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

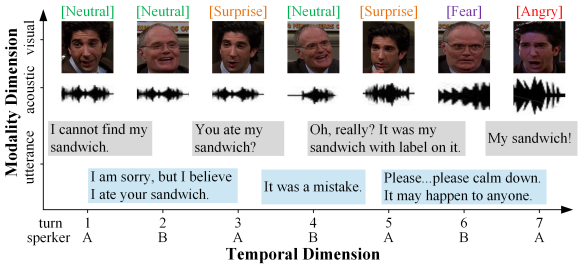

Figure 1: An example of multi-modal conversation.

(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-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2694–2704
November 7–11, 2021.

teristic is ignored by existing studies on ERC. Intuitively, spoken words are highly semantic that require inferences from the context to understand the emtions (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 (Tzirakis 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 (Colnerič 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., 2019), 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 $X = \{x_i^{\lambda_i} | i \in [1, L], \lambda_i \in [1, N]\}$ be a conversation containing a sequence of L utterance-level expressions involving N speakers. At the i-th turn, the emtion of the λ_i-th speaker is conveyed through an expression $x_i^{\lambda_i} = \{x_{i,(u)}^{\lambda_i}, x_{i,(a)}^{\lambda_i}, x_{i,(v)}^{\lambda_i}\}$ in utterance $x_{i,(u)}^{\lambda_i}$, acoustic $x_{i,(a)}^{\lambda_i}$ and visual $x_{i,(v)}^{\lambda_i}$ modalities. According to the speakers that are involved, we define two types of context within a sliding window of K, which are *indi-context*, $\varphi_i^{\lambda_i} = \{x_\tau^\lambda | \tau \in [\max(1, i-K), i), \lambda = \lambda_i\}$, and *conv-context*, $\phi_i = \{x_\tau^\lambda | \tau \in [\max(1, i-K), i), \lambda \in [1, N]\}$. 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	$X = \{x_1^1, x_2^1, x_3^2, x_4^1, x_5^3, x_6^2, x_7^1, x_8^2\}$
target	$x_i^{\lambda_i} = x_7^1$
indi-context	$\varphi_i^{\lambda_i} = \{x_2^1, x_4^1\}$
conv-context	$\phi_i = \{x_2^1, x_3^2, x_4^1, x_5^3, x_6^2\}$

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.

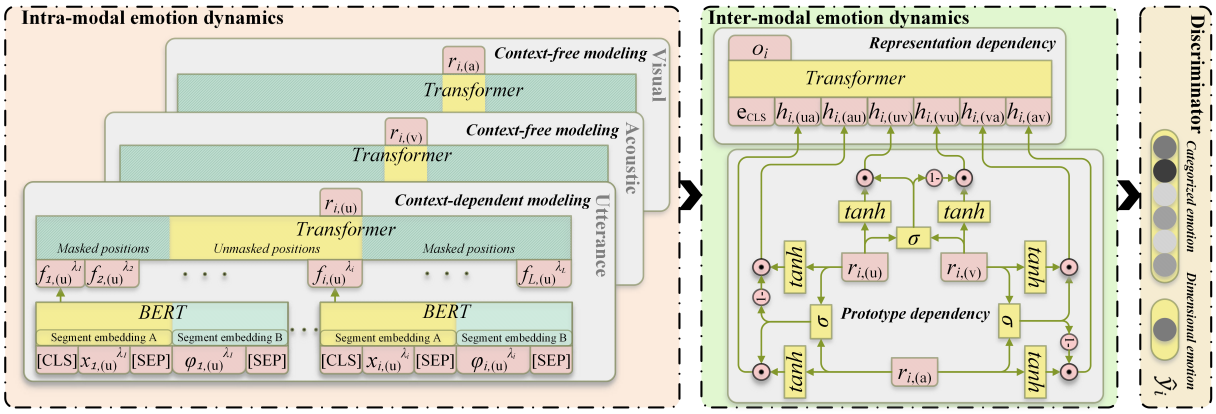

Figure 2: Model architecture.

