Summarize before Aggregate: A Global-to-local Heterogeneous Graph Inference Network for Conversational Emotion Recognition

Dongming Sheng[1,2] **Dong Wang**[1,2] **Ying Shen**[3] **Hai-Tao Zheng**[1,2,*] **Haozhuang Liu**[1,2]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Shenzhen International Graduate School, Tsinghua University
[3]School of Intelligent Systems Engineering, Sun Yat-Sen University
{cdm19, wangd18, lhz19}@mails.tsinghua.edu.cn
sheny76@mail.sysu.edu.cn, zheng.haitao@sz.tsinghua.edu.cn

Abstract

Conversational Emotion Recognition (CER) is a crucial task in Natural Language Processing (NLP) with wide applications. Prior works in CER generally focus on modeling emotion influences solely with utterance-level features, with little attention paid on phrase-level semantic connection between utterances. Phrases carry sentiments when they are referred to emotional events under certain topics, providing a global semantic connection between utterances throughout the entire conversation. In this work, we propose a two-stage Summarization and Aggregation Graph Inference Network (SumAggGIN), which seamlessly integrates inference for topic-related emotional phrases and local dependency reasoning over neighbouring utterances in a global-to-local fashion. Topic-related emotional phrases, which constitutes the global topic-related emotional connections, are recognized by our proposed heterogeneous Summarization Graph. Local dependencies, which captures short-term emotional effects between neighbouring utterances, are further injected via an Aggregation Graph to distinguish the subtle differences between utterances containing emotional phrases. The two steps of graph inference are tightly-coupled for a comprehensively understanding of emotional fluctuation. Experimental results on three CER benchmark datasets verify the effectiveness of our proposed model, which outperforms the state-of-the-art approaches.

1 Introduction

Conversational Emotion Recognition (CER) has attracted increasing interests for its promising applications in intelligent interactive systems with diverse functionalities, including medical-care systems and online recommendation systems (Zhang et al., 2014; Gkotsis et al., 2016; Chen et al., 2018; Shen et al., 2020). As shown in Figure 1, conversations in CER datasets are segmented into multiple utterances based on breaths or pauses of the speaker, and each utterance is associated with an emotion label.

Existing works with deep learning approaches generally capture emotional features solely from an utterance-level perspective by modeling interactions between utterances via Recurrent Neural Network (RNN) structures (Hazarika et al., 2018b; Hazarika et al., 2018a; Majumder et al., 2019) or graph structures (Ghosal et al., 2019). However, phrase-level semantic connection between utterances is still under-explored, creating a substantial barrier for comprehensively understanding of the source of emotional fluctuation. An example of a topic-related emotional phrase is shown in Figure 1, where the noun phrase "fifty dollars" is not an emotional expression at first, but under the topic of compensation, it has been associated with anger and frustration by referring to the unsatisfactory compensation. Therefore, it is crucial to recognize phrase-level emotional patterns globally across different utterances. Furthermore, topic-related emotional phrases generally scatter in different utterances throughout the conversation, and the emotions they convey depend on local context. To accurately distinguish emotions behind these "fifty dollars" phrases, the model is required to be comprehensively aware of both the topic about compensation and the subtle differences in attitude from the responses of the male speaker. By seamlessly

* indicates corresponding author.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4153–4163
Barcelona, Spain (Online), December 8-13, 2020

