Adaptive Graph Learning for Multimodal Conversational Emotion Detection

Geng Tu[1,2*], Tian Xie[1*], Bin Liang[1,3], Hongpeng Wang[1†], Ruifeng Xu[1,2,4†]

[1]Harbin Institute of Technology, Shenzhen, China
[2]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[3]The Chinese University of Hong Kong, Hong Kong, China
[4]Peng Cheng Laboratory, Shenzhen, China
tugeng0313@gmail.com, xuruifeng@hit.edu.cn

Abstract

Multimodal Emotion Recognition in Conversations (ERC) aims to identify the emotions conveyed by each utterance in a conversational video. Current efforts encounter challenges in balancing intra- and inter-speaker context dependencies when tackling intra-modal interactions. This balance is vital as it encompasses modeling self-dependency (emotional inertia) where speakers' own emotions affect them and modeling interpersonal dependencies (empathy) where counterparts' emotions influence a speaker. Furthermore, challenges arise in addressing cross-modal interactions that involve content with conflicting emotions across different modalities. To address this issue, we introduce an adaptive interactive graph network (IGN) called AdaIGN that employs the Gumbel Softmax trick to adaptively select nodes and edges, enhancing intra- and cross-modal interactions. Unlike undirected graphs, we use a directed IGN to prevent future utterances from impacting the current one. Next, we propose Node- and Edge-level Selection Policies (NESP) to guide node and edge selection, along with a Graph-Level Selection Policy (GSP) to integrate the utterance representation from original IGN and NESP-enhanced IGN. Moreover, we design a task-specific loss function that prioritizes text modality and intra-speaker context selection. To reduce computational complexity, we use pre-defined pseudo labels through self-supervised methods to mask unnecessary utterance nodes for selection. Experimental results show that AdaIGN outperforms state-of-the-art methods on two popular datasets. Our code will be available at https://github.com/TuGengs/AdaIGN.

Introduction

Emotion recognition in conversations (ERC) has garnered considerable attention due to its valuable applications in recommendation systems (Zheng et al. 2022), dialogue generation (Zhu et al. 2022), and so on. Most studies on ERC focus primarily on the textual modality, including recurrent neural networks (RNNs) (Majumder et al. 2019), memory networks (Jiao, Lyu, and King 2020), and graph-based models (Saxena, Huang, and Kurohashi 2022).

Despite the progress, text alone cannot provide sufficient cues for deeper feelings compared to multimodal perception (Hazarika et al. 2018). Existing multimodal ERC methods mainly focus on aggregation-based fusion by concatenation (Tu et al. 2022b), tensor product (Mai, Hu, and Xing 2019; Liu et al. 2018), attention network (Rahman et al. 2020; Wang et al. 2019) or heterogeneous graph (Yang et al. 2021; Hu et al. 2022), etc. For instance, Hazarika et al. (2018) proposed a conversational memory network to align features from multiple views. Lian, Liu, and Tao (2021) introduced a cross-modal transformer for implicit enhancement. Hu et al. (2021) explored undirected graph-based fusion to capture intra- and cross-modal interactions.

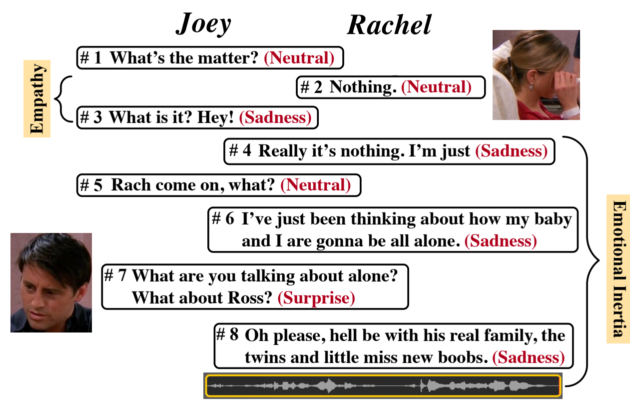

Figure 1: Examples of utterances in a conversation. The golden labels of utterances are highlighted in red font.

However, they have limitations in modeling intra- and cross-modal interactions: (1) **Future utterances affecting the emotion detection of the current one.** Previous approaches in modeling intra-modal interactions have often relied on using future utterances to predict the current one's emotion. However, this approach is not practical in real-world situations. (2) **The difficulty of balancing empathy and emotional inertia.** Emotional dynamics of conversations (Poria et al. 2019) covers two main aspects: self-dependency (emotional inertia), where a speaker's own emotions impact them, and interpersonal dependencies (empathy), where a speaker's emotions are influenced by their counterparts. Striking the right balance between empathy and emotional inertia poses a significant challenge in modeling intra-modal interactions. This balance fundamentally

[*]These authors contributed equally.
[†]Corresponding authors.

involves harmonizing inter- and intra-speaker contexts. As depicted in Fig. 1, discerning the emotion behind Utterance 8 necessitates assigning greater significance to intra-speaker context (Utterances 4, 6, etc.) rather than inter-speaker context (Utterances 5, 7, etc.). Unfortunately, previous studies have overlooked the balance between these two contexts. (3) **Multimodal information involves content with conflicting emotions.** When modeling cross-modal interactions, certain utterances exhibit conflicting emotions across different modalities. As illustrated in Fig. 1, the 2nd utterance visually conveys sadness through tear wiping, while the text itself appears devoid of emotion. Existing research has yet to offer a solution for such discrepancies.

To address the above problems, we present a novel adaptive interactive graph network (IGN), called AdaIGN, which is guided by Node- and Edge-level Selection Policies (NSP and ESP, collectively known as NESP) as well as a Graph-level Selection Policy (GSP). IGN is a directed graph to prevent future utterances from affecting the current one. To jointly optimize these policies with network weights, we employ standard back-propagation along with the Gumbel Softmax trick (Jang, Gu, and Poole 2016). Specifically, the NSP is employed to select nodes across multiple modalities within an utterance. The ESP is capable of selecting distinct contextual edges (including inter- and intra-speaker contexts) of each node within the same modality. The GSP integrates the two graphs by selecting utterance representations at the graph level. Furthermore, we introduce a task-specific loss function based on the keep-or-drop strategy, prioritizing the selection of text modality and intra-speaker context to meet the ERC task. To reduce computational complexity, we leverage pseudo labels generated via self-supervised methods to mask unnecessary utterance nodes for selection and freeze the gradient of their corresponding selection strategies. In summary, our contributions are as follows:

- We propose a novel AdaIGN that enhances intra- and cross-modal interactions by dynamically selecting nodes or edges. And the task-specific loss function based on the keep-or-drop strategies is designed to meet the ERC task.
- To optimize the computational complexity of selection policies, we employ predefined pseudo-labels to mask out utterances that do not require selection.
- Experimental results on two popular ERC datasets show that our AdaIGN outperforms state-of-the-art methods.

