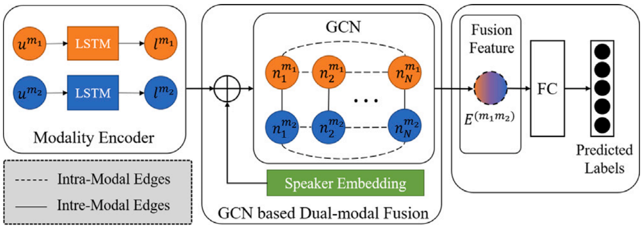


Fig. 2. The details of Pairwise Modalities Fusion for modalities of m_1 and m_2. The input features of m_1 and m_2, after being processed by a bidirectional LSTM, are subsequently merged with speaker information. The merged features are modelled as a graph and fused using GCN to generate integrated features. The optimal fusion feature are extracted with the optimal dual-modal fusion GCN, which is trained combining predicted labels using multiple Fully Connected (FC) layers.

RNN-based methods: BC-LSTM [23] uses a Bidirectional LSTM structure to encode contextual semantic information and does not use speaker information. ICON [1] and CMN [48] both utilise GRUs and memory networks. MFN [49] aligns the features of different modalities and uses multi-views information fusion. HiGRU [24] proposes a hierarchical GRU that uses two levels of GRUs, a lower-level GRU to model the word-level inputs and an upper-level GRU to capture the contexts of utterance-level embeddings. DialogueRNN [50] leverages distinct GRUs to capture speakers' contextual information. COSMIC [51] achieves state-of-the-art performance by combining a structure similar to DialogueRNN with external common-sense knowledge. DialogueCRN [46] uses a bidirectional LSTM and introduces an inference module that can simultaneously understand situation-level and speaker-level contexts to construct an ERC model.

3. Methodology

Audio, vision, and text represent the most commonly utilised modalities; therefore, we use them as illustrative examples to elucidate the proposed methodology. The features of these three modalities are extracted in pairs to obtain the modality features. The fused features extracted with designed pairwise modalities fusion are denoted as (VA) (fusion for video and audio), (TA) (fusion for text and audio), and (TV) (fusion for text and video).

In this section, the designed Pairwise Modalities Fusion (see Fig. 2) is introduced in Section 3.1. In Section 3.2, the designed Multiple Features Fusion is presented. The whole training process for the multimodal attention model is shown in Section 3.3. The detailed framework of our proposed model is shown in Fig. 3.

3.1. *Pairwise modalities fusion*

The data from different modalities are fused pairwise as shown in Fig. 2. In the designed pairwise fusion for different modalities, every two modalities in this method are encoded with the designed multimodal encoder using bidirectional LSTM, and the encoded features are modelled as graphs to provide the structure information during modalities fusion. The two modalities denoted as m_1 and m_2.

3.1.1. *Modality encoder*

To handle the time-series raw data, the designed method adopts bidirectional LSTM for modelling the sequences and capturing long-term dependencies in different modalities. Every two modalities are encoded using a dual-channel bidirectional LSTM model, and each channel processes the input sequence of one modality. Using bidirectional LSTM, the model can capture temporal relationships between the two modalities, aiding in better utilisation of information in sequential data.

For the modalities m_1 and m_2, the context-aware feature encoding $l_i^{m_1}$ and $l_i^{m_2}$ are as:

$$\begin{aligned} l_i^{m_1} &= \left[\overrightarrow{\mathrm{LSTM}}\left(u_i^{m_1}, l_{i-1}^{m_1}\right), \overleftarrow{\mathrm{LSTM}}\left(u_i^{m_1}, l_{i+1}^{m_1}\right)\right], \\ l_i^{m_2} &= \left[\overrightarrow{\mathrm{LSTM}}\left(u_i^{m_2}, l_{i-1}^{m_2}\right), \overleftarrow{\mathrm{LSTM}}\left(u_i^{m_2}, l_{i+1}^{m_2}\right)\right], \end{aligned} \tag{1}$$

where i represents the ith data sample. $u_i^{m_1}$ and $u_i^{m_2}$ are the context-independent raw feature representation in the modalities of m_1 and m_2, respectively. $\overrightarrow{LSTM}$ denotes the contextual relationships of input features obtained from forward sequences, and $\overleftarrow{LSTM}$ denotes from reverse sequences. $[\cdot]$ represents the concatenation operation.

3.1.2. *GCN based dual-modal fusion*

The high-level features across the two modalities are extracted using undirected GCN structure, which can learn to aggregate information from neighbouring nodes, allowing it to capture complex dependencies and patterns in the relationships between encoded features. The encoded features in the modalities are represented as nodes in the graph. The connexions (edges) between nodes are established, based on interactions between corresponding features from different modalities, making it suitable for capturing relationships between nodes.

To enhance feature extraction leveraged by the integrated GCN layers, speaker embeddings are embedded into the nodes of the designed GCN structure. The speaker embeddings are the representations of speaker characteristics obtained through speaker recognition models. It can involve concatenating the speaker embeddings with the features from each node in the graph, effectively providing the GCN with additional information related to speaker characteristics. With encoding the speaker identity information, the nodes in the modelled GCN are generated by weighted adding the speaker embedding s_i with the output results of the bidirectional LSTM $l_i^{m_1}, l_i^{m_2}$ as:

$$\begin{aligned} n_i^{m_1} &= \omega_{ps} s_i + l_i^{m_1}, \\ n_i^{m_2} &= \omega_{ps} s_i + l_i^{m_2}, \end{aligned} \tag{2}$$

where $n_i^{m_1}$ and $n_i^{m_2}$ denote the ith nodes in the modelled GCN. ω_{ps} represents the learnable weight parameters. s_i is the speaker information. Thus, we have the set of nodes $V_p = \{n_1^{m_1}, n_1^{m_2}, n_2^{m_1}, n_2^{m_2}, \dots, n_N^{m_1}, n_N^{m_2}\}$, where N is the number of samples in each modality.

In order to extract temporal and inter-modal information more comprehensively, each sample is linked to various samples within its own modality and is also connected to the sample at the current moment in another modality. These connexions are referred to as intra-modal and inter-modal edges, respectively. The weight of intra-modal edge can be obtained as:

$$W_{ij}^a = 1 - \frac{\arccos\left(\mathrm{sim}\left(n_i^{m_1}, n_j^{m_2}\right)\right)}{\pi}, \tag{3}$$

where $n_i^{m_1}$ and $n_j^{m_2}$ represent the ith node of m_1 and jth node of m_2, respectively. sim is the cosine similarity. The weight of inter-modal edge can be defined as:

$$W_i^e = \gamma(1 - \frac{\arccos\left(\mathrm{sim}\left(n_i^{m_1}, n_i^{m_2}\right)\right)}{\pi}), \tag{4}$$

where $n_i^{m_1}$ and $n_i^{m_2}$ represent the ith nodes of m_1 and m_2, respectively. γ is a hyperparameter.