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, $x^{\lambda_i}_{i,(u)}$, and its *indi-context*, $\varphi^{\lambda_i}_{i,(u)}$, the procedure of feature encoding and self dependency modeling can be formulated as,

$$f^{\lambda_i}_{i,(u)} = \text{BERT}([\text{CLS}]x^{\lambda_i}_{i,(u)}[\text{SEP}]\varphi^{\lambda_i}_{i,(u)}[\text{SEP}]), \quad (1)$$

where $f^{\lambda_i}_{i,(u)}$ is the utterance feature output at the [CLS] position of BERT. The feature retains the λ_i-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. Ginve an L-length feature sequence $F_{(u)}=\{f^{\lambda_i}_{i,(u)}|i\in[1,L]\}$, the interactions are calulated as,

$$r_{i,(u)} = \text{Transformer}(F_{(u)}, \bar{\rho}_i), \quad (2)$$

$$\bar{\rho}_i = \underbrace{00\cdots0}_{i-K-1}\underbrace{11\cdots1}_{K}1\underbrace{00\cdots0}_{L-i}, \quad (3)$$

where $r_{i,(u)}$ is the i-the turn utterance representation. $\bar{\rho}_i$ 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 (Hazarika et al., 2018b) that employs openSMILE (Eyben et al., 2010) and 3D-CNN (Tran et al., 2015) to extract acoustic features, $f^{\lambda_i}_{i,(a)}$, and visual features, $f^{\lambda_i}_{i,(v)}$, respectively. Both sources of features are extracted from utterance-level videos without any context information.

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

$$r_{i,(a)} = \text{Transformer}(F_{(a)}, \dot{\rho}_i), \quad (4)$$

$$r_{i,(v)} = \text{Transformer}(F_{(v)}, \dot{\rho}_i), \quad (5)$$

$$\dot{\rho}_i = \underbrace{00\cdots0}_{i-1}1\underbrace{00\cdots0}_{L-i}, \quad (6)$$

where $r_{i,(a)}$ and $r_{i,(v)}$ are the i-the turn acoustic and visual representations, respectively. $\dot{\rho}_i$ 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,

$$h_{i,(u)} = tanh(W^U r_{i,(u)}), \tag{7}$$

$$h_{i,(a)} = tanh(W^A r_{i,(a)}), \tag{8}$$

$$z_{i,(ua)} = \sigma(W^Z [r_{i,(u)}; r_{i,(a)}]), \tag{9}$$

$$h_{i,(ua)} = z_{i,(ua)} * h_{i,(u)}, \tag{10}$$

$$h_{i,(au)} = (1 - z_{i,(ua)}) * h_{i,(a)}. \tag{11}$$

Here, $z_{i,(ua)}$ and $1 - z_{i,(ua)}$ 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* $z_{i,(ua)}$ *is computed* based on interactions between $r_{i,(u)}$, $r_{i,(a)}$. σ ensures the weights ranging from 0 to 1. $*$ is the Hadamard product. W are the weight matrices. $z_{i,(ua)}$, $h_{i,(u)}$ and $h_{i,(a)}$ are equal-dimension vectors. $h_{i,(ua)}$ and $h_{i,(au)}$ are representations after feature mapping. The above equantions can be refomulated as,

$$h_{i,(ua)}, h_{i,(au)} = \text{GATE}(r_{i,(u)}, r_{i,(a)}) \tag{12}$$

Similarly, we can obtain

$$h_{i,(uv)}, h_{i,(vu)} = \text{GATE}(r_{i,(u)}, r_{i,(v)}), \tag{13}$$

$$h_{i,(av)}, h_{i,(va)} = \text{GATE}(r_{i,(a)}, r_{i,(v)}). \tag{14}$$

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 multi-modal 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,

$$M_i = e_{CLS} h_{i,(ua)} h_{i,(au)} h_{i,(uv)} h_{i,(vu)} h_{i,(av)} h_{i,(va)}, \tag{15}$$

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

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

4.3 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.

Models	MM	IEMOCAP		MELD		AVEC			
		ACC	F1	ACC	F1	Valence	Arousal	Expectancy	Power
c-LSTM-U	×	56.3	56.1	-	56.7	0.16	0.25	0.24	0.10
AGHMN	×	63.5	63.5	59.5	57.5	-	-	-	-
DGCN	×	65.3	64.2	-	58.1	-	-	-	-
BiERU	×	66.1	64.7	-	60.8	0.36	**0.64**	0.38	0.37
DialogueTRM-U	×	68.2	68.1	64.6	63.2	0.73	0.44	0.38	0.32
c-LSTM-M	√	59.8	59.0	-	-	0.14	0.23	0.25	-0.04
CMN	√	61.9	61.4	-	-	0.23	0.30	0.26	-0.02
DRNN	√	63.4	62.7	56.1	57.0	0.35	0.59	0.37	0.37
ICON	√	64.0	63.5	-	-	0.23	0.29	0.26	0.22
DialogueTRM-M	√	**69.5**	**69.7**	**65.7**	**63.5**	**0.76**	0.52	**0.40**	**0.40**

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 $\text{BERT}_{\text{base}}$ 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[1]. 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-

[1]https://github.com/SenticNet/conv-emotion

Table 3: Comparison with recent ERC methods using pre-training techniques on IEMOCAP.