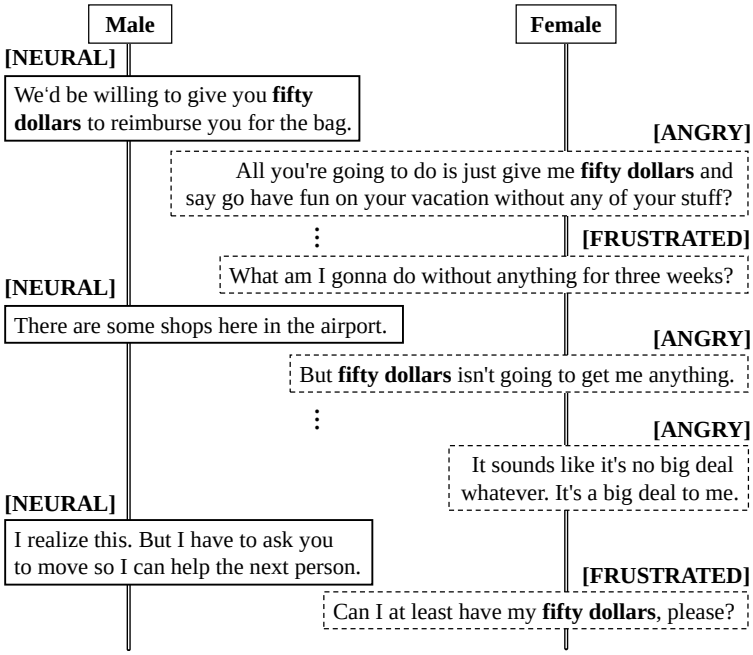


Figure 1: A sample conversation from the IEMOCAP dataset for demonstrating topic-related emotional phrases. The female speaker is complaining that her bag is lost during the flight and the amount of compensation provided by the airline is unsatisfactory. Originally, the noun phrase "fifty dollars" does not contain emotions, however in this context, it has become a topic-related emotional phrase repeatedly used by the female speaker to express her disappointment towards the airline service.

combining global features from topic-related emotional phrases and local features on local dependencies from neighbouring utterances, the emotions associated with each utterance are fully excavated in a global-to-local fashion.

Graphs with multiple types of nodes are called the heterogeneous graph and have been applied to various NLP tasks for aligning information from different domains (Yao et al., 2019; Tu et al., 2019; Yu et al., 2019; Wang et al., 2020), nevertheless it is still a relatively new territory for the CER task. A tricky issue for exploring phrase-level semantic connections is the recognition of topic-related emotional phrases, which requires a meticulous reasoning over different utterances to judge whether a phrase is referred to an emotional event under the topic. Intuitively, the heterogeneous graph can be applied to align among phrase-level and utterance-level features, providing a thorough reasoning of topic-related emotional phrases.

In this paper, we propose a two-stage Summarization and Aggregation Graph Inference Network (SumAggGIN) for ERC, which seamlessly integrates inference for topic-related emotional phrases and local dependency reasoning over neighbouring utterances in a global-to-local fashion. A heterogeneous Summarization Graph which consists of two types of nodes (i.e. utterance nodes and phrase nodes) is proposed to infer topic-related emotional phrases. This Summarization Graph enables information propagation between utterances and phrases via overlapping phrases and utterance phrase relations, thereby enhances utterance representation with summarized phrase-level semantic connections. Subsequently, an Aggregation Graph is constructed to further inject speaker-related local dependencies into the topic-aware utterance representation for capturing short-term emotional effects between neighbouring utterances. The Summarization Graph and the Aggregation Graph are tightly-coupled to bridge global topic-related phrase patterns and local speaker-related utterance-level features for a comprehensively understanding of emotional fluctuation.

The main contributions of our work are highlighted as follows:

(1) We propose a novel two-stage Summarization and Aggregation Graph Inference Network (SumAggGIN), which comprehensively captures emotional influences in a global-to-local fashion.

(2) To the best of our knowledge, we are the first to construct a heterogeneous Summarization Graph for inferencing global emotional interactions across utterances based on phrases.

(3) Extensive experiments on three publicly available CER datasets demonstrate that our model attains a substantial improvement and achieves state-of-the-art performance.

2 Related Work

2.1 Emotion Recognition in Conversation

Emotion recognition, i.e. sentiment analysis, is a fundamental task in NLP with wide applications (Xia et al., 2011; Liu, 2012). For medical-care systems, conversational emotion recognition can be incorporated to accurately monitor patients' emotional fluctuation and detect potential mental health issues (i.e. depression and suicidal intention) in time (Gkotsis et al., 2016; Korkontzelos et al., 2016; Chen et al., 2018). Moreover, emotion recognition can bring benefits to online recommendation on social media platforms by modeling users' short-term preferences (Zhang et al., 2014; Shen et al., 2020). Recently, an increasing number of models have been proposed to solve CER using various structures. Poria et al. (2017) propose to capture context-aware utterance representation via a Bi-directional LSTM network. Additionally, they adopt an attention mechanism to re-weight the outputs for a more informative output. Memory networks proposed by Sukhbaatar et al. (2015) are performed by some prior works to capture speaker-related historical information. In CMN, proposed by Hazarika et al. (2018b), two distinct memory cells are employed to model dialogue history of the two speakers. Hazarika et al. (2018a) further extend CMN by utilizing another memory cell for modeling global emotional influences across speakers. For distinguishing participants in a multiparty conversation, Majumder et al. (2019) propose DialogueRNN with three GRUs (Tang et al., 2015) tracking individual participant states, global context and emotional states respectively. Jiao et al. (2020) proposes an attention gated hierarchical memory network for real-time emotion recognition without future context. The state-of-the-art DialogueGCN (Ghosal et al., 2019) leverages speaker and temporal dependency from an utterance level by performing graph convolution on a homogeneous graph with each utterance as a node. Nevertheless, none of these methods explicitly models global semantic interactions based on topic-related emotional phrases.

