DialogueTRM: Exploring Multi-Modal Emotion Dynamics in Conversations

Yuzhao Mao†, Guang Liu†, Xiaojie Wang‡, Weiguo Gao† and Xuan Li†
†Ping An Life Insurance of China
‡Beijing University of Posts and Telecommunications
maoyuzhao@126.com

Abstract

Emotion dynamics formulates principles explaining the emotional fluctuation during conversations. Recent studies explore the emotion dynamics from the self and inter-personal dependencies, however, ignoring the temporal and spatial dependencies in the situation of multi-modal conversations. To address the issue, we extend the concept of emotion dynamics to multi-modal settings and propose a Dialogue Transformer for simultaneously modeling the intra-modal and inter-modal emotion dynamics. Specifically, the intra-modal emotion dynamics is to not only capture the temporal dependency but also satisfy the context preference in every single modality. The inter-modal emotional dynamics aims at handling multi-grained spatial dependency across all modalities. Our models outperform the state-of-the-art with a margin of 4%-16% for most of the metrics on three benchmark datasets.

1 Introduction

With the development of conversational agents, e.g., Apple Siri, Google Assistant, Microsoft Cortana, etc., there emerges pressing needs for Emotion Recognition in Conversations (ERC). Different from conventional emotion recognition (Tzirakis et al., 2017) that treats emotions as stable traits, ERC involves emotion dynamics (Hazarika et al., 2018b) in conversations. Existing studies propose methods for modeling vanilla emotion dynamics by capturing self and inter-personal dependencies (Morris and Keltner, 2000). The two dependencies are methodologically considered as modeling individual and conversational context using variants of context-dependent models (Cho et al., 2014; Hochreiter and Schmidhuber, 1997). Bidirectional contextual LSTM (Poria et al., 2017) is a straightforward approach but suffers from inadequacy of long-range summarization. To overcome the shortcoming, attention mechanism (Majumder et al., 2019; Jiao et al., 2019) and memory network (Hazarika et al., 2018b,a) are introduced. Besides, variants of hierarchical Recurrent Neural Networks (Majumder et al., 2019; Hazarika et al., 2018a; Ghosal et al., 2019) are proposed to model self and inter-personal dependencies simultaneously. For better context modeling, pre-training techniques are employed to ERC (Ghosal et al., 2020).

Despite the progress of existing studies in modeling vanilla emotion dynamics in conversations, the temporal and spatial dependencies within multiple modalities are ignored. Thus, we extend the concept of emotion dynamics to multi-modal settings, which takes account of the intra-modal and inter-modal emotion dynamics, or multi-modal emotion dynamics for short. The intra-modal emotion dynamics is an emotional influence that one modality received from itself during a conversation. It needs temporal modeling in each modality. The inter-modal emotion dynamics is another emotional influence that one modality received from the other modalities at each conversation turn. It requires spatial modeling across all modalities. The interplays between intra-modal and inter-modal emotion dynamics produce final emotional predictions.

For intra-modal emotion dynamics, the temporal dependency of one modality can be captured through modeling the self and inter-personal dependencies as it is done in vanilla emotion dynamics. However, multi-modal expressions exhibit different dependence on context information. Such charac-
teristic is ignored by existing studies on ERC. Intuitively, spoken words are highly semantic that require inferences from the context to understand the emotions (Poria et al., 2017), while facial attributes or tones of voice are relatively concrete in which emotions are instantly burst in a short time, i.e., within an utterance period (Datcu and Rothkrantz, 2014). The phenomenon is illustrated in Figure 1. Here, the 7th-turn utterance “My sandwich” does not exhibit any anger unless looking back to infer that A is angry because B ate his sandwich. On the contrary, the anger is directly shown up in the frown faces or loud intonations at the 7th utterance period. Thus, the modeling of intra-modal emotion dynamics should satisfy the context preferences of different modalities.