Methodology

In this section, we provide a detailed introduction to each component of the proposed AdaIGN, as depicted in Fig. 2.

Task Definition

Let $U = [u_{(1)}, ..., u_{(N)}]$ be a conversation uttered by $M \geq 2$ speakers, consisting of N utterances. Each utterance $u_{(i)}$ is represented by a triplet $u_{(i)} = \{u^a_{(i)}, u^v_{(i)}, u^t_{(i)}\}$. $u^a_{(i)} \in \mathbb{R}^{d_a}$, $u^v_{(i)} \in \mathbb{R}^{d_v}$, and $u^t_{(i)} \in \mathbb{R}^{d_t}$ denote the acoustic, visual, and text features of $u_{(i)}$, respectively. Multimodal ERC aims to predict the emotion label $e_{(i)}$ of each utterance u_i based on its historical utterances $u_{(j)}$ where $\forall j < i$.

Feature Representation

Following (Ghosal et al. 2020a), we employ layer normalization and average operation on the last four hidden layers of the Roberta Large model (Liu et al. 2019) to obtain textual features. For extracting acoustic and visual features, we utilize OpenSmile (Schuller et al. 2011), an audio feature extraction toolkit, and a pre-trained DenseNet model (Huang et al. 2017) as per previous works (Hu et al. 2022).

Utterance-level Encoder

To capture context information and handle the inconsistent dimensions in multimodal data, we use a bi-directional GRU (BiGRU) $GRU_m \in \mathbb{R}^{d_h \times d_t}$ for text modality and a fully connected layer $\mathcal{F}^\xi \in \mathbb{R}^{d_h \times d_{a/v}}$ for acoustic and visual modalities, to map the feature sequence $u^\eta_{(i)}$ of each modality $\eta \in \{a, v, t\}$ to a fixed-size representation $m^\eta_{(i)} \in \mathbb{R}^{d_h}$.

$$m^t_{(i)}, h^t_{(i)} = \overleftrightarrow{GRU}_m(u^t_{(i)}, h^t_{(i-1)}) \tag{1}$$

$$m^\xi_{(i)} = \mathcal{F}^\xi(u^\xi_{(i)} |\, \theta^\xi_e),\ \xi \in \{a, v\} \tag{2}$$

where $h^t_{(i)}$ is the hidden state. $\mathcal{F}^\xi$ assigns separate parameters θ^ξ_e for acoustic and visual modalities. Considering the significance of speakers in ERC (Ong et al. 2022), we employ another BiGRU $GRU_p \in \mathbb{R}^{d_h \times d_{a/v/t}}$ to capture speaker-specific features $s^\eta_{(i)} \in \mathbb{R}^{d_h}$, as follows:

$$\widehat{m}^\eta_{(i)} = m^\eta_{(i)} + \lambda^\eta s^\eta_{(i)} \tag{3}$$

$$s^\eta_{(i)}, \widehat{h}^\eta_{(i)} = \overleftrightarrow{GRU}_p(u^\eta_{(i)}, \widehat{h}^\eta_{(k)}), 1 \leq k < i, \tag{4}$$

where $h^\eta_{(k)}$ is the hidden state of the k-th utterance spoken by the same speaker as in the i-th utterance. λ^η is a manually set hyperparameter that indicates the weight of the speaker information for each modality. $s^\eta_{(i)} \in \mathbb{R}^{d_h}$ is the speaker-specific features for speaker. $GRU_p \in \mathbb{R}^{d_h \times d_{a/v/t}}$ is used to integrate speaker information.

Adaptive Interactive Graph Network

Graph Structure We suggest a multimodal directed graph network $\mathcal{G}_d = \{\nu_d, \delta_d, \mathcal{P}^\nu_d, \mathcal{P}^\delta_d\}$ to ensure that the prediction of the current utterance is not influenced by future utterances. $\mathcal{P}^\nu_d$ and $\mathcal{P}^\delta_d$ denotes a set of NSP and ESP. And we also build another graph network $\mathcal{G}_\iota = \{\nu_\iota, \delta_\iota\}$. $\nu_{\iota/d}$ and δ_d represent a set of graph nodes and edges, respectively. $\mathcal{G}_{\iota/d}$ comprises $3 \times N$ nodes for a conversation, with $\widehat{m}^\eta_{(i)} \in \mathbb{R}^{d_h}$ represented by three nodes of the i-th utterance. Intra- and cross-modal interactions are modeled using a set of edges $\delta_{\iota/d}$ that follow two rules: (1) Nodes from the same modality are connected in a conversation, and (2) Three nodes from different modalities are connected in an utterance. The weight between nodes i and j, denoted by $\mathcal{W}^\eta_{(ij)}$, $\forall i < j$, is calculated using the cosine similarity function $sim(.)$ as $1 - arccos(sim(\widehat{m}^\eta_{(i)}, \widehat{m}^\eta_{(j)}))/\pi$.

Node- and Edge-level Selection Policies To achieve NSP and ESP, we designed a binary random variable $\varrho^{\nu/\delta}_{(\zeta)} \in \mathbb{R}^{N \times 2}$ for each node and its corresponding edges. Specifically, $\varrho^\nu_{(\zeta)} \in \mathcal{P}^\nu_d$ and $\varrho^\delta_{(\zeta)} \in \mathcal{P}^\delta_d$ determine whether the node