2.2 Applications of Heterogeneous Graph for NLP

Graphs comprised of multiple types of nodes are called the heterogeneous graph. These graphs are constructed to simulate the real-world scenario with multiple granularity levels of information included and have been widely adopted in solving NLP tasks. In text classification, Yao et al. (2019) propose to build a heterogeneous text graph for a corpus based on phrase co-occurrence and term frequency-inverse document frequency (TF-IDF) weights. Tu et al. (2019) design a heterogeneous document-entity graph with candidates, documents and entities in specific document context, facilitating accurate reasoning for multi-hop reading comprehension. For the task of visual commonsense reasoning, Yu et al. (2019) construct a vision-to-answer heterogeneous graph and a question-to-answer heterogeneous graph to bridge the proper semantic alignment between vision and linguistic domains. Wang et al. (2020) propose to capture relations between sentences by constructing a heterogeneous graph network consisting of semantic units of different granularity for extractive document summarization. These prior studies inspire us to align among phrase-level and utterance-level features for recognizing topic-related emotional phrases via a heterogeneous graph.

3 Methodology

Before diving into the details of our proposed model, we begin by introducing the basic mathematical notations and terminologies for the task of CER. The goal of CER is to infer the emotion label (happy, sad, neutral, angry, excited, and frustrated) for each utterance in a conversation. Given a CER dataset D, an example of the dataset is denoted as $\{U_i, s_i, y_i\}_{i=1}^{n}$, where $U = \{U_1, U_2, ..., U_n\}$ represents the n utterances of the conversation with each utterance U_i containing l_i words. Assuming there are M participants, the speaker corresponds to the i-th utterance U_i is represented as $s_i \in \{0, ..., M-1\}$. $y_i \in \{0, ..., N-1\}$ indicates the emotion label for utterance U_i.

An overview of our proposed SumAggGIN model is shown in Figure 2, which consists of an Encoding module, a Summarization Graph, an Aggregation Graph and a Classification module.