For inter-modal emotion dynamics, the spatial dependency can be captured by interactive weighting across multi-modal features. Existing studies on ERC (Majumder et al., 2019; Hazarika et al., 2018a,b) use concatenation to learn the linear weights, which lacks the interactions between modalities. Many studies on multi-modal learning (Gu et al., 2019; Mao et al., 2018; Tsai et al., 2019a) apply interactive weighting to fuse information from multiple modalities. However, most of them consider only one granularity of feature interaction. We argue that interactive weighting should consider both prototype and representation dependencies. The prototype dependency relates to position-wise neuron-grained feature interactions that allocate different weights to neurons in a vector. The representation dependency handles vector-grained feature interactions that allocate a single weight to all neurons in a vector. The modeling of inter-modal emotion dynamics should consider the two granularities of dependencies.

In this paper, we propose a DialogueTRansformer (DialogueTRM) that models the intra-modal and inter-modal emotion dynamics simultaneously. For intra-modal emotion dynamics, we facilitate Transformers for temporal modeling that satisfies the context preferences of different modalities. For inter-modal emotion dynamics, we design a Multi-Grained Interactive Fusion (MGIF) to deal with the prototype and representation dependencies across modalities. Finally, by incorporating the intra-modal and inter-modal emotion dynamics, our DialogueTRM achieves more accurate emotional predictions than State-Of-The-Art (SOTA).

We highlight our contribution as follows:

- We propose a novel understanding of emotion dynamics in multi-modal settings, indicating
  - The intra-modal emotion dynamics, independently modeled under preferred context settings for each modality.
  - The inter-modal emotion dynamics, modeled in a fashion of multi-grained interactive fusion across modalities.
- Our DialogueTRM achieves SOTA performance on three ERC benchmark datasets, and we conduct a series of experiments to verify the effectiveness of each module in our model.

2 Related Work

Emotions are hidden mental states associated with thoughts and feelings (Poria et al., 2019b). Without physiological signals, they are only perceivable through human behaviors like spoken words, tones of voice, and facial attributes.

Emotion recognition is an interdisciplinary field that spans psychology, cognitive science, machine learning, natural language processing, and others (Picard, 2010). It involves handling multi-modal data. Early studies on emotion recognition are usually single-modal oriented (Ekman, 1993; Schröder, 2003; Strapparava et al., 2004). Pioneers have explored the advantages of combining facial expressions and speech signals to predict emotions (Tzirkakis et al., 2017; Wöllmer et al., 2010; Datcu and Rothkrantz, 2014; Zeng et al., 2007). Recent studies (Tsai et al., 2019a; Liang et al., 2018; Wang et al., 2019; Tsai et al., 2019b) have considered all the three modalities, whose primary focus is on fusion strategy while ignoring the emotion dynamics in a conversation. Notice that (Tsai et al., 2019b; Liang et al., 2018) take account of the intra-modal and cross-modal interactions between modalities, however, they ignore the context preference for each modality.

Emotion Recognition in Conversations is different from traditional emotion recognition due to emotion dynamics in conversations. By comparing with the recent proposed ERC approaches (Zhou et al., 2018; Majumder et al., 2019; Hsu et al., 2018), Poria et al. discovered that traditional emotion recognition approaches (Colnerić and Demsar, 2018; Kratzwald et al., 2018; Mohammad and Turney, 2010; Wu et al., 2006; Shaheen et al., 2014) failed to work well on ERC datasets, because the
same utterance within different context may exhibit different emotions (Poria et al., 2019b).