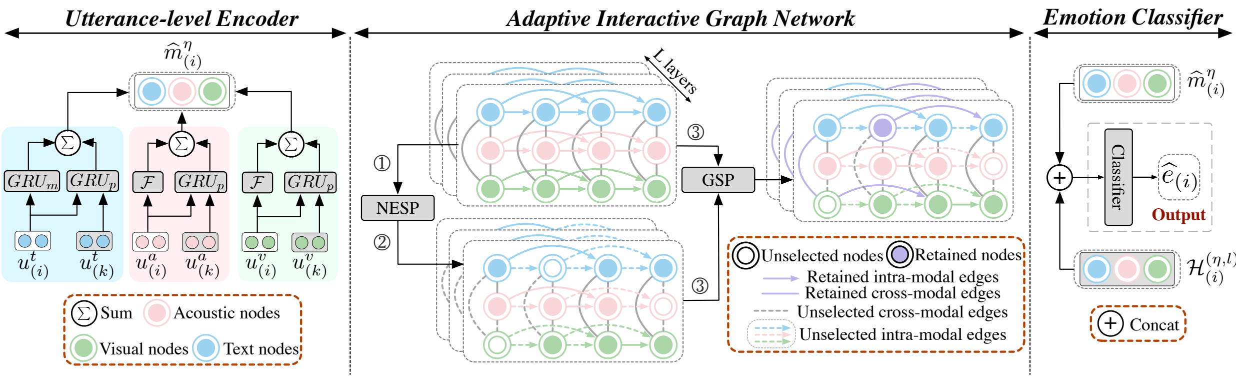

Figure 2: Illustration of AdaIGN framework during the training phase. Mathematical symbols in the illustration are in line with the formulas in paper text. Unselected nodes and edges mean their probability of being selected is less than 0.5, but they still have the potential to retain more than 0.5 after GSP. For example, the probability of NSP for a node is 0.3, while the probability of GSP for $\mathcal{G}_d$ is 0.6 and for $\mathcal{G}_\iota$ is 0.4. So 58% $(0.3 \times 0.6 + 0.4 \times 1)$ node information is retained.

and edges of the ζ-th utterance are selected. With ESP, we divide the context into self- and inter-speaker categories, allowing $\mathcal{P}_d^\delta$ to select which category of contexts. Instead of manually adjusting these selection policies, we use standard back-propagation to jointly learn the network weights θ and $\mathcal{P}_d^{\nu/\delta}$. However, optimizing the non-differentiable policies is challenging. To overcome this problem, we adopt Gumbel Softmax Sampling (Jang, Gu, and Poole 2016).

Gumbel Softmax Sampling Let $\Pi = [\pi_{(1)}, \pi_{(2)}, ..., \pi_{(N)}]$ be a set of distribution vectors of the binary random variable $\gamma \in [0, 1]$ in a conversation, where $\pi_{(\zeta)} = [1-\gamma_{(\zeta)}, \gamma_{(\zeta)}] \in \mathbb{R}^2$. In Gumbel Softmax Sampling, instead of directly sampling $\Gamma_{(\zeta)}$ from $\pi_{(\zeta)}$, we generate it as follows.

$$\Gamma_{(\zeta)}[\kappa] = Argmax(\psi_{(\zeta)}[\kappa] + log(\pi_{(\zeta)}[\kappa])) \quad (5)$$

where $\kappa \in \{0, 1\}$. $\psi_{(\zeta)} = -\log(-\log(\boldsymbol{v}_{(\zeta)})) \in \mathbb{R}^2$. And $\boldsymbol{v}_{(\zeta)} \in \mathbb{R}^2$ are independent and identically distributed samples drawn from the $Unif(0, 1)$ distribution. To remove the non-differentiable Argmax operation, the Gumbel Softmax trick relaxes $\mathcal{E}_o(\Gamma_{(\zeta)})$ to $\mathcal{Y}_{(\zeta)} \in \mathbb{R}^2$. $\mathcal{E}_o$ denotes the one-hot encoding to the non-differentiable results.

$$\mathcal{Y}_{(\zeta)}[k] = \frac{\exp((\log(\pi_{(\zeta)}[k]) + \psi_{(\zeta)}[k])/\tau)}{\sum_{\kappa\in\{0,1\}}\exp((\log(\pi_{(\zeta)}[\kappa]) + \psi_{(\zeta)}[\kappa])/\tau)} \quad (6)$$

where $k \in \{0, 1\}$. $\tau > 0$ is the temperature parameter. Especially, when $\tau \to 0$, $\mathcal{Y}_{(\zeta)}$ becomes the same as $\mathcal{E}_o(\Gamma_{(\zeta)})$ and the corresponding Gumbel Softmax distribution of $\mathcal{Y}_{(\zeta)}$ becomes identical to the discrete distribution $\pi_{(\zeta)}$.

Selection Policies Based on the above, we can assign an attribute value $\mathcal{O}^\nu = [\varrho^\nu_{(1)}, ..., \varrho^\nu_{(N)}]$, which represents a set of distribution vectors of the binary random variable $\gamma \in [0, 1]$ to each node ν in $\mathcal{G}_d$. $\varrho^\nu_{(\zeta)} = [1-\gamma_{(\zeta)}, \gamma_{(\zeta)}]$ and $\gamma_{(\zeta)}$ indicates the probability of the ζ-th nodes being selected in $\mathcal{G}_d$. During the training process, we employ Gumbel Softmax Sampling to generate the $\varrho^\nu_{(\zeta)}$ as follows.

$$\varrho^\nu_{(\zeta)}[k] = \frac{\exp((\log(\varrho^\nu_{(\zeta)}[k]) + \psi^\nu_{(\zeta)}[k])/\tau)}{\sum_{\kappa\in\{0,1\}}\exp((\log(\varrho^\nu_{(\zeta)}[\kappa]) + \psi^\nu_{(\zeta)}[\kappa])/\tau)} \quad (7)$$

where $\varrho^\nu_{(\zeta)}[0]$ and $\varrho^\nu_{(\zeta)}[1]$ are mutually exclusive, so the value of $\varrho^\nu_{(\zeta)}$ can only be $[0, 1]$ or $[1, 0]$ during testing. For ESP, we can add another attribute value $\mathcal{O}^\delta$ to generate the policy $\varrho^\delta_{(\zeta)} \in \mathcal{P}^\delta$. $\varrho^\delta_{(\zeta)}[0]$ and $\varrho^\delta_{(\zeta)}[1]$ indicate the probability of selecting inter- and intra-speaker contexts, respectively.

Pseudo Labels Because of a large number of nodes and edges, training each policy individually leads to high computational complexity. To address this, we use pseudo-labels to identify policies that do not need to be trained.

(1) We train a new graph $\widehat{\mathcal{G}_\iota}$ using two modalities, such as a and t and compared its prediction results against those of the $\mathcal{G}_\iota$. Utterances with the same prediction results are labeled as $MASK^v$, while $MASK^a$ is obtained using a similar method. We omit $MASK^t$ as modality t already exhibits superior performance in ERC (Wu et al. 2021).

(2) We remove intra-modal edges of the same modality (e.g., v) in $\mathcal{G}_\iota$ and then compare predictions to the original $\mathcal{G}_\iota$. Utterances with the same predictions are labeled as $\widehat{MASK}^v$, with ESP set to $[1, 1]$ for a probability of 1 for both intra- and inter-speaker context selection.