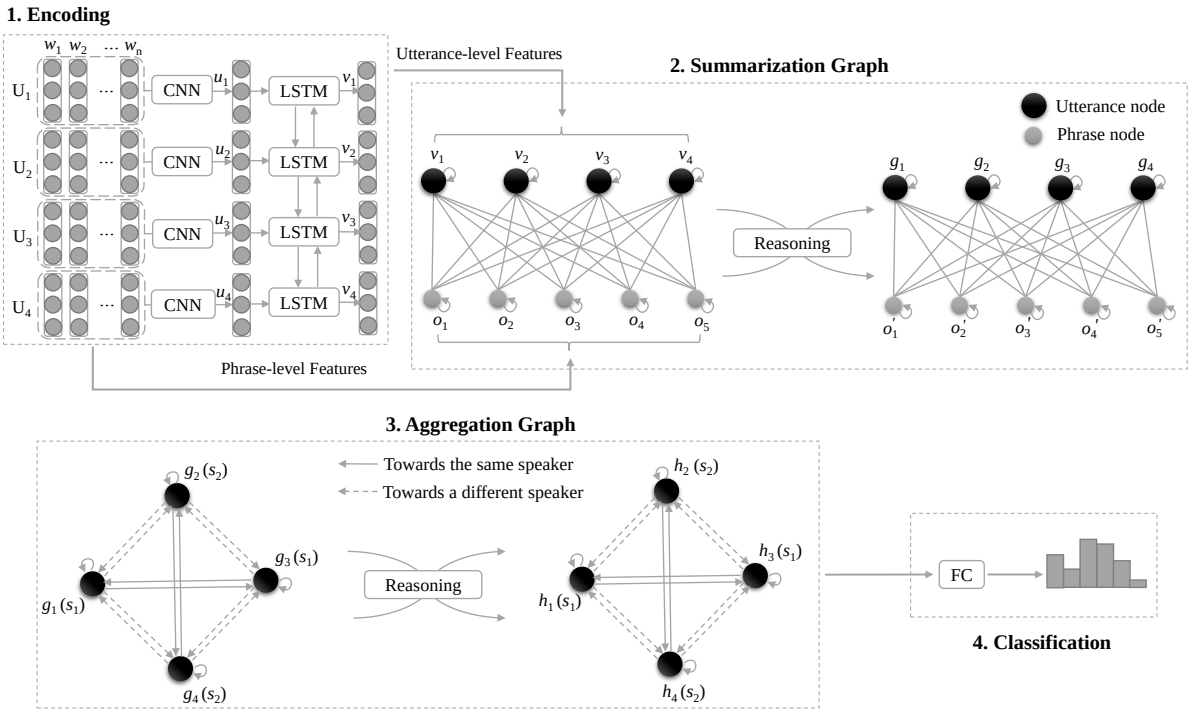


Figure 2: The architecture of our SumAggGIN model. For simplicity, we demonstrate our graph structures via a dyadic conversation with 4 utterances, and each utterance includes the same 5 overlapping phrases. s_1 and s_2 represent the two distinct speakers in the conversation. $\oplus$ denotes the concatenation operation.

3.1 Encoding

For words in utterances, we convert them into 300 dimensional pretrained 840B GloVe word embeddings (Pennington et al., 2014). A TextCNN (Kim, 2014) is performed to capture n-grams information from each utterance U_i. We use convolution filters of sizes 3, 4 and 5 with each filter containing 50 feature maps. The outputs of convolutions are further processed by max-pooling and ReLU activation (Nair and Hinton, 2010). We concatenate these activation results and feed them to a 150 dimensional fully connected layer, whose outputs are denoted as $\{u_i\}_{i=1}^{n}$. Subsequently, based on local utterance features from TextCNN, we apply a bidirectional LSTM (BiLSTM) to capture sequential contextual information. We denote $v_i \in \mathcal{R}^d$ as the sequential context-aware utterance representation for the i-th utterance and d is the hidden size of BiLSTM.

3.2 Summarization Graph

In this section, a heterogeneous Summarization Graph is constructed to recognize topic-related emotional phrases so as to explicitly model topic-related emotional interactions throughout the entire conversation. Phrases are extracted from utterances by TextRank (Mihalcea and Tarau, 2004). Through exchange of information between utterances and phrases on the Summarization Graph, utterance representation can be enhanced with summarized phrase-level semantic connections from a global perspective.

3.2.1 Summarization Graph Construction

We denote our Summarization Graph as $\mathcal{G}_{sum} = (\mathcal{V}, \mathcal{E}_{sum})$, where $\mathcal{V} = \mathcal{V}_o \cup \mathcal{V}_u$ represents a node set composing of phrase nodes and utterance nodes and $\mathcal{E}_{sum}$ stands for edges between nodes. $\mathcal{V}_o = \{o_1, ..., o_m\}$ and $\mathcal{V}_u = \{U_1, ..., U_n\}$ represent the m key phrases of the utterances and n utterances in the conversation, respectively. $e_{ij} \geq 0$ denotes the weight of the edge between the i-th phrase and the j-th utterance. In particular, $e_{ij} = 0$ indicates that the i-th phrase does not appear in the j-th utterance. Self-loops are included to ensure that the original features of each node can be preserved in the course of message propagation. For phrase nodes in $\mathcal{V}_o$, their feature vectors are initialized by averaging pretrained GloVe embeddings of the constituting words. As for utterance nodes $U_i \in \mathcal{V}_u$, they

are initialized with the corresponding sequential context-aware utterance representation v_i obtained from BiLSTM. Therefore, the feature matrices of phrase and utterance nodes are denoted as $X_o \in \mathcal{R}^{m \times d_w}$ and $X_u \in \mathcal{R}^{n \times 2d}$ respectively, where d_w is the dimension of the word embedding. In our experiments, we have $d_w = 2d$.

To infuse relation importance about the edge between a phrase node and an utterance node, we use TF-IDF weights of the phrase in the utterance as suggested by Yao et al. (2019). Term frequency is the number of times phrase o_i occurs in an utterance U_j, while inverse document frequency represents the logarithmically scaled inverse fraction of the number of utterances containing the phrase o_i.

3.2.2 Message Propagation

We apply a variant of Graph Attention Network (GAT) (Veličković et al., 2018) to propagate information among nodes in the Summarization Graph. The hidden states of input nodes are denoted as $g_i \in \mathcal{R}^{2d \times 1}$, $i \in \{1, ..., (m+n)\}$. A Multi-Layer Perceptron (MLP) is applied to compute attention coefficients between a node i and its neighbor j ($j \in \mathcal{N}_i$) at layer t:

$$p_{ij}^{(t)} = W_a^{(t)}(\mathrm{ReLU}(W_b^{(t)}[g_i^{(t-1)} \oplus g_j^{(t-1)} \oplus e_{ij}])), \tag{1}$$

where W_a^t and W_b^t are trainable parameters at the t-th layer, $\oplus$ denotes the concatenation operation, and $\mathcal{N}_i$ denotes the set of neighbors of node i. Subsequently, the coefficients are normalized using the softmax function:

$$\alpha_{ij}^{(t)} = \mathrm{softmax}_j(p_{ij}^{(t)}) = \frac{\exp(p_{ij}^{(t)})}{\sum_{k \in \mathcal{N}_i} \exp(p_{ik}^{(t)})}. \tag{2}$$