ERC is advancing in the recent few years. scLSTM (Poria et al., 2017) is an RNN-based approach that captures the self-dependency using a bi-directional LSTM. CMN (Hazarika et al., 2018b) and ICON (Hazarika et al., 2018a) distinguish the self and inter-personal dependencies by leveraging memory network. DialogueRNN (Majumder et al., 2019) uses multiple GRUs with global attention and further develops ERC to multi-party conversations. DialogueGCN (Ghosal et al., 2019) uses the Graph Convolutional Network (GCN) to model complex interactions between interlocutors. BiERU (Li et al., 2020) focus on the party-ignorant transferring of emotion in a conversation. Recently, several pieces of work, e.g., transfer learning ERC (Hazarika et al., 2019), and commonsense knowledge ERC (Ghosal et al., 2020), have employed pre-training models to the task of ERC. However, those approaches ignore the multi-modal emotion dynamics in conversations. Our dialogueTRM is specially designed to model such kinds of emotion dynamics.

Multi-modal Fusion seeks to generate a single representation to boost a specific task involving multiple modalities when building classifiers or other predictors. Many surveys (Guo et al., 2019; Kaur and Kautish, 2019; Angadi and Reddy, 2019) have investigated the strategies of multi-modal analysis with different kinds of clues. We divide fusion techniques into two groups.

The first is combination approaches, including concatenation (Majumder et al., 2019), hadamard product (Kiros et al., 2014), summing up (Mao et al., 2014), differential operation (Wu et al., 2018), gate (Mao et al., 2018), attention (Tsai et al., 2019a). According to whether there are interactions between features, those approaches can be categorized into linear weighting fusion (first three) and interactive weighting fusion (latter three). The second is learning approaches. According to the learning objective, approaches can be categorized as task-oriented and self-learning fusion. Task-oriented fusion (Frome et al., 2013; Hazarika et al., 2018a; Majumder et al., 2019) is for supervised learning, whose hidden states are the learned features. Self-learning fusion (Feng et al., 2014; Socher et al., 2014, 2013) is often unsupervised learned by structures like Restricted Boltzmann Machines (Srivastava and Salakhutdinov, 2012) or autoencoders (Ngiam et al., 2011). The strategy is to reconstruct source representation to target representation. Both source and target representations could be one or any combination of the multiple modalities (Feng et al., 2014; Ngiam et al., 2011).

Our MGIF has a similar idea with the Sub-View Attention (SVA) mechanism (Gu et al., 2019). The main differences are, 1) Our MGIF considers both prototype and representation granularities of feature interactions while SVA considers only the sub-view granularity; 2) Our MGIF can deal with multiple modalities while SVA is dyadic fusion.

3 Task Formulation
Let \( X = \{ x_i^\lambda \mid i \in [1, L], \lambda \in [1, N] \} \) be a conversation containing a sequence of \( L \) utterance-level expressions involving \( N \) speakers. At the \( i \)-th turn, the emotion of the \( \lambda_i \)-th speaker is conveyed through an expression \( x_i^\lambda = \{ x_{i,1}(\lambda), x_{i,2}(\lambda), \ldots, x_{i,K}(\lambda) \} \) in utterance \( x_{i,1}(\lambda), \) acoustic \( x_{i,2}(\lambda), \) and visual \( x_{i,3}(\lambda) \) modalities. According to the speakers that are involved, we define two types of context within a sliding window of \( K \), which are indi-context, \( \varphi_i^\lambda = \{ x_i^\lambda \mid \tau \in [\max(1, i - K), i], \lambda = \lambda_i \} \), and conv-context, \( \phi_i = \{ x_i^\lambda \mid \tau \in [\max(1, i - K), i], \lambda \in [1, N] \} \). Table 1 presents an example of the two types of contexts in a conversation.

| Table 1: Context examples in a conversation when \( L = 8, S = 3, \) and \( K = 5 \) |
|-------------------------------|-----------------|
| conversation | \( X = \{ x_1^1, x_2^1, x_3^1, x_4^1, x_5^1, x_6^1, x_7^1, x_8^1 \} \) |
| target | \( x_1^1 = x_1^1 \) |
| indi-context | \( \varphi_i^\lambda = \{ x_i^1, x_i^2 \} \) |
| conv-context | \( \phi_i = \{ x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5} \} \) |