The intra-modal edges are used to capture the contextual information within the modality, while inter-modal edges are used to capture the interactive information across modalities. Thus, the set of weights $\mathcal{E}_p = \{W_{11}^a, W_{12}^a, \dots, W_{1n}^a, W_{22}^a, W_{23}^a, \dots, W_{nn}^a, W_1^e, W_2^e, \dots, W_n^e\}$ can be obtained via Eqs. (3) and (4).

By depicting the outcomes of modality encoding as an undirected graph $G_p = (V_p, \mathcal{E}_p)$, the model can adeptly grasp and exploit the interrelationships among features originating from diverse modalities. This process facilitates improved feature extraction for subsequent tasks. The spectral domain GCN is constructed to encode multimodal contextual information, which uses learnable speaker embeddings to encode speaker-level contextual information (see Fig. 3).

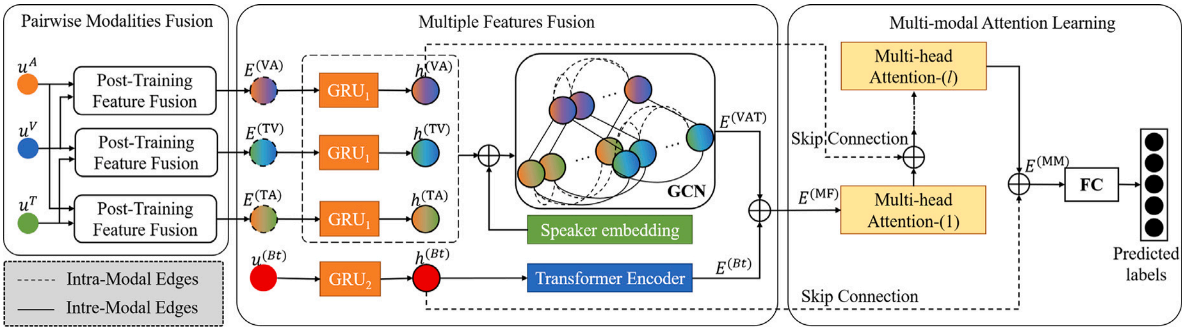


Fig. 3. The details of our proposed fusing pairwise modalities for ERC. The proposed model consists of three parts, including Pairwise Modalities Fusion (PMF), Multiple Features Fusion (MFF), and Multimodal Attention Learning (MAL).

3.1.3. Training

After constructing the undirected graph, in order to better encode the context, for the designed $G_p = (V_p, \mathcal{E}_p)$, the following method is used to obtain the renormalised graph Laplacian matrix [52], and the fusion feature for the two modalities are calculated as:

$$E^{(m_1 m_2)} = \mathrm{GCN}_p(V_p, \mathcal{E}_p), \tag{5}$$

where $E^{(m_1 m_2)}$ is the PMF result for the modalities of m_1 and m_2.

To train the GCN_p model for feature fusion, the fully connected (FC) layer is added after the features of GCN_p for the recognition training. In addition, we design the density loss $\mathcal{L}_{pd}$ to represent the density of fusion features for reducing the redundancy after the pairwise modalities fusion.

$$\mathcal{L}_{pd} = \|E^{(m_1 m_2)}\|_1, \tag{6}$$

where $\|\cdot\|_1$ denotes the L1 norm. Thus, the trained loss function can be formularised with the categorical cross-entropy $\mathcal{L}_{pc}$, $\mathcal{L}_{pd}$, and L2-regularisation as

$$\mathcal{L}_{\mathrm{PMF}} = \mathcal{L}_{pc} + \lambda_{p1}\mathcal{L}_{pd} + \lambda_{p2}\|\theta_p\|_2, \tag{7}$$

where λ_{p1} is the weight for density. λ_{p2} is the weight for L2-regularisation. θ_p denotes all the trainable parameters in the trained model. The introduction of regularisation can prevent model overfitting in training [53].

By optimising the designed $\mathcal{L}_{\mathrm{PMF}}$, the input features of modalities m_1 and m_2 can be fused, and then obtaining the fused feature $E^{(m_1 m_2)}$ before the fully connected layer. The redundant features between the modalities of m_1 and m_2 can also be reduced by optimising the designed $\mathcal{L}_{pd}$. We take the most widely used modalities as the examples, so three fused features $E^{(\mathrm{VA})}$, $E^{(\mathrm{TA})}$, and $E^{(\mathrm{TV})}$ can be obtained with the designed method.

3.2. Multiple features fusion

The results of pairwise fused modalities are processed with GRU [54] to capture the time-dependent relationship in the time-series features. GRU performs well in capturing short-term dependencies for sequential data. It is particularly effective in tasks where remembering recent information is crucial for making predictions. With fewer parameters and a simpler structure, GRU is generally less prone to overfitting, especially when working with limited amounts of training data. It has demonstrated good performance in tasks involving sequential data due to its ability to capture dependencies over time [55].

$$\begin{aligned}
h_t^{(\mathrm{VA})} &= \mathrm{GRU}_1(h_{t-1}^{(\mathrm{VA})}, E_t^{(\mathrm{VA})}),\\
h_t^{(\mathrm{VT})} &= \mathrm{GRU}_1(h_{t-1}^{(\mathrm{VT})}, E_t^{(\mathrm{VT})}),\\
h_t^{(\mathrm{VA})} &= \mathrm{GRU}_1(h_{t-1}^{(\mathrm{VA})}, E_t^{(\mathrm{VA})}),
\end{aligned} \tag{8}$$

where t represents the tth data sample. $E_t^{(\mathrm{VA})}$, $E_t^{(\mathrm{VT})}$, and $E_t^{(\mathrm{AT})}$ are derived from the fused VA, VT, and AT fusion features, respectively. $h_t^{(\mathrm{VA})}$, $h_t^{(\mathrm{VT})}$, and $h_t^{(\mathrm{VA})}$ are the GRU results for the VA, VT, and AT fusion features, respectively.

The results of these GRU modules are also modelled with GCN structure to extract the high-level features with structure information. Similarly, the original speaker information is also embedded with speaker embedding to encode the speaker identity information.

$$\begin{aligned}
n_t^{(\mathrm{VA})} &= w_{ms}s_t + h_t^{(\mathrm{VA})},\\
n_t^{(\mathrm{VT})} &= w_{ms}s_t + h_t^{(\mathrm{VT})},\\
n_t^{(\mathrm{AT})} &= w_{ms}s_t + h_t^{(\mathrm{AT})},
\end{aligned} \tag{9}$$

where w_{ms} represents the learnable weight parameters, and s_t represents the speaker information. Thus, we can have the set of nodes $V_f = \{n_1^{(\mathrm{VA})}, n_1^{(\mathrm{VT})}, n_1^{(\mathrm{AT})}, \dots, n_N^{(\mathrm{VA})}, n_N^{(\mathrm{VT})}, n_N^{(\mathrm{AT})}\}$.