Models	ACC	F1
$\text{BERT}_{\text{base}}$	56.7	56.4
TL-ERC	-	58.8
DRNN§	-	64.7
COSMIC	-	65.2
DialogueTRM-U	**68.2**	**68.1**
remove transformer	67.4	67.4
remove segment embedding	65.8	65.7
remove [SEP] tokens	65.3	65.2

Table 4: Analysis of (u)tterance, (a)coustic, (v)isual expressions in different context settings on IEMOCAP. $\dot{*}$ and $\bar{*}$ denote context-free and context-dependent settings. ‡ means our context settings. † means context settings in other studies

	ACC	F1		ACC	F1
$\dot{u}$	56.7	56.4	$\bar{u}$	68.2	68.1
$\dot{a}$	46.8	44.9	$\bar{a}$	44.7	42.9
$\dot{v}$	33.6	36.8	$\bar{v}$	32.2	33.7
$\dot{a}$+$\dot{v}$	50.5	49.4	$\bar{a}$+$\bar{v}$	47.7	47.0
$\dot{a}$+$\dot{v}$+$\dot{u}$	58.8	58.3	$\bar{a}$+$\bar{v}$+$\dot{u}$	57.2	57.1
$\dot{a}$+$\dot{v}$+$\bar{u}$‡	**69.5**	**69.7**	$\bar{a}$+$\bar{v}$+$\bar{u}$†	68.9	68.8

ical emotions on IEMOCAP and MELD and use pearson's corRelation coefficient (R) to evaluate the dimensional emotions on AVEC. From the result, the ACC and F1 of our DialogueTRM-M markedly outperform SOTA, indicating 5%, 7% improvements on IEMOCAP and 10%, 4% improvements on MELD, respectively. On AVEC, DialogueTRM-M outperforms SOTA in most of the criteria, which are 16%, 5%, 8% improvements in Valence, Expectancy, and Power, respectively. Since the utterance features of traditional baselines are based on CNN, the improvement is partly due to the boosting from BERT.

For fair comparisons, we investigate some very recent ERC approaches that incorporate pretraining techniques. The results are presented in Table 3. *BERT_{base}* is identical to the utterance encoder without modeling self dependency. *TL-ERC* (Hazarika et al., 2019) leverage BERT to transfer affective knowledge from a general-domain conversational corpus to the task of ERC. *COSMIC* (Ghosal et al., 2020) is based on RoBERTa, a more powerful pre-training model than BERT, and incorporates DRNN with commonsense for ERC. *DRNN§* is DRNN with RoBERTa features reported in (Ghosal et al., 2020). Since BiERU is not open-sourced, we cannot present its result in pre-training settings. All the methods are in utterance-only settings on IEMOCAP. DialogueTRM-U markedly outperforms those methods. We believe our results can help build a comparable baseline for future studies addressing ERC with pre-training techniques.

5.4 Analysis

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

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 $\dot{*}$ and $\bar{*}$ 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 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. *Bilnear* (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

Table 5: Fusion results on IEMOCAP showing the F1 performance (gain) from (U)tterance-only to (M)ulti-modal settings.

Fusion Techniques	$U \rightarrow M_{gain}$
Additive	$68.1 \rightarrow 68.6_{0.5\uparrow}$
Concat	$68.1 \rightarrow 68.5_{0.4\uparrow}$
Max-pooling	$68.1 \rightarrow 68.7_{0.6\uparrow}$
Bilnear	$68.1 \rightarrow 69.0_{0.9\uparrow}$
GMU	$68.1 \rightarrow 68.8_{0.7\uparrow}$
MulT	$68.1 \rightarrow 68.4_{0.3\uparrow}$
Our MGIF	$68.1 \rightarrow \mathbf{69.7_{1.6\uparrow}}$
w/o representation dependency	$68.1 \rightarrow 69.2_{1.1\uparrow}$
w/o prototype dependency	$68.1 \rightarrow 68.9_{0.8\uparrow}$

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.