4 Model
4.1 Intra-Modal Emotion Dynamics
The intra-modal emotion dynamics needs to not only capture the temporal dependency but also satisfy the context preference of different modalities. Transformer (Vaswani et al., 2017) can be easily switched to sequential structure for context-dependent modeling or feed-forward structure for context-free modeling. Thus, we use Transformer as the backbone. The modeling of intra-modal emotion dynamics is depicted on the left of Figure 2.
4.1.1 Context-Dependent Modeling

Emotions expressed in utterance modality prefer to be modeled in context-dependent settings. The self and inter-personal dependencies are two factors for context-dependent modeling.

**Self dependency.** Unlike traditional ERC approaches that separate the process of utterance encoding and dependency modeling, i.e., CNN encodes utterances and RNN learns dependencies among utterances (Majumder et al., 2019), we unify the two processes in one BERT (Devlin et al., 2019).

Specifically, BERT encodes each utterance by receiving a sequence of raw lexical input, containing information from not only the utterance itself but also the indi-context. Since the utterance-context pairs are spoken by the same speaker, the information relates to the self dependency is naturally preserved in the output representations of BERT. Additionally, there exists a length imbalance between the utterance and its context, we leverage the segment embeddings and [SEP] token in BERT to explicitly distinguish the utterance-context pair in a sequence rather than directly concatenating them.

Given an utterance, $x_{i(u)}$, and its indi-context, $\varphi_{i(u)}$, the procedure of feature encoding and self dependency modeling can be formulated as,

$$F_{i(u)}^\lambda = \text{BERT}([\text{CLS}]x_{i(u)}^\lambda[\text{SEP}]\varphi_{i(u)}^\lambda[\text{SEP}]),$$  

where $F_{i(u)}^\lambda$ is the utterance feature output at the [CLS] position of BERT. The feature retains the $\lambda$-th speaker information at the $i$-th conversation turn. [CLS] and [SEP] are special tokens in BERT.

**Inter-personal dependency.** Since the speaker information is retained, the inter-personal dependency can be modeled through interactions within speaker-based features obtained from last stage.

Rather than using graph convolutional networks to connect those features (Ghosal et al., 2019), we deploy deep layers of multi-head attention in a Transformer to calculate the interactions. Given an $L$-length feature sequence $F_{(u)} = \{F_{i(u)}|i \in [1, L]\}$, the interactions are calculated as,

$$r_{i(u)} = \text{Transformer}(F_{(u)}, \hat{\rho}_i),$$  

$$\hat{\rho}_i = \begin{bmatrix} 0 & 0 & \cdots & 0.11 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix},$$  

where $r_{i(u)}$ is the $i$-the turn utterance representation. $\hat{\rho}_i$ is an $L$-length attention mask. It masks the future and distant historical information, enforcing emotional interactions to be within a $K$-length conv-context. More information about attention mask can be found in (Kaitao et al., 2019).

4.1.2 Context-Free Modeling

Emotions expressed in acoustic and visual modalities prefer to be modeled in context-free settings. We follow (Hazariak et al., 2018b) that employs openSMILE (Eyben et al., 2010) and 3D-CNN (Tran et al., 2015) to extract acoustic features, $f_{i(a)}^\lambda$, and visual features, $f_{i(v)}^\lambda$, respectively. Both sources of features are extracted from utterance-level videos without any context information.

Given feature sequences $F_{(a)} = \{f_{i(a)}^\lambda|i \in [1, L]\}$ and $F_{(v)} = \{f_{i(v)}^\lambda|i \in [1, L]\}$, the acoustic and visual representations can be calculated as

$$r_{i(a)} = \text{Transformer}(F_{(a)}, \hat{\rho}_i),$$  

$$r_{i(v)} = \text{Transformer}(F_{(v)}, \hat{\rho}_i),$$  

$$\hat{\rho}_i = \begin{bmatrix} 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix},$$  

where $r_{i(a)}$ and $r_{i(v)}$ are the $i$-the turn acoustic and visual representations, respectively. $\hat{\rho}_i$ turns
on context-free settings, so that the interactions are within the target expression itself.