The edges in the GCN structure also connect the nodes in the same modality to form intra-modal edges and the nodes at the same moment between different modalities for inter-modal edges as (3) and (4), respectively. Thus, we can obtain the set of weights $\mathcal{E}_f$, and construct the GCN structure of $G_f = (V_f, \mathcal{E}_f)$.

Therefore, the pairwise modalities fusion features can be processed with GCN_f.

$$E^{(\mathrm{VAT})} = \mathrm{GCN}_f(V_f, \mathcal{E}_f), \tag{10}$$

where $E^{(\mathrm{VAT})}$ is the modalities fusion result for all the modalities.

To enhance the contextualised semantic [56] and syntactic [57] understanding, the text feature is extracted via a RoBERTa pre-trained language model [58], which benefits from extensive pre-training that helps the model capture a diverse range of linguistic patterns and representations. RoBERTa uses a dynamic masking strategy during pre-training, where different masks are applied to different training instances, so it can enhance the model's ability to generalise across various tasks [59,60].

The features processed with RoBERTa are denoted as $u^{(Bt)}$ and also capture the time-dependent relationship with the GRU structure as:

$$h_t^{(Bt)} = \mathrm{GRU}_2(h_{t-1}^{(Bt)}, u_t^{(Bt)}), \tag{11}$$

The results of GRU are processed by the Transformer encoder [61], allowing for parallelisation of training due to its self-attention mechanism, which enables the model to attend to all positions in the input sequence simultaneously, leading to faster training times. The Transformer encoder is scalable to different input lengths, and the self-attention mechanism in Transformers allows for efficient processing of both short and long sequences without significantly increasing computation.

A Transformer encoder with a variable stack is used to learn discourse context information for $h_t^{(Bt)}$, resulting in $E^{(Bt)}$. The results of $E^{(Bt)}$ are concatenated with the $E^{(\mathrm{VAT})}$, so that the features fusion can be obtained combined with the text information extracted by the RoBERTa model.

$$E^{(\mathrm{MF})} = \left[E^{(\mathrm{VAT})}, E^{(Bt)}\right], \tag{12}$$

where $E^{(\mathrm{MF})}$ is the multiple features fusion for these modalities.

3.3. Multimodal attention learning

The designed multimodal attention learning jointly models and learns attention mechanisms across multiple modalities. The attention mechanism [61] is employed to capture and emphasise important information in a sequence or set of features [62,63]. Dealing with multimodal data, learning attention across these modalities becomes crucial for effective information fusion and understanding. Attention mechanisms facilitate the fusion of information from different modalities. By giving varying degrees of importance to different modalities or regions within modalities, the model can create a more informative and contextually rich representation. The attention weights assigned to different modalities or features can be dynamic and adapt to the input data. This adaptability is particularly useful when dealing with diverse or changing relationships between modalities.

In order to improve feature transfer efficiency and model training stability [64], skip-connection is adopted in the designed model and combined with the multi-head attention structure. The fused feature $h_t^{(\mathrm{VA})}$ is skip-connected to the multi-head attention directly, and $h_t^{(Bt)}$ is skip-connected to the result of multi-head attention. After applying the attention mechanism to the features fusion E_{FF}, we perform sentiment classification on the dataset. The attention mechanism is used to capture features that contain emotional tendencies or emotional factors.

$$\begin{aligned} C_1 &= \mathrm{ATT}_1(E^{(\mathrm{MF})}), \\ C_i &= \mathrm{ATT}_i(\left[C_{i-1}, h_t^{(\mathrm{VA})}\right]), \end{aligned} \tag{13}$$

where ATT_i is the multi-head self-attention mechanism for the ith layer. The h_t^{Bt} is skip-connected to the result of the last layer C_l in the multi-head self-attention mechanism to form the fused feature of multimodal $E_{\mathrm{MM}} = \left[C_l, h_t^{(Bt)}\right]$. $E^{(\mathrm{MM})}$ is set into several fully connected layers to get the predicted labels. The model is also trained with the density loss $\mathcal{L}_{md}$ to remove redundancy among different modalities.

$$\mathcal{L}_{md} = \|E^{(\mathrm{MM})}\|_1. \tag{14}$$

The training loss function is formed with the categorical cross-entropy $\mathcal{L}_{mc}$, density loss $\mathcal{L}_{md}$, and L2-regularisation as

$$\mathcal{L}_{\mathrm{MAL}} = \mathcal{L}_{mc} + \lambda_{m_1}\mathcal{L}_{md} + \lambda_{m_2}\|\theta_m\|_2, \tag{15}$$

where λ_{m_1} is the weight for density. λ_{m_2} is the weight for L2-regularisation. θ_m denotes all the trainable parameters in the trained model. The designed multimodal attention learning involves integrating information from different modalities. This enables the model to understand the relationships and dependencies between different modalities, leading to more comprehensive representations.

4. Experiments

4.1. Datasets

The proposed method is evaluated on two conversational multimodal emotion recognition benchmark datasets with aligned acoustic, visual, and textual information for each utterance in the dialogue. The IEMOCAP dataset is split into Training, Validation, and Test sets with a distribution of 100:20:31. In the MELD dataset, the partition ratio is set as 1038:114:280 for Training, Validation, and Test sets, respectively. Our dataset setup is consistent with that in [65].

IEMOCAP [25]: The IEMOCAP dataset is a multimodal dataset for emotion recognition and speech emotion recognition. The dataset includes approximately 12 h of audio and video recordings from 10 speakers, including speech and facial expression data. It includes a total of 7433 utterances and 151 dialogues. The IEMOCAP dataset contains 6 emotion categories: happiness, sadness, neutral, anger, excitement, and frustration. Each emotion category includes approximately 200 speech recordings.

MELD [26]: MELD (Multimodal Emotion Lines Dataset) is a dataset for emotion recognition and multimodal sentiment analysis that includes dialogue segments from a television show with video, audio, and text data. The MELD dataset contains 7 emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise, with a total of approximately 1400 dialogue segments and over 13 000 utterances. MELD has three or more speakers in one conversation.

4.2. Setups

In the IEMOCAP dataset, specific hyperparameters are employed during the Pairwise Modalities Fusion (PMF) for training fusion features involving audio and video, video and text, and text and audio. The hyperparameter configuration includes a dropout rate of 0.15, a learning rate of 0.0002, λ_{p1} set to 0.00002, γ at 0.7, and λ_{p2} at 0.00002. In the Multiple Features Fusion, the dropout rate is adjusted to 0.25, the learning rate becomes 0.00015, λ_{m_1} is set to 0.00002, γ is maintained at 0.7, and λ_{m_2} is held at 0.00002. The transformer encoder configuration includes one header and one layer. For the MELD dataset, a similar pattern is followed in the PMF with a dropout rate set to 0.2, a learning rate of 0.0002, λ_{p1} at 0.0002, γ set to 0.7, and λ_{p2} of 0.00002 during the training of fusion features. In the Multimodal Attention Learning, the dropout rate is set to 0.1, the learning rate is increased to 0.0004, λ_{m_1} is set to 0.0002, γ is maintained at 0.7, and λ_{m_2} is adjusted to 0.000015. The Transformer encoder in this case has four headers and one layer. These hyperparameter settings are specifically tailored for effective training and fusion of multimodal features in each dataset.