After the above steps, we initialize NSP and ESP:

$$\varrho^\nu_{(\zeta,\eta)} = \begin{cases} [0,1], & \zeta \in MASK^\eta \\ Gumbel{-}Softmax(\mathcal{O}^\nu_{(\zeta)}), & otherwise \end{cases} \quad (8)$$

$$\varrho^\xi_{(\zeta,\eta)} = \begin{cases} [1,1], & \zeta \in \widehat{MASK}^\eta \\ Gumbel{-}Softmax(\mathcal{O}^\xi_{(\zeta)}), & otherwise \end{cases} \quad (9)$$

where $\eta \in \{a, v, t\}$. The representations of node and edge weights are updated as follows:

$$\nu^\eta_{(i)} = \nu^\eta_{(i)} \circ (\varrho^\nu_{(i,\eta)}[1]) \quad \mathcal{W}^\eta_{ij} = \begin{cases} \mathcal{W}^\eta_{ij} \circ (\varrho^\xi_{(i,\eta)}[1]), & p_i == p_j \\ \mathcal{W}^\eta_{ij} \circ (\varrho^\xi_{(i,\eta)}[0]), & otherwise \end{cases} \quad (10)$$

where $\circ$ denotes the elementwise multiplication operator. ν refers to the nodes corresponding to the i-th utterance for η modality. i and j are the two nodes connected by an edge. During the testing process, $\varrho^\nu_{(i)}[0]$ or $\varrho^\nu_{(i)}[1]$ is determined by whether it is greater than 0.5, and similarly for $\varrho^\xi_{(i)}[0]$ or $\varrho^\xi_{(i)}[1]$. It can take on either 1 or 0 based on this condition.

Graph Convolution Operation According to (Chen et al. 2020), the graph convolution operation of $\mathcal{G}_{d/\iota}$ in a minibatch data as follows:

$$\mathcal{H}_{\alpha}^{(l)} = (1-\alpha)\widetilde{P}H^{(l-1)} + \alpha H^{(0)} \tag{11}$$

$$\mathcal{H}_{\beta}^{(l)} = (1-\Gamma(l-1))\mathcal{E} + \Gamma(l-1)W_{\beta}^{(l-1)} \tag{12}$$

$$\mathcal{H}_{d/\iota}^{(l)} = \sigma(\mathcal{H}_{\alpha}^{(l)}\mathcal{H}_{\beta}^{(l)}) \tag{13}$$

where $\widetilde{P} = \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2} \in \mathbb{R}^{3d_q\times 3d_q}$ is the graph convolution matrix with the renormalization trick (Kipf and Welling 2016) for three modalities, where $\widetilde{D} \in \mathbb{R}^{3d_q\times 3d_q}$ is the degree matrix of $\widetilde{A}$. d_q denotes the maximum sequence length of the minibatch data. $\widetilde{A} \in \mathbb{R}^{3d_q\times 3d_q}$ is the adjacency matrix of $\mathcal{G}_{d/\iota}$. $\mathcal{E} \in \mathbb{R}^{3d_q\times 3d_q}$ is the identity matrix, which represents the connection relationship between nodes. $\mathcal{H}^{(0)} \in \mathbb{R}^{3d_q\times 3d_m}$ is initialized with $\widehat{m}^{\eta}W_m, \eta \in \{a, v, t\}$. $W_m \in \mathbb{R}^{3d_h\times 3d_m}$ is a trainable parameter. $\mathcal{H}_{d/\iota}^{(l)} \in \mathbb{R}^{3d_q\times 3d_q}$ is the output of the l-th layer and $\Gamma(l) = \log(\frac{\beta}{l}) + 1$. α and β are two hyperparameters. $W_{\beta} \in \mathbb{R}^{3d_m\times 3d_m}$ is the weight matrix for the $(l-1)$-th layer. $\sigma(.)$ denotes the ReLU activation function (Agarap 2018).

IGN with GSP To avoid the high dimensionality resulting from concatenating the two representations in $\mathcal{G}_{\iota}$ and $\mathcal{G}_d$, we utilize the GSP similar to NSP and ESP by setting a policy $\varrho^g \in \mathbb{R}^2$, which is also a binary random variable and adaptively selects between the $\mathcal{H}_d^{(l)}$ and $\mathcal{H}_{\iota}^{(l)}$ to obtain the final utterance representation $\widehat{\mathcal{H}}^{(l)} \in \mathbb{R}^{N\times 3d_m}$ at the graph level.

Emotion Classifier

We utilize a linear unit to predict the emotion distributions:

$$\widehat{e}_{(i)} = Argmax(Softmax(W_c\,\mathcal{X}_{(i)} + b_c)) \tag{14}$$

$$\mathcal{X}_{(i)} = \widehat{m}_{(i)}^{(a,v,t)} \oplus \widehat{\mathcal{H}}_{(i)}^{(a,l)} \oplus \widehat{\mathcal{H}}_{(i)}^{(v,l)} \oplus \widehat{\mathcal{H}}_{(i)}^{(t,l)} \tag{15}$$

where $\oplus$ denotes the concatenation operation. $W_c \in \mathbb{R}^{d_o\times(3d_h+3d_m)}$ and $b_c \in \mathbb{R}^{d_o}$ are trainable parameters, where d_o is the number of categories of emotions. $\widehat{\mathcal{H}}_{(i)}^{(\eta,l)} \in \mathbb{R}^{d_m}$ represents the i-th utterance representation after the stack of l layers for modality η. $\widehat{e} \in \mathbb{R}^N$ is the predicting emotional label set of utterances in a conversation. The graph learning of AdaIGN is performed by minimizing $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{ce} + \underbrace{\gamma\mathcal{L}_m + \omega\mathcal{L}_k}_{keeping} + \underbrace{\phi\mathcal{L}_d + \mu\mathcal{L}_n}_{dropping} \tag{16}$$

$$\mathcal{L}_{ce} = CrossEntropy(\widehat{e}, e) + \varsigma\,\|\Theta\|_2 \tag{17}$$

where Θ is a set of projection parameters. ς represents the coefficient of L_2-regularization. $\mathcal{L}_{ce}$ is the classification loss. $\gamma, \phi, \omega, \mu$ are hyperparameters that determine the contribution of each component to $\mathcal{L}$. These four loss items are mainly used to lay constraints on selection policy learning based on the keep-or-drop strategy.