4.2 Inter-Modal Emotion Dynamics

The inter-modal emotion dynamics should consider multi-grained feature interactions to combine more predictive features from different modalities. The prototype and representation dependencies are two granularities for fusing multi-modal features. The modeling of inter-modal emotion dynamics is depicted in the middle of Figure 2.

4.2.1 Prototype Dependency

The prototype dependency can be learned through position-wise interactions between neurons of two equal-dimension vectors. We design a multi-modal gate to learn the prototype dependency, allocating different weights to neurons in each vector. Specifically, the multi-modal gate enforces a position-wise trade-off between two vectors, so that more predictive neurons are amplified in one vector, while the counterpart do the opposite. Instead of directly applying Hadamard product between two equal-dimension vectors (Fukui et al., 2016), our strategy has to compute a pair of weights. We adopt a neural network to compute the weights, taking the two candidate vectors as input. Furthermore, inspired by the softmax in the attention mechanism, we propose a position-wise normalization, that force a position-wise comparison for better learning the neuron importance. Given utterance \( r_i(u) \) and acoustic \( r_i(a) \) representations, our multi-modal gate is calculated as,

\[
\begin{align*}
    h_i(u) &= \tanh(W_i^u r_i(u)), \\
    h_i(a) &= \tanh(W_i^a r_i(a)), \\
    z_i(u,a) &= \sigma(W_i^z [r_i(u); r_i(a)]), \\
    h_i(u,a) &= z_i(u,a) * h_i(u), \\
    h_i(a,u) &= (1 - z_i(u,a)) * h_i(a),
\end{align*}
\]

Here, \( z_i(u,a) \) and \( 1 - z_i(u,a) \) are a pair of weights for neurons in \( h_i(u) \) and \( h_i(a) \), where “1-” operation behaves as the position-wise normalization. The normalization relates to a weight trade-off and enforces an explicit position-wise comparison between neurons in \( h_i(u) \) and \( h_i(a) \). The weight \( z_i(u,a) \) is computed based on interactions between \( r_i(u), r_i(a) \). \( \sigma \) ensures the weights ranging from 0 to 1. * is the Hadamard product. \( W \) are the weight matrices, \( z_i(u,a), h_i(u) \) and \( h_i(a) \) are equal-dimension vectors. \( h_i(u,a) \) and \( h_i(a,u) \) are representations after feature mapping. The above equations can be reformulated as,

\[
h_i(u,a), h_i(a,u) = \text{GATE}(r_i(u), r_i(a)) \tag{12}
\]

Similarly, we can obtain

\[
\begin{align*}
    h_i(u), h_i(v) &= \text{GATE}(r_i(u), r_i(v)), \tag{13} \\
    h_i(a), h_i(v) &= \text{GATE}(r_i(a), r_i(v)). \tag{14}
\end{align*}
\]

4.2.2 Representation Dependency

The representation dependency is modeled through interactions in a sequence of six gated representations, allocating one weight to one representation. The interactions are calculated via deep layers of multi-head attention in a Transformer. Specifically, the procedure is as follows, (1) packing the multimodal representations into a sequence with a fixed order; (2) inserting a special embedding, \( e_{CLS} \), at the head of the sequence, similar to that in BERT; (3) feeding the sequence to a Transformer and calculating deep multi-head attention for representation dependency, formulated as,

\[
M_i = e_{CLS} h_i(u) h_i(a) h_i(u) h_i(v) h_i(a) h_i(v), \tag{15}
\]

\[
o_i = \text{Transformer}(M_i, \rho_i), \tag{16}
\]

where \( o_i \) is the final representation output at the \( e_{CLS} \) position, and \( \rho_i \) is the attention mask that sets all positions to ones.