Finally, we utilize the normalized attention coefficients to compute a linear combination of the neighbouring features. The updated feature vector for node i at the t-th layer is formulated as:

$$g_i^{(t)} = \sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(t)} g_j^{(t-1)}. \tag{3}$$

Although utterance nodes are not directly connected, stacking 2 layers of GAT enables the indirect exchange of information between pairs of utterances through co-appearing phrases. Inspired by Transformer (Vaswani et al., 2017), we further apply a position-wise feed-forward (FFN) layer after each GAT layer. The summarized representation for the i-th utterance after propagation is denoted as $g_i = g_i^{(2)}$.

3.3 Aggregation Graph

The utterance representation obtained from Summarization Graph mainly captures global topic-related emotional interactions throughout the whole conversation. To further explore short-term emotional effects between neighbouring utterances, we construct an Aggregation Graph for modeling speaker-related context dependencies from a local perspective.

3.3.1 Aggregation Graph Construction

An Aggregation Graph can be denoted as $\mathcal{G}_{agg} = (\mathcal{V}_u, \mathcal{E}_{agg}, \mathcal{R})$, where $\mathcal{V}_u$ represents the node set containing utterance nodes solely, $\mathcal{E}_{agg}$ stands for edges between nodes, and $\mathcal{R}$ denotes the type of the edges. Each utterance node $U_i \in \mathcal{V}_u$ is initialized with the corresponding summarized utterance representation g_i obtained from the Summarization Graph.

To explicitly model speaker dependencies between utterances, we divide edges in $\mathcal{E}_{agg}$ into 2 categories, i.e. edges towards the same speaker and edges towards a different speaker. To capture emotional patterns only from neighbouring utterances, we construct the edges by keeping a context window size of W. As a result, each utterance node U_i only links to W utterances in the past ($U_{i-W}, U_{i-W+1}, ..., U_{i-1}$) and W utterances in the future ($U_{i+1}, U_{i+2}, ..., U_{i+W}$). The edge weights z_{ij} are obtained from the cosine similarity between the feature vectors h_i and h_j of the two utterance nodes U_i and U_j:

$$z_{ij} = \frac{h_i^T h_j}{\|h_i\|_2 \cdot \|h_j\|_2}. \tag{4}$$

To ensure that for each utterance node, the incoming set of edges receives a total weight contribution of 1. The edge weights are further normalized by the softmax function:

$$\beta_{ij} = \mathrm{softmax}_j(z_{ij}) = \frac{\exp(z_{ij})}{\sum_{k=i-W}^{i+W} \exp(z_{ik})}. \tag{5}$$

3.3.2 Message Propagation

To pass messages between neighbouring utterance nodes, graph convolution is performed on the basis of the Aggregation Graph. A message passing strategy concerning different types of edges from (Schlichtkrull et al., 2018) is adopted:

$$h_i^{(t)} = \mathrm{ReLU}(\sum_{r\in\mathbb{R}} \sum_{j\in\mathcal{N}_i^r} \frac{\beta_{i,j}}{c_{i,r}} W_r^{(t)} h_j^{(t-1)} + \beta_{i,i} W_0^{(t)} h_i^{(t-1)}), \tag{6}$$

where $\mathcal{N}_i^r$ represents the set of neighbors of node i under edge type $r \in \mathcal{R}$, $\beta_{i,j}$ and $\beta_{i,i}$ are the normalized edge weights. The normalization constant $c_{i,r}$ is set to be $|N_i^r|$, the number of neighbouring nodes of node i under edge type r. $W_r^{(t)}$ and $W_0^{(t)}$ are trainable parameters. A two-step graph convolution process is applied and the final aggregated representation for the i-th utterance is denoted as $h_i = h_i^{(2)}$.