Keeping Loss The loss terms $\mathcal{L}_m$ and $\mathcal{L}_k$ are components of the keeping loss. Minimizing $\mathcal{L}_k$ encourages the selection of text modalities and intra-speaker context edges for each modality. Minimizing $\mathcal{L}_m$ encourages the selection of the other two modalities and the original graph $\mathcal{G}_{\iota}$.

$$\begin{aligned}\mathcal{L}_m =& \sum\nolimits_{n\le N}\frac{N-n}{N}\left|\varrho_{(n,a)}^{\nu}[1] - \varrho_{(n,v)}^{\nu}[1]\right| \\ &+ \sum\nolimits_{\xi\in\{a,v\}}\sum\nolimits_{n\le N}\log(\varrho_{(n,\xi)}^{\nu}[0]) + \log(\varrho^g[0])\end{aligned} \tag{18}$$

$$\begin{aligned}\mathcal{L}_k =& \sum\nolimits_{\xi\in\{a,v\}}\sum\nolimits_{n\le N}\frac{N-n}{N}\left|\varrho_{(n,t)}^{\nu}[1] - \varrho_{(n,\xi)}^{\nu}[1]\right| \\ &+ \sum\nolimits_{n\le N}\log(\varrho_{(n,t)}^{\nu}[0]) \\ &+ \sum\nolimits_{\eta,\widehat{\eta}\in\{a,v,t\}}\sum\nolimits_{n\le N}\frac{N-n}{N}\left|\varrho_{(n,\eta)}^{\delta}[1] - \varrho_{(n,\widehat{\eta})}^{\delta}[1]\right| \\ &+ \sum\nolimits_{\eta\in\{a,v,t\}}\sum\nolimits_{n\le N}\log(\varrho_{(n,\eta)}^{\delta}[0]),\ \eta\neq\widehat{\eta}\end{aligned} \tag{19}$$

where the sum of $\varrho_{(n,\eta)}^{\nu} \in \mathbb{R}^2$, $\varrho_{(n,\eta)}^{\delta} \in \mathbb{R}^2$, and $\varrho_{(n,\eta)}^{g} \in \mathbb{R}^2$ is all 1. $\varrho_{(n,\eta)}^{\nu}[0]$ and $\varrho_{(n,\eta)}^{\nu}[1]$ denotes the probability of unselecting and selecting the η modality of the n-th utterance. $\varrho_{(n,\eta)}^{\delta}[1]$ and $\varrho_{(n,\eta)}^{\delta}[0]$ represents the probability of selecting contexts of the n-th utterance, belonging to the same and different speakers within the modality η. $\varrho^g[0]$ and $\varrho^g[1]$ denotes the probability of selecting $\mathcal{G}_{\iota}$ and $\mathcal{G}_d$.

Dropping Loss By minimizing the dropping loss $\mathcal{L}_d$ and $\mathcal{L}_n$, the policies run counter to the keeping loss $\mathcal{L}_m$ and $\mathcal{L}_k$, respectively. To meet the ERC task, it is necessary for γ to be smaller than ω in the keeping loss. Similarly, in the dropping loss, μ should be smaller than ϕ. Furthermore, both γ and ϕ need to be smaller than ω and μ, respectively, as well.

$$\mathcal{L}_d = \sum\nolimits_{\xi\in\{a,v\}}\sum\nolimits_{n\le N}\log(\varrho_{(n,\xi)}^{\nu}[1]) + \log(\varrho^g[1]) \tag{20}$$

$$\mathcal{L}_n = \sum\nolimits_{n\le N}(\log(\varrho_{(n,t)}^{\nu}[1]) + \sum\nolimits_{\eta\in\{a,v,t\}}\log(\varrho_{(n,\eta)}^{\delta}[1])) \tag{21}$$

Experiments

Datasets

We benchmark AdaIGN on two well-known conversational datasets: **IEMOCAP** (Busso et al. 2008) is a dataset of interactive emotional binary motion capture recordings with ten actors in dialogues. It has 151 dialogues, and 7433 utterances, each labeled with six emotions: neutral, happy, angry, sad, excited, and frustrated. **MELD** (Poria et al. 2018) has multi-party conversation videos from the Friends TV series, with 1,433 conversations, 13,708 utterances, and 304 speakers. Utterances are labeled with emotions: anger, disgust, sadness, joy, surprise, fear, or neutral, and sentiment: positive, negative, or neutral. The data split of datasets in Table 1 is as follows (Ghosal et al. 2020a).

Experimental Settings

We perform a hyperparameter search for AdaIGN on each dataset using the validation set. The learning rates is 3e-4 for IEMOCAP and 1e-3 for MELD. We train our model using a batch size of 32 conversations with Adam optimizers. NESP and GSP are randomly initialized for policy initialization. For policy learning, we employ an Adam optimizer

Dataset	Dialogues			Utterances			Classes
	train	val	test	train	val	test	
MELD	1039	114	280	9,989	1,109	2610	7
IEMOCAP	120		31	5,810		1,623	6

Table 1: Statistics of two datasets. As the IEMOCAP dataset does not come with a predefined train/validation split, we allocate 10% of the training dialogues for validation.

with a learning rate of 2e-2. For other hyperparameters, d_a is 1582 for IEMOCAP and 300 for MELD. d_v=342, d_t=1024, d_h=200, and d_m=100. γ=0.6, ϕ=0.2, ω=0.9, and μ=0.1. λ^η is 3 (a), 0 (v), and 1 (t) for IEMOCAP; and 0.5 (a), 0.5 (v), and 1.5 (t) for MELD. The number of GCN layers l is 16 for IEMOCAP and 32 for MELD. The selection policy distribution size is set to 200 * batch size, where 200 is the max sequence length. All experiments are conducted at a single Tesla V100s-PCIE-32GB GPU. The results reported in our experiments are averages of 5 random runs on the test set.

Baselines

Aggregation-based Fusion DialogueRNN (Majumder et al. 2019) utilizes three GRUs to track speaker states and context, while DialogueGCN (Ghosal et al. 2019) tackles context propagation through a graph network; both use concatenated multimodal features. CTNet (Zhang et al. 2020) utilizes a transformer-based structure to model inter- and intra-modal interaction. SCMM (Yang et al. 2023) combines context modeling, modal interaction, and self-adaptive path selection for enhanced multi-modal representation.

Graph-based Fusion MMDFN (Hu et al. 2022) utilizes a multimodal graph with a uniform structure to represent relationships between modalities. MMGCN (Hu et al. 2021) employs a graph-based fusion module for capturing both intra- and inter-modal contextual features. CMCF-SRNet (Zhang and Li 2023) is a framework combining cross-modal interaction through a locality-constrained transformer and enhancing semantic relationships between utterances using a graph-based refinement transformer.

Overall Results

Following (Zhang and Li 2023; Chudasama et al. 2022; Hu et al. 2021), we utilize weighted F1 scores as evaluation metrics for ERC models and we also report F1 scores per class, except for Fear and Disgust classes on MELD due to insufficient training samples for statistically significant results.