5.3 Discriminator

The discriminator uses a two-layer perceptron with hidden layer activated by \( \tanh \). As shown in the right of Figure 2, we use the softmax for Categorical Emotion (CE) and linear layer for Dimensional Emotion (DE), denoted by,

\[
\mathcal{P}_i = \begin{cases} 
\text{softmax}(W^c \tanh(W^o o_i)), & \text{for CE;} \\
W^d \tanh(W^o o_i), & \text{for DE,} 
\end{cases} \tag{17}
\]

\[
\hat{y}_i = \begin{cases} 
\arg \max_j \mathcal{P}_i[j], & \text{for CE;} \\
\mathcal{P}_i, & \text{for DE,}
\end{cases} \tag{18}
\]

where \( W \) are the weight matrices, \( \hat{y}_i \) is the predicted emotion.

5. Experiment

5.1 Datasets

Three benchmark datasets, IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), and
Table 2: Main results on three benchmarks. “-M” and “-U” denote models using multi-modal or utterance-only settings. MM denotes if models use multi-modal settings. “-” represents no results reported in original paper.

<table>
<thead>
<tr>
<th>Models</th>
<th>MM</th>
<th>IEMOCAP ACC</th>
<th>MELD ACC</th>
<th>AVEC Valence</th>
<th>AVEC Arousal</th>
<th>AVEC Expectancy</th>
<th>AVEC Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c-LSTM-U</td>
<td>×</td>
<td>56.3</td>
<td>56.1</td>
<td>-</td>
<td>56.7</td>
<td>0.16</td>
<td>0.25</td>
</tr>
<tr>
<td>AGHMN</td>
<td>×</td>
<td>63.5</td>
<td>63.5</td>
<td>59.5</td>
<td>57.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DGCN</td>
<td>×</td>
<td>65.3</td>
<td>64.2</td>
<td>-</td>
<td>58.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BiERU</td>
<td>×</td>
<td>66.1</td>
<td>64.7</td>
<td>-</td>
<td>60.8</td>
<td>0.36</td>
<td>0.64</td>
</tr>
<tr>
<td>DialogueTRM-U</td>
<td>×</td>
<td>68.2</td>
<td>68.1</td>
<td>64.6</td>
<td>63.2</td>
<td>0.73</td>
<td>0.44</td>
</tr>
<tr>
<td>c-LSTM-M</td>
<td>√</td>
<td>59.8</td>
<td>59.0</td>
<td>-</td>
<td>-</td>
<td>0.14</td>
<td>0.23</td>
</tr>
<tr>
<td>CMN</td>
<td>√</td>
<td>61.9</td>
<td>61.4</td>
<td>-</td>
<td>-</td>
<td>0.23</td>
<td>0.30</td>
</tr>
<tr>
<td>DRNN</td>
<td>√</td>
<td>63.4</td>
<td>62.7</td>
<td>56.1</td>
<td>57.0</td>
<td>0.35</td>
<td>0.59</td>
</tr>
<tr>
<td>ICON</td>
<td>√</td>
<td>64.0</td>
<td>63.5</td>
<td>-</td>
<td>-</td>
<td>0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>DialogueTRM-M</td>
<td>√</td>
<td>69.5</td>
<td>69.7</td>
<td>65.7</td>
<td>63.5</td>
<td>0.76</td>
<td>0.52</td>
</tr>
</tbody>
</table>

AVEC (Schuller et al., 2012), are adopted to evaluate our model. IEMOCAP consists of 151 dyadic conversation videos with 6 emotion types. Following (Majumder et al., 2019), we apply the first four sessions for training and the last for testing. The validation is randomly selected from the training set with a ratio of 0.1. MELD consists of 1433 multi-party conversation videos with 7 emotion types. We apply the official splits for training, validation, and testing. The visual source may involve multiple speakers and is hard to use. Thus, experiments on MELD do not use visual information. AVEC consists of 95 dyadic conversation videos with four real-value annotations per utterance in terms of Valence, Arousal, Expectancy, and Power (Mehrabian, 1996). We apply the official splits for training and testing. The validation is randomly selected from the training set with a ratio of 0.1.