3.4 Classification

The aggregated utterance representation h_i from the Aggregation Graph is fed into a fully-connected network to obtain the final prediction results for emotion classification:

$$\tilde{y_i} = \mathrm{softmax}(W_c h_i + b_c), \tag{7}$$

where W_c and b_c are trainable parameters of the classifier.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the performance of our SumAggGIN model on three benchmark datasets for the CER task, i.e. IEMOCAP (Busso et al., 2008), AVEC (Schuller et al., 2012), and MELD (Poria et al., 2019), which are also used by several prior works, including DialogueRNN(Majumder et al., 2019) and DialogueGCN(Ghosal et al., 2019). These multimodal datasets originally contain textual, visual and acoustic information about the utterances in each conversation. However, for the task of CER, we only focus on textual modality to conduct our experiments.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture database (Busso et al., 2008) is a multimodal dataset consisting of videos of dyadic sessions. Each video contains a single conversation, which is segmented into multiple utterances. Each utterance is annotated with one of six emotion labels, i.e. happy, sad, neural, angry, excited and frustrated. The dataset is officially split into a training set consisting of 120 conversations with 5,810 utterances and a test set containing 31 conversations with 1,623 utterances. We randomly select 10% of training conversations as evaluation split for selecting hyperparameters. We use weighted average of accuracy and f1-score for evaluating the overall performance.

AVEC: The continuous Audio/Visual Emotion Challenge (Schuller et al., 2012) dataset is a modification of SEMAINE database (McKeown et al., 2012), which records interactions between human users and virtual agents. Different from IEMOCAP, utterances in the AVEC dataset are labeled every 0.2 second with four real valued attributes: valence ($[-1, 1]$), arousal ($[-1, 1]$), expectancy ($[-1, 1]$), and power ($[0, \infty)$). Following Majumder et al. (2019), the attributes are averaged over the span of an utterance to obtain utterance-level annotations. The standard split of AVEC dataset contains 63/32 conversations (4,368/1,430 utterances) for training and testing. 10% of the training conversations are randomly selected for evaluation. Mean Absolute Error (MAE) for the regression task is applied for model evaluation.

MELD: the Multimodal EmotionLines Dataset (Poria et al., 2019) is a multimodal and multiparty dataset extended from the EmotionLines dataset (Hsu et al., 2018). It contains more than 1,400 conversations and 13,000 utterances from the Friends TV series. The utterances are annotated with one of

the seven emotion labels (anger, disgust, sadness, joy, surprise, fear and neutral). Following Poria et al. (2019), we split the dataset into 1,039/114/280 conversations (9,989/1,109/2,610 utterances) for training, evaluation and testing. We use weighted average of f1-score to evaluate our model.

4.2 Baselines

We choose the following baseline models to evaluate the performance of our SumAggGIN model:

TextCNN (Kim, 2014) is the baseline convolutional neural network based model. It is a sub-component of our *Encoding* module (Section 3.1), which captures n-grams information from each utterance independently. However, it is not capable of capturing context-aware information from the surrounding utterances.

bc-LSTM+Att (Poria et al., 2017) utilizes Bi-directional LSTM network for capturing context-aware utterance representations. These representations are speaker agnostic for the reason that they are encoded irrespective of their speakers. Additionally, an attention mechanism is adopted to re-weight features and provide a more informative output.

CMN (Hazarika et al., 2018b) exploits two distinct GRUs for modeling the historical utterances of two speakers, and generates speaker-aware utterance representation.

ICON (Hazarika et al., 2018a) extends CMN by incorporating inter-speaker emotional influences into the original output of CMN, which only takes self-speaker historical information into consideration.

DialogueRNN (Majumder et al., 2019) is a recurrent network with three GRUs capturing speaker states, global context and emotional states respectively. It can be applied on multiparty datasets for distinguishing different participants in a conversation interactively.

AGHMN (Jiao et al., 2020) proposes an attention gated hierarchical memory network which keeps track of the individual party states throughout the conversation for real-time emotion recognition without future context.

DialogueGCN (Ghosal et al., 2019) is the state-of-the-art model for the CER task. It employs graph convolutional network for capturing self- and inter-speaker influences between utterances.

4.3 Implementation Details

All experiments are carried out using a single NVIDIA Tesla M40 24GB card. The batch size is set to 16. We adopt Adam (Kingma and Ba, 2015) as the optimizer with an initial learning rate of 3e-4 and L2 weight decay of 1e-5. We use an early stopping strategy on f1-score (IEMOCAP and MELD) and MAE (AVEC) of the validation set, with a patience of 20 epochs. Hyper-parameters are tuned on the validation set. The hidden size of the Summarization Graph and the Aggregation Graph is set to 100 for IEMOCAP/AVEC and 200 for MELD. We set the context window size W in the Aggregation Graph to 10 for IEMOCAP/AVEC and 5 for MELD. Dropout rate with 0.5 is applied to avoid over-fitting.