The features in different modalities are extracted before input the model. *Vision*: The visual facial expression features are extracted using the DenseNet [66] pre-trained on the Facial Expression Recognition Plus (FER+) corpus [67]. *Audio*: The acoustic raw features are extracted using the OpenSmile toolkit with the IS10 configuration [68]. *Text*: Two extraction methods are used to obtain text features. (1) The text feature is extracted using TextCNN following ICON [1]. TextCNN uses a Convolutional Neural Network (CNN) for text classification, which is a simple model that can serve as a baseline for text classification. (2) The RoBERTa-Large [58] model is used to extract context-independent utterance-level feature vectors. A special token $[CLS]$ is added at the beginning of the utterance. These vectors are averaged and passed through a linear layer to obtain a context-independent utterance feature vector.

4.3. Baselines

Our method is compared with several baselines.

CMN [48]: CMN uses two GRUs to encode the conversational context of the speaker, which achieves the state of the art. CMN considers the utterance histories of both speakers to model emotional dynamics.

BC-LSTM [23]: BC-LSTM uses a Bidirectional LSTM structure to encode contextual semantic information, which is the earliest method used on the MELD dataset and IEMOCAP dataset. BC-LSTM ignores speaker information because it does not attach any information related to the speaker to their model.

ICON [1]: ICON uses a global GRU to model the variance of in emotion within a conversation. ICON introduce a multimodal approach that provides comprehensive features from modalities such as language, visual, and audio in utterance videos.

DialogueRNN [50]: DialogueRNN uses three GRUs to capture distinct aspects of speaker-related information. Specifically, these three GRUs are dedicated to extracting contextual information, modelling speaker identity, and discerning the emotional content of the speech.

DialogueGCN [30]: DialogueGCN is the first work applied GCN to emotion classification, using GCN to construct the contextual relationship of utterances. DialogueGCN only uses information from the text modality,

Table 1
Experimental results (Accuracy and Average-weight F1 score) on the IEMOCAP dataset. Average-F1 means Average-weight F1 score.

	IEMOCAP		MELD	
	Acc	F1	Acc	F1
CMN [48]	56.32	56.19	–	–
BC-LSTM [23]	55.21	54.95	59.62	56.80
ICON [1]	64.00	63.50	–	–
DialogueRNN [50]	63.50	63.50	59.54	57.03
DialogueGCN [30]	63.22	64.18	58.62	58.10
COSMIC [51]	–	65.30	–	65.21
MMGCN [20]	66.36	66.22	60.42	58.65
HFGCN [69]	–	67.24	–	59.71
DAG-ERC [47]	–	68.03	–	63.65
DialogueCRN [46]	66.05	66.33	60.73	58.39
MM-DFN [32]	68.21	68.18	62.49	59.4
COGMEN [31]	68.2	67.6	–	–
GraphCFC [21]	69.13	68.91	61.42	58.86
Ours	**69.57***	**69.34***	**67.13***	**66.05***

* Denotes the improvements are statistically significant on a two-tailed t-test ($p < 0.001$).

which prevents it from taking advantage of complementary information between different modalities.

COSMIC [51]: COSMIC is a state-of-the-art method that improves emotion classification by obtaining commonsense knowledge using ATOMIC. Similar to DialogueGCN, COSMIC only uses information from the text modality and ignores complementary information between different modalities.

MMGCN [20]: MMGCN effectively uses GCN to fuse audio, video, and text features. It also uses speaker information. MMGCN first uses a graph structure to obtain the interaction relationships of the three modalities.

HFGCN [69]: HFGCN uses a new graph fusion method, including two-stage graph construction, attention-based edge weights, and relationship graph transformation to capture multimodal interaction. It also proposes a multi-task loss to guide the joint prediction of emotion labels and Valence-Arousal (VA) levels.

DAG-ERC [47]: DAG-ERC uses a directed acyclic graph to collect information from adjacent and remote utterances in the query utterance. DAG-ERC provides an effective way to model the information flow between long-distance conversation context and nearby context.

DialogueCRN [46]: DialogueCRN designs an inference module that can understand both situation-level and speaker-level contexts, which first proposes to understand the conversational context from a cognitive perspective. It can extract and integrate emotional cues from contextual information.

MM-DFN [32]: MM-DFN designs a dynamic fusion module to fuse multimodal context features in the conversation.

COGMEN [31]: COGMEN proposes a contextualised GNN to solve the impact of context on utterances and the internal and external dependencies for predicting the emotions of each speaker's utterance during the conversation.

GraphCFC [21]: GraphCFC introduces a module for cross-modal feature complementarity, utilising a directed graph to proficiently capture contextual and interaction information. This approach attains a state-of-the-art performance.

5. Results

The benchmarking results are shown in Table 1. It can be observed that our proposed method achieves the highest accuracy and F1 scores among the existing methods.

Table 2
Ablation study of the different components impact of the proposed method on performance (Average-weight F1 score and Accuracy).

	PMF	DL	MAL	IEMOCAP		MELD	
				F1	Acc	F1	Acc
$\mathcal{N}_{11}$	×	×	×	64.70	64.39	64.07	64.42
$\mathcal{N}_{12}$	✓	×	×	66.17	66.85	64.54	64.72
$\mathcal{N}_{13}$	×	×	✓	64.97	65.37	64.25	65.07
$\mathcal{N}_{14}$	✓	×	✓	68.04	68.02	65.13	66.05
$\mathcal{N}_{15}$	✓	✓	×	67.22	67.31	64.82	65.43
$\mathcal{N}_{16}$	✓	✓	✓	**69.34**	**69.57**	**66.05**	**67.13**

On the IEMOCAP dataset, our method demonstrates a notable average F1 score of 69.34% and achieves the accuracy of 69.57%. Particularly, COSMIC is a representative method based on RNNs, incorporating commonsense elements such as mental states, events, and causal relations with enhanced text features. However, our proposed method outperforms COSMIC by 4.04% F1 scores. MMGCN is a representative multimodal fusion method, utilising graph convolution on undirected graphs with speaker information. GraphCFC is a leading cross-modal feature complementary method, based on directed graphs. Compared with these methods, our proposed method exhibits superior results, surpassing MMGCN and GraphCFC by 3.12% and 0.43% in weighted average F1 scores, respectively.