5.2 Implementation Details

We use the off-the-shelf pre-trained BERT base model with default parameters and finetune it during training. It outputs 768-dimensional utterance features. The visual and acoustic features are fixed 512- and 100-dimensional vectors, respectively, obtained from an open-source project. Those vectors are projected to 768 dimensions to match the input size. The intra-modal component, a 6-layer, 12-head-attention, and 768 hidden-unit Transformer encoder, is implemented with PyTorch API using default parameters. For inter-modal modeling, we construct a 4-layer, 8-head-attention, and 768-hidden-unit Transformer encoder. We use AdamW (Loshchilov and Hutter) as the optimizer with initial learning rate= 6e-6, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, learning rate warmup over the first 1, 200 steps, and linear decay of the learning rate. To make it easy for reproduction, our model does not apply to multi-GPU settings. Our hardware (11GB GPU memory) affords a maximum context window of 14. A larger context can achieve better performance (Jiao et al., 2019) which is beyond the concern of this paper.

5.3 Main Results

Traditional baselines of ERC can be divided into two groups. One is utterance-only based models, including c-LSTM-U (Poria et al., 2017), the earliest study we can track in ERC, AGHMN (Jiao et al., 2019), an attention gated hierarchical memory network, DGCN (Ghosal et al., 2019), using graph neural network to address context propagation issue, and BiERU (Li et al., 2020), applying a party-ignorant bidirectional emotional recurrent unit for ERC. The other is multi-modal based models, including c-LSTM-M, the multi-modal version of c-LSTM-U, CMN (Hazarika et al., 2018b), the first memory network based ERC model, DRNN (Majumder et al., 2019), the first approach for multi-party ERC, ICON (Hazarika et al., 2018a), developing CMN with more emotional interactions.

The results are based on an average of 5 runs and are presented in Table 2. Following (Majumder et al., 2019), we use weighted average ACCuracy (ACC) and F1 Score (F1) to evaluate the categor-
To test the effect of our multi-grained interactive fusion, we perform a comparison with other fusion strategies. Additive, Concat, and Max-pooling are three simple fusion methods that add, concatenate and max-pool multi-modal features, respectively. Bilinear (Fukui et al., 2016), GMU (Arevalo et al., 2020), and MulT (Tsai et al., 2019a) are three advanced single-grained fusion methods, in which the first two approaches only capture the prototype dependency, and the last one only captures the representation dependency. The results are shown in Table 5, and we focus on the performance gained from utterance-only to multi-modal settings. The performance of MulT is limited because the model forces all the modalities to use context-dependent settings. The

The temporal aspect. To verify that different modalities exhibit different dependence on context information, we present results for different combinations of modalities in Table 4. We manage the context setting using attention masks in Transformers. We use $\dot{u}$ and $\ddot{u}$ to denote context-free and context-dependent settings, respectively. As seen, emotions in visual and acoustic modalities prefer context-free settings. An intuitive explanation is that identifying emotions from acoustic or visual modalities is based on very concrete features, e.g., frown or loudness for “angry”. If we incorporate previously extracted features, e.g., tear or sob for “sad”, it becomes ambiguous for predicting the “angry”. The emotion modeling in utterance modality strongly depends on context information and dominates the performance. Thus, our strategy of using multi-modal information is to satisfy their context preference, while previous methods indiscriminately apply context-dependent settings.

The spatial aspect. To better understand multi-modal emotion dynamics, we conduct a series of experiments to test its effect from different aspects.