4.4 Overall Performance

As shown in Table 1, we compare the performance of our proposed SumAggGIN model with the state-of-the-art DialogueGCN and other strong baseline models on textual modality using three CER datasets. The results reveal that our SumAggGIN model attains a substantial improvement compared to the state-of-the-art on all three datasets. Our SumAggGIN model pushes up state-of-the-art results by 1.51% and 2.43% in terms of weighted average accuracy and f1-score on IEMOCAP dataset, respectively. On AVEC dataset, our model outperforms on all four attributes compared to the state-of-the-art DialogueGCN. As for the MELD dataset, our SumAggGIN model attains the state-of-the-art average f1-score of 58.45%.

4.5 Ablation Study

In order to examine the effectiveness of our proposed SumAggGIN model, we perform ablation test with respect to the two graph inference modules on IEMOCAP dataset, and the results are listed in Table 2. The two modules are removed one at a time to examine its contribution on the entire model.

w/o Aggregation Graph In this setting, the Aggregation Graph is removed from the full model. From the result, we can observe a 2.87% and 2.50% drop in terms of f1-score and accuracy, indicating the

Models	IEMOCAP		AVEC				MELD
	Average		*Valence*	*Arousal*	*Expectancy*	*Power*	*Average*
	Accuracy	F1	MAE				F1
CNN	48.92	48.18	0.545	0.542	0.605	8.71	55.02
bc-LSTM+Att	56.32	56.19	0.189	0.213	0.190	8.67	56.70
CMN	56.56	56.13	0.192	0.213	0.195	8.74	-
ICON	59.09	58.54	0.180	0.190	0.180	8.45	-
DialogueRNN	63.40	62.75	0.168	0.165	0.175	7.90	57.03
AGHMN	63.50	63.50	-	-	-	-	58.10
DialogueGCN	65.25	64.18	0.157	0.161	0.168	7.68	58.10
SumAggGIN	**66.76**	**66.61**	**0.145**	**0.156**	**0.160**	**7.55**	**58.45**

Table 1: Comparison of our model with baselines on IEMOCAP, AVEC and MELD datasets. For IEMOCAP, *Average* represents the weighted average of all six labels (i.e. happy, sad, neural, angry, excited and frustrated). As for MELD, *Average* represents the weighted average of all seven labels (i.e. anger, disgust, sadness, joy, surprise, fear and neutral). '-' denotes that the results are not reported in the original paper.

importance of the Aggregation Graph for capturing speaker-related local dependencies on short-term emotional effects between neighbouring utterances.

w/o Summarization Graph If we only consider local dependencies from the Aggregation Graph and remove global connections from the Summarization Graph, f1-score and accuracy degrade by 4.03% and 3.98%, respectively. This proves that the global connections via topic-related emotional phrases are of vital importance for emotional recognition.

w/o inference graphs Here, we remove both the Aggregation Graph and the Summarization Graph, and use the sequential context-aware utterance representation output by BiLSTM for emotion recognition. The experimental results show a significant drop of 10.18% and 8.60% in terms of f1-score and accuracy, which demonstrates the efficacy of modeling emotional influences via a two-stage graph inference network in a global-to-local fashion.

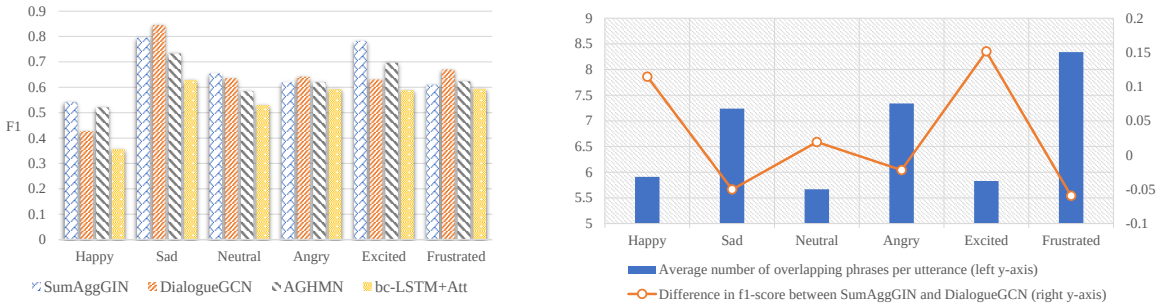
Model	F1		Accuracy	
	Test	Δ	Test	Δ
Full model	66.61	-	66.76	-
w/o Aggregation Graph	63.74	2.87	64.26	2.50
w/o Summarization Graph	62.58	4.03	62.78	3.98
w/o inference graphs	56.43	10.18	58.16	8.60

Table 2: Ablation results on the test set of the IEMOCAP dataset.

