DialogueRNN: An Attentive RNN for Emotion Detection in Conversations

Navonil Majumder†‡, Soujanya Poria‡‡, Devamanyu HazarikaΦ, Rada MihalceaΦ, Alexander Gelbukh†, Erik Cambria‡

†Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
‡School of Computer Science and Engineering, Nanyang Technological University, Singapore
ΦSchool of Computing, National University of Singapore, Singapore
‡‡Computer Science & Engineering, University of Michigan, Ann Arbor, USA

Abstract

Emotion detection in conversations is a necessary step for a number of applications, including opinion mining over chat history, social media threads, debates, argumentation mining, understanding consumer feedback in live conversations, and so on. Currently systems do not treat the parties in the conversation individually by adapting to the speaker of each utterance. In this paper, we describe a new method based on recurrent neural networks that keeps track of the individual party states throughout the conversation and uses this information for emotion classification. Our model outperforms the state-of-the-art by a significant margin on two different datasets.

1 Introduction

Emotion detection in conversations has been gaining increasing attention from the research community due to its applications in many important tasks such as opinion mining over chat history, social media threads, debates, argumentation mining, understanding consumer feedback in live conversations, and so on. In this paper, we present a method based on recurrent neural networks (RNN) that can cater to these needs by processing the huge amount of available conversational data.

Current systems, including the state of the art (Hazarika et al. 2018), do not distinguish different parties in a conversation in a meaningful way. They are not aware of the speaker of a given utterance. In contrast, we model individual parties with party states, as the conversation flows, by relying on the utterance, the context, and current party state. Our model is based on the assumption that there are three major aspects relevant to the emotion in a conversation: the speaker, the context from the preceding utterances, and the emotion of the preceding utterances. These three aspects are not necessarily independent, but their separate modeling significantly outperforms the state-of-the-art (Table 2). In dyadic conversations, the parties have distinct roles. Hence, to extract the context, it is crucial to consider the preceding turns of both the speaker and the listener at a given moment (Fig. 1).

Our proposed DialogueRNN system employs three gated recurrent units (GRU) (Chung et al. 2014) to model these aspects. The incoming utterance is fed into two GRUs called global GRU and party GRU to update the context and party state, respectively. The global GRU encodes corresponding party information while encoding an utterance. Attending over this GRU gives contextual representation that has information of all preceding utterances by different parties in the conversation. The speaker state depends on this context through attention and the speaker’s previous state. This ensures that at time $t$, the speaker state directly gets information from the speaker’s previous state and global GRU which has information on the preceding parties. Finally, the updated speaker state is fed into the emotion GRU to decode the emotion representation of the given utterance, which is used for emotion classification. At time $t$, the emotion GRU cell gets the emotion representation of $t-1$ and the speaker state of $t$.

The emotion GRU, along with the global GRU, plays a pivotal role in inter-party relation modeling. On the other hand, party GRU models relation between two sequential states of the same party. In DialogueRNN, all these three different types of GRUs are connected in a recurrent manner. We believe that DialogueRNN outperforms state-of-the-art contextual emotion classifiers such as (Hazarika et al. 2018; Poria et al. 2017) because of better context representation.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 3 provides detailed description of our model; Sections 4 and 5 present the experimental results; finally, Section 6 concludes the paper.

2 Related Work

Emotion recognition has attracted attention in various fields such as natural language processing, psychology, cognitive science, and so on (Picard 2010), Ekman (1993) found correlation between emotion and facial cues. Datcu and Rothkrantz (2008) fused acoustic information with visual cues for emotion recognition. Alm, Roth, and Sproat (2005) introduced text-based emotion recognition, developed in the work of Strapparava and Mihalcea (2010). Wöllmer et al. (2010) used contextual information for emotion recognition in multimodal setting. Recently, Poria et al. (2017) successfully used RNN-based deep networks for multimodal emotion recognition, which was followed by other works (Chen et al. 2017; Zadeh et al. 2018a; 2018b).

Reproducing human interaction requires deep understanding of conversation. Ruusuvuori (2013) states that emotion plays a pivotal role in conversations. It has been ar-
guessed that emotional dynamics in a conversation is an inter-
personal phenomenon (Richards, Butler, and Gross 2003). 
Hence, our model incorporates inter-personal interactions in 
an effective way. Further, since conversations have a natu-
ral temporal nature, we adopt the temporal nature through 
recurrent network (Poria et al. 2017).