Table 2 presents the results of the comparison between AdaIGN and other baseline methods. Our proposed AdaIGN demonstrates superior performance over previous approaches in terms of the weighted F1 score, establishing a new state-of-the-art benchmark. As depicted in Table 3, the exclusion of multiple selection policies from AdaIGN leads to a dynamic decrease of 3.86% and 2.76% in the F1 score on the IEMOCAP and MELD datasets respectively. This reduction serves as compelling evidence for the effectiveness of integrating multiple selection policies. Furthermore, AdaIGN achieves remarkable enhancements compared to alternative graph-based models. Specifically, on the MELD dataset, AdaIGN surpasses the CMCF-SRNet model, showcasing a substantial improvement of 4.49% in the weighted F1 score. A similar positive trend is observed on the IEMOCAP dataset, further reinforcing the efficacy of employing multiple selection policies for the ERC task.

Analysis of Various Modalities and Contexts

Table 3 presents the experimental results of IGN with different modalities and contexts removed, highlighting the importance of using multimodal data for ERC. Removing the textual modality led to significant F1 score drops of 19.63% and 20.36% on the IEMOCAP and MELD datasets, respectively, in line with previous research findings (Wu et al. 2021). Additionally, the impact of intra-speaker context on ERC performance was found to be greater than inter-speaker context (Ghosal et al. 2020a). Hence, task-specific loss items were included in selection policies to prioritize selecting the textual modality and self-speaker context, while avoiding pseudo-label annotation for the text modality.

Analysis of Emotional Conflicts

To calculate the ratio of utterances displaying emotional conflict issues, we train the IGN solely using unimodality data as input. Subsequently, we generate three sets of predictions for each modality. Inconsistencies observed among pairwise predictions indicate conflicts among these modalities. The conflict ratios in the IEMOCAP dataset are 83.73% for VA (acoustic-visual) modality, 50.71% for AT (acoustic-text) modality, and 77.45% for VT (visual-text) modality. Turning to the MELD dataset, conflict ratios within the VA, AT, and VT modalities are determined to be 27.32%, 43.60%, and 47.20% respectively, emphasizing the significant and valuable nature of exploring selection policies.

Ablation Study

In this section, we analyze the impact of various components within AdaIGN. Ablation experiments in Table 4 demonstrate that all components of AdaIGN have significantly improved results. This is further supported by the statistical analysis, where the p-value $\ll$ is 0.05 for the paired t-test.

Analysis of Selection Policies As shown in Table 4, the removal of selection policies leads to a decrease in the overall performance of the model. Experimental results on the IEMOCAP dataset demonstrate a reduction in accuracy of 2.41%, 2.28%, and 1.68%, and a decrease in F1 score of 2.72%, 2.60%, and 2.00% when NSP, ESP, and GSP are not utilized. This decrease in performance underscores the significance of these selection policies in enabling the model to dynamically select positive information. Moreover, we noted that NSP yielded the most favorable results, highlighting the effectiveness of the model in handling utterances involving emotional conflicts across various modalities.

Analysis of Loss Items Table 4 indicates that incorporating any loss item leads to an improvement in the F1 score.

Methods	IEMOCAP							MELD					
	Happy	Sad	Neutral	Angry	Excited	Frustrated	w-F1	Neutral	Surprise	Sadness	Happy	Anger	w-F1
DialogueRNN $^{\sharp}$	32.20	80.26	57.89	62.82	73.87	59.76	62.89	76.97	47.69	20.41	50.92	45.52	57.66
DialogueGCN $^{\sharp}$	51.57	80.48	57.69	53.95	72.81	57.33	62.89	75.97	46.05	19.60	51.20	40.83	56.36
CTNet $^{\natural}$	51.30	79.90	65.80	67.20	**78.70**	58.80	67.00	77.40	52.70	32.50	56.00	44.60	60.50
MMGCN $^{\sharp}$	45.14	77.16	64.36	68.82	74.71	61.40	66.26	76.33	48.15	26.74	53.02	46.09	58.31
MMDFN $^{\sharp}$	42.22	78.98	66.42	69.77	75.56	66.33	68.18	77.76	50.69	22.93	54.78	47.82	59.46
SCMM $^{\flat}$	45.37	78.76	63.54	66.05	76.70	66.18	67.53	-	-	-	-	-	59.44
CMCF-SRNet $^{\flat}$	52.20	80.90	68.80	**70.30**	76.70	61.60	69.60	-	-	-	-	-	62.30
AdaIGN (ours)	**53.04**	**81.47**	**71.26**	65.87	76.34	**67.79**	**70.74**	**79.75**	**60.53**	**43.70**	**64.54**	**56.15**	**66.79**

Table 2: Comparison Results under the multimodal setting (acoustic, visual, and textual modalities). w-F1 denotes the weighted average F1 score. $^{\sharp}$, $^{\natural}$, and $^{\flat}$ results come from (Hu et al. 2022), (Lian, Liu, and Tao 2021), and original papers, respectively.

Methods	IEMOCAP	MELD
IGN (Ours)	66.88	64.03
w/o A	65.97	63.15
w/o V	66.10	63.68
w/o T	47.25	43.67
w/o intra-speaker context	65.64	62.94
w/o inter-speaker context	66.32	63.75

Table 3: Analysis of IGN on various modalities and contexts. A, V, and T denote acoustic, visual, and textual modalities.

Methods	IEMOCAP		MELD	
	Acc	w-F1	Acc	w-F1
AdaIGN (Ours)	70.49	70.74	67.62	66.79
w/o $\mathcal{L}_k$	66.99	67.16	65.44	64.37
w/o $\mathcal{L}_m$	68.08	68.37	66.81	65.76
w/o $\mathcal{L}_d$	67.42	67.64	66.59	65.08
w/o $\mathcal{L}_n$	68.23	68.44	66.97	66.09
w/o NSP	68.08	68.02	66.02	64.57
w/o ESP	68.21	68.14	66.40	64.96
w/o GSP	68.81	68.74	66.55	65.39

Table 4: Ablation results of AdaIGN.

Specifically, adding $\mathcal{L}_k$ (selection of text modality and intra-speaker context) and $\mathcal{L}_d$ (unselection of acoustic and visual modalities) results in a significant F1 score improvement of 3.58% and 3.10% on the IEMOCAP dataset and 2.42% and 1.71% on the MELD dataset, respectively, underscoring the effectiveness of the keep-or-drop strategy for the ERC task.