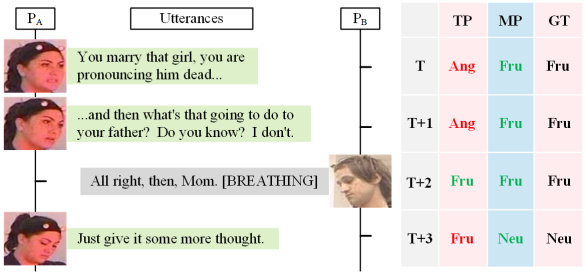

Figure 3: Conversation cases with MP (Multi-modal-Predicted), TP (Text-Predicted) and GT (Ground-Truth) emotions, where 'Neu', 'Exc', 'Fru' stands for neural, excited and frustrated, respectively.

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.

Table 6: Performance of cases that exhibit Intra-ES and Inter-ES on IEMOCAP. Numbers in parenthesis indicate the average count of the corresponding shifts per conversation. We present the OriGinal (OG) performance for comparision.

Models	OG		Intra-ES (13.2)		Inter-ES (22.0)	
	ACC	F1	ACC	F1	ACC	F1
DialogueTRM-U	68.2	68.1	52.9	52.9	73.8	73.8
DialogueTRM-M	69.5	69.7	55.1	55.4	74.7	74.3

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

Sujay Angadi and R Venkata Siva Reddy. 2019. Survey on sentiment analysis from affective multimodal content. In *Smart intelligent computing and applications*, pages 599–607. Springer.

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González. 2020. Gated multimodal networks. *Neural Computing and Applications*, pages 1–20.

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.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Niko Colneriĉ and Janez Demsar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*.

Dragos Datcu and Leon JM Rothkrantz. 2014. Semantic audio-visual data fusion for automatic emotion recognition. *Emotion recognition: a pattern analysis approach*, pages 411–435.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.

A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of EMNLP*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP*, Hong Kong, China. Association for Computational Linguistics.

Yue Gu, Xinyu Lyu, Weijia Sun, Weitian Li, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2019. Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 157–166.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

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.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2019. Emotion recognition in conversations with transfer learning from generative conversation modeling. *arXiv preprint arXiv:1910.04980*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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. 2019. Real-time emotion recognition via attention gated hierarchical memory network. *arXiv preprint arXiv:1911.09075*.

Song Kaitao, Tan Xu, Qin Tao, Lu Jianfeng, and Liu Tie-yan. 2019. Mass:masked sequence to sequence pre-training for language generation. In *ICML*.

Ramandeep Kaur and Sandeep Kautish. 2019. Multimodal sentiment analysis: A survey and comparison. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, 10(2):38–58.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International conference on machine learning*, pages 595–603.

Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Decision support with text-based emotion recognition: Deep learning for affective computing. *arXiv preprint arXiv:1803.06397*.

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

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161.

I Loshchilov and F Hutter. Fixing weight decay regularization in adam, corr, abs/1711.05101. In *Proceedings of the ICLR 2018 Conference Blind Submission, Vancouver, BC, Canada*, volume 30.

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.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. 2018. Show and tell more: Topic-oriented multi-sentence image captioning. In *IJCAI*, pages 4258–4264.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Michael W Morris and Dacher Keltner. 2000. How emotions work: The social functions of emotional expression in negotiations. *Research in organizational behavior*, 22:1–50.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning.

Rosalind W Picard. 2010. Affective computing: from laughter to ieee. *IEEE Transactions on Affective Computing*, 1(1):11–17.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. 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.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Marc Schröder. 2003. Experimental study of affect bursts. *Speech communication*, 40(1-2):99–116.

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*, pages 449–456.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, pages 383–392. IEEE.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

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.

Nitish Srivastava and Russ R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. In *International Conference on Representation Learning*.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

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.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365.

Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Ruifan Li. 2019. Differential networks for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8997–9004.

Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2):165–183.

Zhihong Zeng, Yuxiao Hu, Glenn I Roisman, Zhen Wen, Yun Fu, and Thomas S Huang. 2007. Audio-visual spontaneous emotion recognition. In *Artifical intelligence for human computing*, pages 72–90. Springer.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.