5.4 Analysis

To better understand multi-modal emotion dynamics, we conduct a series of experiments to test its effect from different aspects.
Table 5: Fusion results on IEMOCAP showing the F1 performance (gain) from (U)terence-only to (M)ulti-modal settings.

<table>
<thead>
<tr>
<th>Fusion Techniques</th>
<th>$U \rightarrow M_{gain}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive</td>
<td>68.1 $\rightarrow$ 68.6.0,5↑</td>
</tr>
<tr>
<td>Concat</td>
<td>68.1 $\rightarrow$ 68.5.4↑</td>
</tr>
<tr>
<td>Max-pooling</td>
<td>68.1 $\rightarrow$ 68.7.6↑</td>
</tr>
<tr>
<td>Bilinear</td>
<td>68.1 $\rightarrow$ 69.0.9↑</td>
</tr>
<tr>
<td>GMU</td>
<td>68.1 $\rightarrow$ 68.8.7↑</td>
</tr>
<tr>
<td>MuT</td>
<td>68.1 $\rightarrow$ 68.4.0.3↑</td>
</tr>
<tr>
<td>Our MGIF</td>
<td>68.1 $\rightarrow$ 69.7.1.6↑</td>
</tr>
<tr>
<td>w/o representation dependency</td>
<td>68.1 $\rightarrow$ 69.2.1,1↑</td>
</tr>
<tr>
<td>w/o prototype dependency</td>
<td>68.1 $\rightarrow$ 68.9.0.8↑</td>
</tr>
</tbody>
</table>

gain of our MGIF is markedly higher than those of single-grained approaches. Furthermore, we conduct an ablation study on MGIF. The results are presented in the last two rows of Table 5, including w/o representation dependency, i.e., concatenating the six gated representations without using the Transformer, and w/o prototype dependency, i.e., directly using the Transformer to wrap representations without multi-modal gate. We find that prototype dependency contributes more to MGIF.

Utterance context modeling. Since utterance modality dominates the performance, we conduct an ablation study on utterance context modeling. Specifically, we step by step remove some key operations in DialogueTRM-U. The results are listed in the last three rows of Table 3. We can find that differentiating utterance and context is effective, and segment embedding contributes more to such differentiation.

5.5 Case Study

Short utterance cases. “yeah.” appears 23 times in the test set. Given only the target utterance, the accuracy is 43.48%. After adding utterance context, it increases to 65.22%. After adding multi-modal information, it arrives at 73.91%.

Multi-modal rectified cases. We analyze cases that incorrectly predicted in utterance-only settings but correctly predicted in multi-modal settings. Among the cases, “neutral” and “frustrated” are in the majority with the ratios of about 30.38% and 27.85%, respectively. Moreover, about 85.41% “neutral” and 70.45% “frustrated” cases are rectified from negative emotions. It means multi-modal provides easy-to-distinguish information for negative emotions. The reason is probably that human tends to use neutral words to cover their negative emotions yet show up in the faces or intonations.

Emotion shift cases. We analyze cases that exhibit Intra-speaker Emotion Shift (Intra-ES), e.g., the emotion shift from person A at $T+1$ to person A at $T+3$ in Figure 3, and Inter-speaker Emotion Shift (Inter-ES), e.g., the emotion shift from person B at $T+2$ to person A at $T+3$ in Figure 3. We present the results in table 6. Note that our model mainly improves the performance of Inter-ES cases and is relatively poor for Intra-ES cases. It provides a direction for future studies.

6 Conclusion and future work

This paper describes a novel understanding of emotion dynamics in multi-modal conversations. The proposed DialogueTRM provides a straightforward yet effective strategy to model both intra-modal and inter-modal emotion dynamics for the task of ERC. Satisfying context preferences of different modalities and multi-grained interactive fusion are two major factors that our model addresses. In the future, we would formulate more principles for analyzing complex emotional behaviors in conversations, e.g., addressing the limitation of our model for intra-speaker emotion shift.
References


Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2020. Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. CoRR.


Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2:207–218.