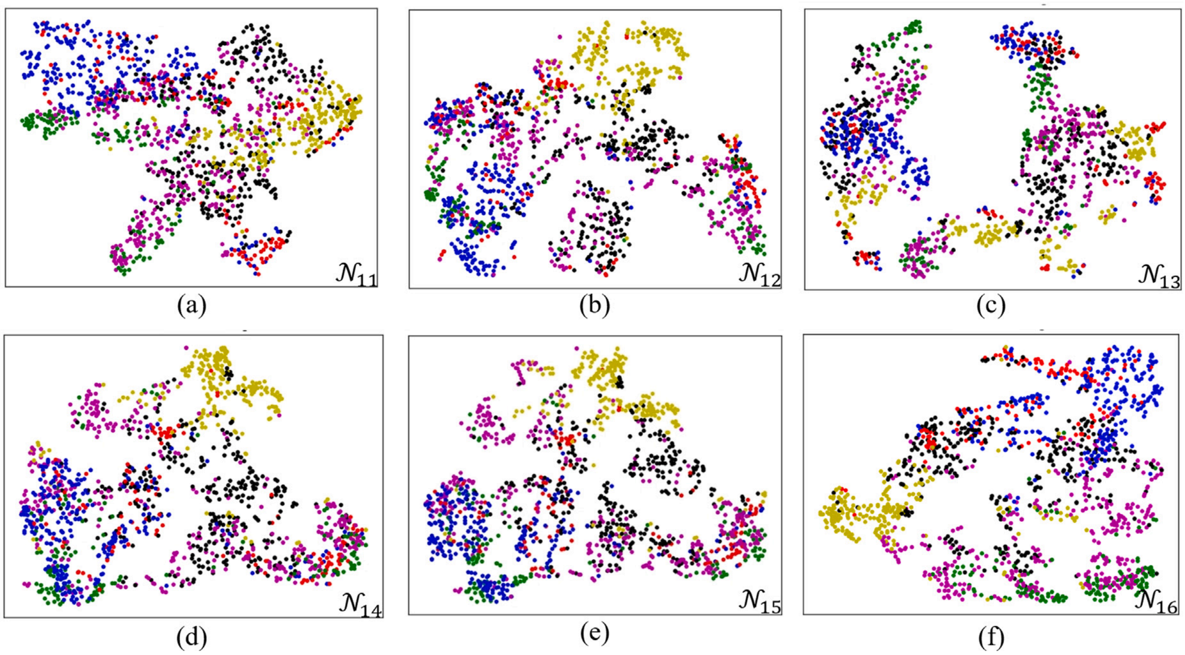
On the MELD dataset, our method exhibits an impressive average F1 score of 66.05% and attains an accuracy performance of 67.13%. Notably, our method surpasses existing methods by at least 0.84% in terms of F1 scores. Particularly, our method outperforms COSMIC by over 0.84% on the F1 score, highlighting its superior performance. For instance, baseline methods such as HFGCN, MMGCN, DialogueGCN, and DialogueRNN without leveraging pre-trained language models achieve the F1 scores of 59.71%, 58.65%, 58.10%, and 57.03%, respectively. In contrast, our proposed method yields 66.05% F1 scores, showcasing the effectiveness of integrating the RoBERTa text features. Notably, among the baseline methods, DAG-ERC and COSMIC, which leverage RoBERTa text features, attain F1 scores of 63.65% and 65.21%, respectively. However, these scores still lag behind our proposed method.

To analyse the utility of each module in our framework, multiple ablation studies are conducted on the IEMOCAP and MELD datasets, respectively. The ablation experiments scrutinise the influence of various modules, encompassing the impact of distinct components in the proposed method (Section 5.1); the effects of Pairwise Modalities Fusion (Section 5.2); the consequences of Multiple Features Fusion (Section 5.3); (4) the implications of Multimodal Attention Learning (Section 5.4).

5.1. *Effect of different components*

To verify the effect of different components in the proposed method, the roles of the Pairwise Modalities Fusion (PMF), the Density Loss (DL) (*i.e.*, $\mathcal{L}_{pd}$ and $\mathcal{L}_{md}$), and Multimodal Attention Learning (MAL) modules are studied. The model is trained with different combinations. The performance of different networks ($\mathcal{N}_{11} \sim \mathcal{N}_{16}$) is listed in Table 2. Symbol ✓denotes that the component is used for quantisation training. The symbol × indicates that the component is excluded for training.

The comparison between the networks of $\mathcal{N}_{11}$ and $\mathcal{N}_{12}$ reveals that the incorporation of the PMF module significantly enhances model performance. The absence of PMF results in a notable degradation in performance on both IEMOCAP and MELD datasets, providing empirical evidence for the efficacy of the designed PMF. Similarly, the comparison between $\mathcal{N}_{11}$ and $\mathcal{N}_{13}$, as well as $\mathcal{N}_{15}$ and $\mathcal{N}_{16}$, demonstrates that the adoption of the density loss efficiently improves network performance, which is caused by the redundancy removing with the density loss in enhancing the overall performance of the network.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Feature visualisation using t-SNE. The t-SNE plots of the ablation experiments $\mathcal{N}_{11}$, $\mathcal{N}_{12}$, $\mathcal{N}_{13}$, $\mathcal{N}_{14}$, $\mathcal{N}_{15}$, and $\mathcal{N}_{16}$ on the IEMOCAP dataset. The features in same categories can be clustered more tightly in $\mathcal{N}_{16}$ (Ours), which means the proposed method can help to learn more information for improving the classification performance.

Table 3
Performance (Average-weight F1 score and Accuracy) under different multimodal settings. T represents text modality, A represents audio modality, and V represents visual modality.

	PMF	Modalities			IEMOCAP		MELD	
		A	V	T	F1	Acc	F1	Acc
$\mathcal{N}_{21}$	×	✓	×	×	53.40	54.22	44.57	48.64
$\mathcal{N}_{22}$	×	×	✓	×	34.61	36.35	32.83	35.15
$\mathcal{N}_{23}$	×	×	×	✓	65.01	65.68	64.47	64.82
$\mathcal{N}_{24}$	✓	✓	×	✓	66.29	66.59	64.84	65.34
$\mathcal{N}_{25}$	✓	×	✓	✓	64.76	65.22	64.75	65.07
$\mathcal{N}_{26}$	✓	✓	✓	✓	**69.34**	**69.57**	**66.05**	**67.13**

Table 4
Ablation study of the different variants impact on multiple features fusion (Average-weight F1 score and Accuracy).