Additionally, we visualize the performance of AdaIGN across varying weights for keeping and dropping loss, as illustrated in Fig. 3. With a fixed γ, as μ increases, the model's performance improves steadily until reaching its peak, after which it starts to decline. When γ is less than μ, leading to a collapse in the model's performance. Similar phenomena are also observable in ω and ϕ, emphasizing the importance of selecting text modality and intra-speaker context.

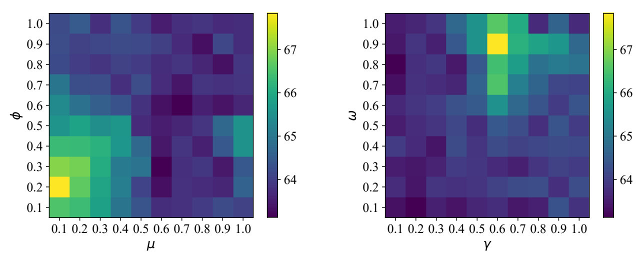

Figure 3: Performance of AdaIGN on the validation set of the IEMOCAP dataset under different loss weights.

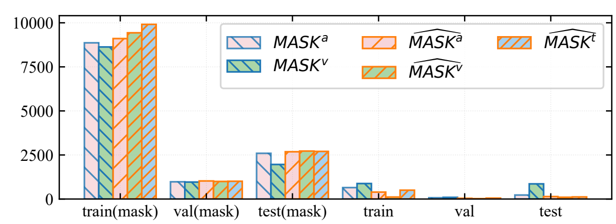

Figure 4: Numbers of selection policies that require training and non-training on the IEMOCAP dataset.

Analysis of Pseudo-labels

To address the problem of high computational complexity in training individual policies, pseudo-labels have been utilized to eliminate non-training policies. The results presented in Fig. 4 exhibit the number of selection policies that require training versus non-training ones (indicated by 'mask'). A considerable decrease is observed in the count of policies needing training, proving the significance of pseudo-labels in reducing the computational complexity of NESP.

Case Study

Unlike the training phase, the weight of selection policies is either 1 or 0 during testing. We extract mini-batch data with a weight of 1 on the GSP from the IEMOCAP dataset for the case study as shown in Fig. 5. In the 2-nd utterance, although the woman appears happy, the emotion labeled is neutral. Therefore, it is reasonable for the NSP to unselect

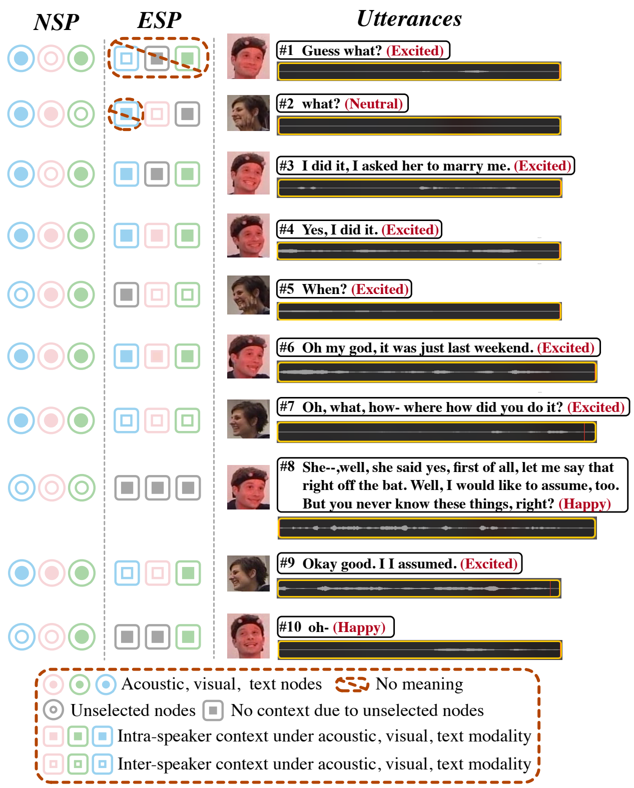

Figure 5: Visualization of selection policies during the test phase. The 1st utterance lacks context, and the 2nd utterance lacks intra-speaker context, rendering ESP meaningless.

the visual modality information. In the 5-th utterance, the woman empathizes with this man because of utterances 1, 3, and 4. Thus the ESP introduces a more inter-speaker context for visual modality data because they are the most effective for recognizing the emotion 'excited'. For the 8-th utterance, the model prioritizes the most original features (the output of the utterance-level encoder). This is likely because the utterance contains substantial, distinctive content that offers ample information for accurate emotion analysis.

Error Analysis

After conducting an error analysis per dataset, we discovered that the majority of errors can be attributed to the problem of class imbalance. Specifically, the 'fear' emotion had only 268 samples while 'neutral' had 4710, leading to an F1 score for 'fear' as low as 15.15 in the MELD dataset. Additionally, we focus on the issue of emotional shifts, where two consecutive utterances exhibit different emotions. Existing methods struggled with addressing emotional shifts (Shen et al. 2021b). Our AdaIGN faces similar challenges, as evident in Table 5, performing comparably worse on samples with emotional shifts compared to those without.

Methods	IEMOCAP		MELD	
	Acc	w-F1	Acc	w-F1
AdaIGN	70.49	70.74	67.62	66.79
w/ Emotion Shift	58.76	58.83	61.27	59.40
w/o Emotion Shift	75.74	75.93	76.91	77.63

Table 5: Analysis of AdaIGN on Emotional Shifts.

Related Work

Context Modeling in ERC Contextual information in ERC provides significant clues for emotion analysis, as evidenced by (Tu et al. 2023b). Unlike vanilla sentence-level emotion analysis, the ERC model requires modeling context- and speaker-sensitive dependencies (Tu et al. 2022a), including recurrent-based network (Majumder et al. 2019; Hu, Wei, and Huai 2021; Li et al. 2022), transformer-based network (Lian, Liu, and Tao 2021; Shen et al. 2021a; Jiang et al. 2022), and graph-based network (Ghosal et al. 2020b; Shen et al. 2021b; Tu et al. 2023a). However, modeling contexts among different modalities remains a significant challenge. Recent research efforts (Kang and Kong 2022; Hu et al. 2021; Lian et al. 2023) have explored the modeling of intra- and cross-modal interactions within a graph framework. Despite progress in ERC, these methods have not yet effectively tackled the essential need to balance inter- and intra-speaker contextual dependencies, striking a balance between empathy and emotional inertia.