4.6 Visualization

Figure 3(a) shows the comparison of f1-score between our SumAggGIN model and other three baselines on six emotion labels of the IEMOCAP dataset. The bc-LSTM+Att model, which only takes sequentially encoded context information into consideration, attains an inferior performance on all six labels compared to the other models, which proves the importance of modeling speaker-related local dependencies between utterances in the conversation.

Our SumAggGIN model surpasses AGHMN on four emotion labels (i.e. 'happy', 'sad', 'neutral' and 'excited') and achieves competitive results on 'angry' and 'frustrated'. We believe our improvement comes from the graph structures we designed for modeling utterance connections. For the conversations with over 100 utterances in the IEMOCAP dataset, the RNN structure in AGHMN for dependency modeling suffers from long-term information propagation issues resulted from gradient vanishing problem. On the contrary, our Summarization Graph and Aggregation Graph directly aggregate information from neighbouring utterances, which greatly shortens the path for information propagation and achieves a better result in modeling emotions.



(a) (b)

Figure 3: (a) The f1-score of our SumAggGIN and other three baselines on six emotion labels of the IEMOCAP dataset. (b) Plots of the average number of overlapping phrases per utterance in Summarization Graph (left y-axis), and the difference in f1-score between our SumAggGIN and DialogueGCN (right y-axis) on the test set of the IEMOCAP dataset.

The comparison between our SumAggGIN and another graph-based method DialogueGCN shows that our model outperforms DialogueGCN on the positive emotion labels (i.e. 'happy' and 'excited') by a large margin, and attains comparable results on the other four labels. We surmise that the improvement comes from our proposed heterogeneous Summarization Graph, which effectively explores global semantic connections provided by topic-related emotional phrases.

We further investigate the effect of our Summarization Graph by examining the average number of overlapping phrases per utterance for each emotion label, as shown in Figure 3(b). For the utterances labeled with negative emotions (i.e. 'sad', 'angry' and 'frustrated'), the average number of overlapping phrases per utterance is much higher than that of the utterances with positive emotions. We argue that global semantic connections are beneficial to capture emotional patterns, but excessive connections from overlapping phrases result in bringing noise signals into the Summarization Graph and causing performance degradation.

5 Conclusion

In this work, we present a two-stage Summarization and Aggregation Graph Inference Network (SumAggGIN) for the CER task. Combining global topic-related emotional interactions from the heterogeneous Summarization Graph and short-term emotional effects from the Aggregation Graph, our model is capable of comprehensively capturing emotional influences in a global-to-local fashion. Experimental results show that our proposed SumAggGIN outperforms the state-of-the-art approaches on three CER benchmark datasets. In the future, we plan on applying our two-stage graph inference framework to other multi-turn dialogue tasks.

Acknowledgements

We thank the reviewers for their helpful comments. This research is supported by National Natural Science Foundation of China (Grant No. 61773229), Shenzhen Giiso Information Technology Co. Ltd., the Basic Research Fund of Shenzhen City (Grand No. JCYJ20190813165003837), Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202032) and Overseas Cooperation Research Fund of Graduate School at Shenzhen, Tsinghua University (Grant No. HW2018002).

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. 2016. Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 95–105.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. 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)*, pages 2122–2132, New Orleans, Louisiana, June. Association for Computational Linguistics.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wenxiang Jiao, Michael R Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H Gonzalez. 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of biomedical informatics*, 62:148–158.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, Jul.

G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, Jan.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. 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)*, pages 873–883, Vancouver, Canada, July. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July. Association for Computational Linguistics.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012: The Continuous Audio/Visual Emotion Challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 449–456, New York, NY, USA. ACM.

Tiancheng Shen, Jia Jia, Yan Li, Yihui Ma, Yaohua Bu, Hanjie Wang, Bo Chen, Tat-Seng Chua, and Wendy Hall. 2020. Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 206–213.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end Memory Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2440–2448, Cambridge, MA, USA. MIT Press.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuan-Jing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*.

Rui Xia, Chengqing Zong, and Shoushan Li. 2011. Ensemble of feature sets and classification algorithms for sentiment classification. *Information sciences*, 181(6):1138–1152.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao. 2019. Heterogeneous graph learning for visual commonsense reasoning. In *Advances in Neural Information Processing Systems*, pages 2769–2779.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92.