	GCN	s_t	$h_t^{(Bt)}$	$E^{(Bt)}$	IEMOCAP		MELD	
					F1	Acc	F1	Acc
$\mathcal{N}_{31}$	×	×	✓	✓	65.99	66.67	64.28	64.42
$\mathcal{N}_{32}$	✓	✓	×	×	67.89	68.15	58.56	59.25
$\mathcal{N}_{33}$	×	✓	✓	✓	66.67	66.61	64.67	65.41
$\mathcal{N}_{34}$	✓	✓	✓	×	67.64	68.08	65.37	66.17
$\mathcal{N}_{35}$	✓	✓	×	✓	68.14	68.52	59.17	59.72
$\mathcal{N}_{36}$	✓	×	✓	✓	68.10	68.27	65.73	66.47
$\mathcal{N}_{37}$	✓	✓	✓	✓	**69.34**	**69.57**	**66.05**	**67.13**

Moreover, through a comparative analysis of the outcomes achieved by $\mathcal{N}_{14}$ and $\mathcal{N}_{16}$ employing the MAL module, it becomes apparent that the incorporation of MAL leads to enhanced performance on both the IEMOCAP and MELD datasets. This underscores the crucial role of MAL in augmenting model performance and emphasises the positive impact of these modules on overall model effectiveness.

In order to better visualise the performance of different networks, Fig. 4 presents the feature distributions of the network $\mathcal{N}_{11} \sim \mathcal{N}_{16}$ on the IEMOCAP dataset (a~f) with t-SNE [70]. Fig. 4 shows the distribution for classification by transferring data from high dimensions into the two-dimensional space [71]. It is evident that the features in the same categories can be clustered more tightly in $\mathcal{N}_{16}$ (Ours in Fig. 4f), which means the proposed method can help to learn more information for improving the classification performance.

5.2. Impact of pairwise modalities fusion

To evaluate the impact of different components in Pairwise Modalities Fusion, Table 3 presents the performance of different networks ($\mathcal{N}_{21} \sim \mathcal{N}_{26}$), trained with and without PMF, employing different modality combinations. The networks ($\mathcal{N}_{21} \sim \mathcal{N}_{23}$) are individually trained with a single modality: $\mathcal{N}_{21}$ with the Audio modality, $\mathcal{N}_{22}$ with the Video modality, and $\mathcal{N}_{23}$ with the Text modality. Networks ($\mathcal{N}_{24}$ and $\mathcal{N}_{25}$) are trained with one pairwise modalities fusion: $\mathcal{N}_{24}$ with Audio and Text modalities, and $\mathcal{N}_{25}$ with Video and Text modalities.

The network $\mathcal{N}_{26}$ is trained with all three modalities, and it is the proposed method. By comparing the performance of $\mathcal{N}_{21} \sim \mathcal{N}_{23}$, it becomes evident that within the unimodal framework, the text modality significantly outperforms both the audio and video modalities. This finding suggests that the features inherent to the text modality exhibit superior effectiveness, particularly in the context of the MELD and IEMOCAP datasets. Analysing the outcomes of $\mathcal{N}_{24} \sim \mathcal{N}_{25}$ in comparison to $\mathcal{N}_{21} \sim \mathcal{N}_{23}$ highlights the effectiveness of the designed Pairwise Modalities Fusion (PMF) in enhancing network performance. The results show that the multimodal configuration outperforms the unimodal settings. The performance improvement is evident from the outcomes of $\mathcal{N}_{24} \sim \mathcal{N}_{25}$ and $\mathcal{N}_{26}$, showcasing the proposed method's capability to fuse information from all three modalities.

5.3. Impact of multiple features fusion

To evaluate the impact of each part in Multiple Features Fusion, the networks of $\mathcal{N}_{31} \sim \mathcal{N}_{37}$ are trained with different parts in the Multiple Features Fusion (*i.e.*, GCN structure, s_t, $h_t^{(Bt)}$ and $E^{(Bt)}$), and the results are listed in Table 4.

To assess the influence of employing GCN, the results of $\mathcal{N}_{31}$ and $\mathcal{N}_{33}$ are compared with $\mathcal{N}_{36}$ and $\mathcal{N}_{37}$. Notably, omitting GCN leads to decreases in F1 score on both the IEMOCAP and MELD datasets, emphasising the substantial impact of GCN on model performance.

Table 5
Ablation study of the different variants impact on multimodal attention learning (Average-weight F1 score and Accuracy).

	$h_t^{(Bt)}$	$h_t^{(\mathrm{VA})}$	ATT	IEMOCAP		MELD	
				F1	Acc	F1	Acc
$\mathcal{N}_{41}$	×	×	×	67.22	67.31	64.75	65.16
$\mathcal{N}_{42}$	×	×	✓	67.74	67.93	64.82	65.43
$\mathcal{N}_{43}$	✓	×	×	67.57	67.74	64.91	65.24
$\mathcal{N}_{44}$	✓	✓	×	68.23	68.33	65.27	65.74
$\mathcal{N}_{45}$	✓	×	✓	68.14	68.02	65.42	66.13
$\mathcal{N}_{46}$	×	✓	✓	68.47	68.75	65.15	65.68
$\mathcal{N}_{47}$	✓	✓	✓	**69.34**	**69.57**	**66.05**	**67.13**

Furthermore, the comparison between $\mathcal{N}_{31}$ and $\mathcal{N}_{33}$, as well as $\mathcal{N}_{36}$ and $\mathcal{N}_{37}$, reveals that the inclusion of s_t contributes to enhanced model performance on both the IEMOCAP and MELD datasets. Analysing different networks (*i.e.*, $\mathcal{N}_{32}$ and $\mathcal{N}_{34}$, $\mathcal{N}_{35}$ and $\mathcal{N}_{37}$), it is evident that incorporating $h_t^{(Bt)}$ significantly improves the Average-weight F1-score on both the IEMOCAP and MELD datasets. This highlights the substantial impact of $h_t^{(Bt)}$ features, especially on the MELD dataset. It can be observed that the enhancement in performance with $h_t^{(Bt)}$ is evident on both the MELD and IEMOCAP datasets, with a more significant improvement observed in the case of the MELD dataset. Notably, the IEMOCAP dataset is derived from continuous dialogues, whereas the MELD dataset, being drawn from TV shows, often lacks the continuity typical of continuous dialogues [47]. In this context, $h_t^{(Bt)}$ demonstrates effectiveness in alleviating performance degradation resulting from non-continuous dialogues. Additionally, the removal of $E^{(Bt)}t$ results in F1 score decreases on the IEMOCAP and MELD datasets when comparing the results of $\mathcal{N}_{34}$ and $\mathcal{N}_{37}$, indicating the positive role of the Transformer encoder in improving model performance.