Multimodal Fusion Multimodal fusion aims to combine information from different modalities through feature, decision, and model-level fusion strategies. Feature-level fusion involves concatenating multimodal features into a joint feature vector at the input level (Jiang et al. 2023), but it faces data sparseness due to high-dimensional feature sets (Wu, Lin, and Wei 2014). Decision-level fusion combines unimodal decision values through voting (Morvant, Habrard, and Ayache 2014), averaging (Shutova, Kiela, and Maillard 2016), or weighted sum (Glodek et al. 2011), but overlooks correlations between modalities. Model-level fusion, a middle ground, fuses intermediate representations of different modalities (Hsu et al. 2023). Recently, researchers have explored graph-based fusion to capture intra- and inter-modal interactive information (Hu et al. 2021, 2022; Yang et al. 2023). However, these graph structures predict emotions using future utterances, which is impractical in real-world scenarios. Furthermore, they face limitations in handling content with conflicting emotions across different modalities.

Conclusion

In this paper, we propose a novel adaptive IGN termed AdaIGN, that learns a selection pattern for nodes and edges in a multimodal heterogeneous graph. This selection process is guided by our proposed selection policies NSP and ESP. These policies prioritize selecting the text modality and intra-speaker context to meet the ERC task. Furthermore, we introduce GSP to integrate the utterance representation from the original IGN and NESP-enhanced IGN. To mitigate the computational complexity of policy learning, we leverage pseudo-labels to mask unnecessary utterance nodes for selection. Experimental results show that our method outperforms state-of-the-art methods on two well-known datasets.

Acknowledgments

We thank the anonymous reviewers for their valuable suggestions to improve the quality of this work. This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of Guangdong 2023A1515012922, Shenzhen Foundational Research Funding JCYJ20220818102415032, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005, The Major Key Project of PCL PCL2023A09.

References

Agarap, A. F. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; and Li, Y. 2020. Simple and deep graph convolutional networks. In *International conference on machine learning*, 1725–1735. PMLR.

Chudasama, V.; Kar, P.; Gudmalwar, A.; Shah, N.; Wasnik, P.; and Onoe, N. 2022. M2FNet: multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4652–4661.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020a. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2020b. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Glodek, M.; Tschechne, S.; Layher, G.; Schels, M.; Brosch, T.; Scherer, S.; Kächele, M.; Schmidt, M.; Neumann, H.; Palm, G.; et al. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*, 359–368. Springer.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2594–2604.

Hsu, P. Y.; Lei, H.-T.; Cheng, M.-S.; and Yuan, T. F. 2023. Applying Data Driven Approach to Cluster Components for Preventive Maintenance. In *Intelligent and Transformative Production in Pandemic Times: Proceedings of the 26th International Conference on Production Research*, 197–205. Springer.

Hu, D.; Hou, X.; Wei, L.; Jiang, L.; and Mo, Y. 2022. 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)*, 7037–7041. IEEE.

Hu, D.; Wei, L.; and Huai, X. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. 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)*, 5666–5675.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jiang, D.; Liu, H.; Wei, R.; and Tu, G. 2023. CSAT-FTCN: A Fuzzy-Oriented Model with Contextual Self-attention Network for Multimodal Emotion Recognition. *Cognitive Computation*, 1–10.

Jiang, D.; Wei, R.; Wen, J.; Tu, G.; and Cambria, E. 2022. AutoML-Emo: Automatic Knowledge Selection using Congruent Effect for Emotion Identification in Conversations. *IEEE Transactions on Affective Computing*.

Jiao, W.; Lyu, M.; and King, I. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8002–8009.

Kang, J.; and Kong, F. 2022. DialogueTRGAT: Temporal and Relational Graph Attention Network for Emotion Recognition in Conversations. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, 460–472. Springer.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, W.; Shao, W.; Ji, S.; and Cambria, E. 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, 467: 73–82.

Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lian, Z.; Liu, B.; and Tao, J. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 985–1000.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.

Mai, S.; Hu, H.; and Xing, S. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 481–492.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6818–6825.

Morvant, E.; Habrard, A.; and Ayache, S. 2014. Majority vote of diverse classifiers for late fusion. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, S+ SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*, 153–162. Springer.

Ong, D.; Su, J.; Chen, B.; Luu, A. T.; Narendranath, A.; Li, Y.; Sun, S.; Lin, Y.; and Wang, H. 2022. Is Discourse Role Important for Emotion Recognition in Conversation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11121–11129.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Poria, S.; Majumder, N.; Mihalcea, R.; and Hovy, E. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7: 100943–100953.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Saxena, P.; Huang, Y. J.; and Kurohashi, S. 2022. Static and dynamic speaker modeling based on graph neural network for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 247–253.

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

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13789–13797.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021b. Directed Acyclic Graph Network for Conversational Emotion Recognition. 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)*, 1551–1560.

Shutova, E.; Kiela, D.; and Maillard, J. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 160–170.

Tu, G.; Liang, B.; Jiang, D.; and Xu, R. 2022a. Sentiment-Emotion-and Context-guided Knowledge Selection Framework for Emotion Recognition in Conversations. *IEEE Transactions on Affective Computing*.

Tu, G.; Liang, B.; Lyu, X.; Gui, L.; and Xu, R. 2023a. Do topic and causal consistency affect emotion cognition? a graph interactive network for conversational emotion detection. In *In The 26th European Conference on Artificial Intelligence (ECAI'23)*, 2362–2369.

Tu, G.; Liang, B.; Mao, R.; Yang, M.; and Xu, R. 2023b. Context or Knowledge is Not Always Necessary: A Contrastive Learning Framework for Emotion Recognition in Conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 14054–14067.

Tu, G.; Wen, J.; Liu, H.; Chen, S.; Zheng, L.; and Jiang, D. 2022b. Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models. *Knowledge-Based Systems*, 235: 107598.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7216–7223.

Wu, C.-H.; Lin, J.-C.; and Wei, W.-L. 2014. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3: e12.

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

Yang, H.; Gao, X.; Wu, J.; Gan, T.; Ding, N.; Jiang, F.; and Nie, L. 2023. Self-adaptive Context and Modal-interaction Modeling For Multimodal Emotion Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6267–6281.

Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; and 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*, 1009–1021.

Zhang, D.; Zhang, W.; Li, S.; Zhu, Q.; and Zhou, G. 2020. Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 503–511.

Zhang, X.; and Li, Y. 2023. A Cross-Modality Context Fusion and Semantic Refinement Network for Emotion Recognition in Conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13099–13110.

Zheng, X.; Zhao, G.; Zhu, L.; and Qian, X. 2022. PERD: Personalized Emoji Recommendation with Dynamic User Preference. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1922–1926.

Zhu, L. Y.; Zhang, Z.; Wang, J.; Wang, H.; Wu, H.; and Yang, Z. 2022. Multi-Party Empathetic Dialogue Generation: A New Task for Dialog Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 298–307.