5.4. Impact of multimodal attention learning

Table 5 illustrates the impact of different variants in multimodal attention learning, encompassing the utilisation of multi-head self-attention, the incorporation of $h_t^{(\mathrm{VA})}$ with skip-connection, and the integration of $h_t^{(Bt)}$ with skip-connection. By examining the performance of different networks (*i.e.*, $\mathcal{N}_{41}$ with $\mathcal{N}_{43}$, $\mathcal{N}_{42}$ with $\mathcal{N}_{45}$, $\mathcal{N}_{46}$ with $\mathcal{N}_{47}$), the incorporation of $h_t^{(Bt)}$ is shown to enhance network performance. Notably, a significant performance decrease is observed when omitting $h_t^{(\mathrm{VA})}$, as seen in the comparison between $\mathcal{N}_{42}$ and $\mathcal{N}_{46}$, as well as $\mathcal{N}_{45}$ and $\mathcal{N}_{47}$. This indicates that by concatenating $h_t^{(\mathrm{VA})}$ and $h_t^{(\mathrm{Bt})}$ through a skip-connection, the degradation phenomenon resulting from an increase in network layers can be mitigated to some extent. Moreover, comparing the performance of different networks (*i.e.*, $\mathcal{N}_{41}$ with $\mathcal{N}_{42}$, $\mathcal{N}_{43}$ with $\mathcal{N}_{45}$, $\mathcal{N}_{44}$ with $\mathcal{N}_{47}$) indicates that the utilisation of multi-head self-attention improves network performance. This underscores the effectiveness of the multi-head attention mechanism in comprehending contextual information. The multi-head attention mechanism can capture correlations between all elements in the sequence, addressing the limitation of RNN-based networks in capturing distant contextual relationships. Therefore, integrating the multi-head attention mechanism can effectively enhance model performance. These experimental results underscore the effectiveness of incorporating skip-connection and multi-head self-attention in enhancing the performance of our model in the context of attention emotion recognition.

6. Conclusion

In this paper, we propose a novel multimodal fusion framework to tackle the challenges associated with integrating information from diverse modalities. The introduction of a novel pairwise fusion approach enhances integration efficiency, deviating from conventional direct fusion methods.

The designed density loss, incorporating L1 norm regularisation, successfully addresses redundancy issues, preventing overfitting and facilitating better generalisation. Our contributions are substantiated by extensive experiments on the IEMOCAP and MELD datasets, where our framework consistently outperforms existing methods. The demonstrated superior performance validates the efficacy of our approach in handling real-world multimodal data. In essence, our work makes significant strides in advancing the field of multimodal fusion by offering practical solutions to challenges such as complexity reduction, redundancy mitigation, and adaptive information capture. These contributions are poised to impact a wide range of applications, from emotion recognition to dialogue generation, providing researchers and practitioners with effective tools to enhance the efficiency and accuracy of multimodal models. As multimodal data becomes increasingly prevalent, our framework stands as a valuable contribution to the ongoing evolution of this important research area. In future work, we will improve the explainability of our model in decision-making [72] and leverage external knowledge and pragmatic processing techniques [73] to enhance the performance of the proposed method.

CRediT authorship contribution statement

Chunxiao Fan: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Jie Lin:** Software, Resources, Methodology, Investigation. **Rui Mao:** Writing – review & editing, Methodology, Conceptualization. **Erik Cambria:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

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

Data availability

The work is implemented on public data.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China (No. 2022YFC3803200), the Natural Science Foundation of China (No. 61802105), the University Synergy Innovation Program of Anhui Province, China (No. GXXT-2021-005 and GXXT-2022–033), and the Fundamental Research Funds for the Central Universities, China (No. JZ2022HGTB0250 and PA2023IISL0096).

References

[1] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.

[2] C. Huang, O.R. Zaiane, A. Trabelsi, N. Dziri, Automatic dialogue generation with expressed emotions, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, pp. 49–54.

[3] A. Chatterjee, K.N. Narahari, M. Joshi, P. Agrawal, SemEval-2019 task 3: EmoContext contextual emotion detection in text, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 39–48.

[4] T. Yue, R. Mao, H. Wang, Z. Hu, E. Cambria, KnowleNet: Knowledge fusion network for multimodal sarcasm detection, Inf. Fusion 100 (2023) 101921, http://dx.doi.org/10.1016/j.inffus.2023.101921.

[5] Y. Ma, R. Mao, Q. Lin, P. Wu, E. Cambria, Multi-source aggregated classification for stock price movement prediction, Inf. Fusion 91 (2023) 515–528, http://dx.doi.org/10.1016/j.inffus.2022.10.025.

[6] Y. Ma, R. Mao, Q. Lin, P. Wu, E. Cambria, Quantitative stock portfolio optimization by multi-task learning risk and return, Inf. Fusion 104 (2024) 102165, http://dx.doi.org/10.1016/j.inffus.2023.102165.

[7] Y. Li, M. El Habib Daho, P.-H. Conze, H. Al Hajj, S. Bonnin, H. Ren, N. Manivannan, S. Magazzeni, R. Tadayoni, B. Cochener, et al., Multimodal information fusion for glaucoma and diabetic retinopathy classification, in: International Workshop on Ophthalmic Medical Image Analysis, Springer, 2022, pp. 53–62.

[8] T. Zhang, M. Shi, Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease, J. Neurosci. Methods 341 (2020) 108795, http://dx.doi.org/10.1016/j.jneumeth.2020.108795.

[9] S.Y. Boulahia, A. Amamra, M.R. Madi, S. Daikh, Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, Mach. Vis. Appl. 32 (6) (2021) 121.

[10] S. Pang, D. Morris, H. Radha, CLOCs: Camera-LiDAR object candidates fusion for 3D object detection, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2020, pp. 10386–10393.

[11] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, A.G. Hauptmann, Multimedia classification and event detection using double fusion, Multimedia Tools Appl. 71 (2014) 333–347.

[12] S. Tang, Z. Luo, G. Nan, J. Baba, Y. Yoshikawa, H. Ishiguro, Fusion with hierarchical graphs for multimodal emotion recognition, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2022, pp. 1288–1296.

[13] J.K. Chen, Z. Chen, Z. Chi, H. Fu, Emotion recognition in the wild with feature fusion and multiple kernel learning, ACM (2014).

[14] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, M. Bartlett, Multiple kernel learning for emotion recognition in the wild, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 517–524.

[15] F. Liu, L. Zhou, C. Shen, J. Yin, Multiple kernel learning in the primal for multimodal Alzheimer's disease classification, IEEE J. Biomed. Health Inf. 18 (3) (2013) 984–990.

[16] Z. Ghahramani, M. Jordan, Factorial hidden Markov models, Adv. Neural Inf. Process. Syst. 8 (1995).

[17] T. Baltrušaitis, N. Banda, P. Robinson, Dimensional affect recognition using continuous conditional random fields, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG, IEEE, 2013, pp. 1–8.

[18] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 689–696.

[20] J. Hu, Y. Liu, J. Zhao, Q. Jin, MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation, 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), Association for Computational Linguistics, 2021, pp. 5666–5675.

[21] J. Li, X. Wang, G. Lv, Z. Zeng, Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition, IEEE Trans. Multimed. (2023).

[22] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling, in: Proceedings of the International Speech Communication Association (Interspeech) 2010, 2010, pp. 2362–2365, http://dx.doi.org/10.21437/Interspeech.2010-646.

[23] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 873–883.

[24] W. Jiao, H. Yang, I. King, M.R. Lyu, HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 397–406, http://dx.doi.org/10.18653/v1/N19-1037.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.

[26] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019, pp. 527–536, http://dx.doi.org/10.18653/v1/P19-1050.

[27] R. Mao, X. Li, Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification, Proc. AAAI Conf. Artif. Intell. 35 (15) (2021) 13534–13542, http://dx.doi.org/10.1609/aaai.v35i15.17596.

[28] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3829–3839.

[29] R. Mao, Q. Liu, K. He, W. Li, E. Cambria, The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, IEEE Trans. Affect. Comput. 14 (3) (2023) 1743–1753, http://dx.doi.org/10.1109/TAFFC.2022.3204972.

[30] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, DialogueGCN: A graph convolutional neural network for emotion recognition in conversation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 154–164, http://dx.doi.org/10.18653/v1/D19-1015.

[31] A. Joshi, A. Bhat, A. Jain, A. Singh, A. Modi, COGMEN: COntextualized GNN based multimodal emotion recognition, in: M. Carpuat, M.-C. de Marneffe, I.V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4148–4164, http://dx.doi.org/10.18653/v1/2022.naacl-main.306.

[32] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2022, pp. 7037–7041.

[33] Y. Fu, S. Okada, L. Wang, L. Guo, Y. Song, J. Liu, J. Dang, CONSK-GCN: conversational semantic-and knowledge-oriented graph convolutional network for multimodal emotion recognition, in: 2021 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2021, pp. 1–6.

[34] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, W. Gao, Late fusion incomplete multi-view clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (10) (2018) 2410–2423.

[35] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl.-Based Syst. 161 (2018) 124–133.

[36] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1103–1114, http://dx.doi.org/10.18653/v1/D17-1115.

[37] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, 2018, arXiv preprint arXiv:1806.00064.

[38] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018, pp. 5634–5641.

[39] F. Chen, Z. Sun, D. Ouyang, X. Liu, J. Shao, Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1064–1073.

[40] S. Poria, E. Cambria, A. Gelbukh, 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, 2015, pp. 2539–2544.

[41] W. Li, L. Zhu, R. Mao, E. Cambria, SKIER: A symbolic knowledge integrated model for conversational emotion recognition, Proc. AAAI Conf. Artif. Intell. 37 (11) (2023) 13121–13129, http://dx.doi.org/10.1609/aaai.v37i11.26541.

[42] Q. Lin, J. Liu, R. Mao, F. Xu, E. Cambria, TECHS: Temporal logical graph networks for explainable extrapolation reasoning, in: Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL, Vol. 1, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1281–1293, http://dx.doi.org/10.18653/v1/2023.acl-long.71.

[43] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, stat 1050 (20) (2017) 10–48550.

[44] J. Chen, T. Ma, C. Xiao, FastGCN: Fast learning with graph convolutional networks via importance sampling, in: Proceedings of the International Conference on Learning Representations, ICLR, 2018, pp. 1–15.

[45] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. 30 (2017).

[46] D. Hu, L. Wei, X. Huai, DialogueCRN: Contextual reasoning networks for emotion recognition in conversations, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 7042–7052, http://dx.doi.org/10.18653/v1/2021.acl-long.547.

[47] W. Shen, S. Wu, Y. Yang, X. Quan, Directed acyclic graph network for conversational emotion recognition, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 1551–1560, http://dx.doi.org/10.18653/v1/2021.acl-long.123.

[48] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Vol. 2018, 2018, p. 2122.

[49] S. Sahay, S.H. Kumar, R. Xia, J. Huang, L. Nachman, Multimodal relational tensor network for sentiment and emotion classification, in: A. Zadeh, P.P. Liang, L.-P. Morency, S. Poria, E. Cambria, S. Scherer (Eds.), Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 20–27, http://dx.doi.org/10.18653/v1/W18-3303.

[50] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive RNN for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6818–6825.

[51] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, COSMIC: COmmonsense knowledge for emotion identification in conversations, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2470–2481.

[52] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017, pp. 1–14.

[53] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[54] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.

[55] L. Zhu, W. Li, R. Mao, E. Cambria, HIPPL: Hierarchical intent-inferring pointer network with pseudo labeling for consistent persona-driven dialogue generation, IEEE Comput. Intell. Mag. (2024).

[56] R. Mao, K. He, X. Zhang, G. Chen, J. Ni, Z. Yang, E. Cambria, A survey on semantic processing techniques, Inf. Fusion 101 (2024) 101988, http://dx.doi.org/10.1016/j.inffus.2023.101988.

[57] X. Zhang, R. Mao, E. Cambria, A survey on syntactic processing techniques, Artif. Intell. Rev. 56 (2023) 5645–5728, http://dx.doi.org/10.1007/s10462-022-10300-7.

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[59] M. Ge, R. Mao, E. Cambria, Explainable metaphor identification inspired by conceptual metaphor theory, Proc. AAAI Conf. Artif. Intell. 36 (10) (2022) 10681–10689, http://dx.doi.org/10.1609/aaai.v36i10.21313.

[60] R. Mao, X. Li, K. He, M. Ge, E. Cambria, MetaPro Online: A computational metaphor processing online system, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 127–135.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[62] S. Han, R. Mao, E. Cambria, Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings, in: Proceedings of the 29th International Conference on Computational Linguistics, COLING, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 94–104.

[63] X. Zhang, R. Mao, K. He, E. Cambria, Neurosymbolic sentiment analysis with dynamic word sense disambiguation, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 8772–8783.

[64] F. Liu, X. Ren, Z. Zhang, X. Sun, Y. Zou, Rethinking skip connection with layer normalization, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 3586–3598.

[65] T. Kim, P. Vossen, Emoberta: Speaker-aware emotion recognition in conversation with RoBERTa, 2021, arXiv preprint arXiv:2108.12009.

[66] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[67] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283.

[68] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, Speech Commun. 53 (9–10) (2011) 1062–1087.

[69] S. Tang, Z. Luo, G. Nan, J. Baba, Y. Yoshikawa, H. Ishiguro, Fusion with hierarchical graphs for multimodal emotion recognition, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022, pp. 1288–1296, http://dx.doi.org/10.23919/APSIPAASC55919.2022.9979932.

[70] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[71] R. Mao, K. Du, Y. Ma, L. Zhu, E. Cambria, Discovering the cognition behind language: Financial metaphor analysis with MetaPro, in: 2023 IEEE International Conference on Data Mining, ICDM, IEEE, 2023, pp. 1–6.

[72] E. Cambria, R. Mao, M. Chen, Z. Wang, S.-B. Ho, Seven pillars for the future of artificial intelligence, IEEE Intell. Syst. 38 (6) (2023) 62–69.

[73] R. Mao, X. Li, M. Ge, E. Cambria, MetaPro: A computational metaphor processing model for text pre-processing, Inf. Fusion 86–87 (2022) 30–43, http://dx.doi.org/10.1016/j.inffus.2022.06.002.