Memory networks (Sukhbaatar et al. 2015) has been 
successful in several NLP areas, including question an-
wering (Sukhbaatar et al. 2015; Kumar et al. 2016), ma-
chine translation (Bahdanau, Cho, and Bengio 2014), speech 
recognition (Graves, Wayne, and Danihelka 2014), and so 
on. Thus, Hazarika et al. (2018) used memory networks for 
emotion recognition in dyadic conversations, where two 
distinct memory networks enabled inter-speaker interaction, 
yielding state-of-the-art performance.

3 Methodology

3.1 Problem Definition

Let there be \( M \) parties/participants \( p_1, p_2, \ldots, p_M \) (\( M = 2 \)
for the datasets we used) in a conversation. The task is to pre-
dict the emotion labels (happy, sad, neutral, angry, excited, 
and frustrated) of the constituent utterances \( u_1, u_2, \ldots, u_N \), 
where utterance \( u_t \) is uttered by party \( p_s(u_t) \), while \( s \) being 
the mapping between utterance and index of its correspond-
ing party. Also, \( u_t \in \mathbb{R}^{D_m} \) is the utterance representation, 
obtained using feature extractors described below.

3.2 Unimodal Feature Extraction

For a fair comparison with the state-of-the-art method, con-
versational memory networks (CMN) (Hazarika et al. 2018), 
we follow identical feature extraction procedures.

Textual Feature Extraction We employ convolutional 
near networks (CNN) for textual feature extraction. Fol-
lowing Kim (2014), we obtain n-gram features from each 
utterance using three distinct convolution filters of sizes 3, 
4, and 5 respectively, each having 50 feature-maps. Outputs 
are then subjected to max-pooling followed by rectified lin-
ear unit (ReLU) activation. These activations are concate-
nated and fed to a 100 dimensional dense layer, which is re-
garded as the textual utterance representation. This network 
is trained at utterance level with the emotion labels.

Audio and Visual Feature Extraction Identical to Haz-
arika et al. (2018), we use 3D-CNN and openSMILE (Ey-
ben, Wöllmer, and Schuller 2010) for visual and acoustic 
feature extraction, respectively.

3.3 Our Model

We assume that the emotion of an utterance in a conversation 
depends on three major factors:
1. the speaker.
2. the context given by the preceding utterances.
3. the emotion behind the preceding utterances.

Our model DialogueRNN, shown in Fig. 2a, models these 
three factors as follows: each party is modeled using a party 
state which changes as and when that party utters an utter-
ance. This enables the model to track the parties’ emotion 
dynamics through the conversations, which is related to the 
emotion behind the utterances. Furthermore, the context of 
an utterance is modeled using a global state (called global, 
because of being shared among the parties), where the pre-
ceding utterances and the party states are jointly encoded 
for context representation, necessary for accurate party state 
representation. Finally, the model infers emotion represen-
tation from the party state of the speaker along with the pre-
ceding speakers’ states as context. This emotion representation 
is used for the final emotion classification.

We use GRU cells (Chung et al. 2014) to update the states 
and representations. Each GRU cell computes a hidden state 
defined as \( h_t = GRU_s(h_{t-1}, x_t) \), where \( x_t \) is the current 
input and \( h_{t-1} \) is the previous GRU state. \( h_t \) also serves as 
the current GRU output. We provide the GRU computation 
details in the supplementary. GRUs are efficient networks 
with trainable parameters: \( W_s^{r,z,c} \) and \( b_s^{r,z,c} \).

We model the emotion representation of the current ut-
terance as a function of the emotion representation of the 
previous utterance and the state of the current speaker. Fi-
nally, this emotion representation is sent to a softmax layer 
for emotion classification.

Global State (Global GRU) Global state aims to capture 
the context of a given utterance by jointly encoding utterance 
and speaker state. Each state also serves as speaker-specific 
utterance representation. Attending on these states facilitates 
the inter-speaker and inter-utterance dependencies to pro-
duce improved context representation. The current utterance 
\( u_t \) changes the speaker’s state from \( q_{s(u_t),t-1} \) to \( q_{s(u_t),t} \). We 
capture this change with GRU cell \( GRU_G \) with output size 
\( D_g \), using \( u_t \) and \( q_{s(u_t),t-1} \):

\[
q_t = GRU_G(q_{t-1}, (u_t \oplus q_{s(u_t),t-1})), \tag{1}
\]

where \( D_g \) is the size of global state vector, \( D_P \) is the 
size of party state vector, \( W_{g,h}^{r,z,c} \in \mathbb{R}^{D_g \times D_g} \), \( W_{g,x}^{r,z,c} \in \mathbb{R}^{D_g \times (D_m + D_P)} \), \( b_g^{r,z,c} \in \mathbb{R}^{D_g} \), \( q_{s(u_t),t-1} \in \mathbb{R}^{D_P} \), \( g_t, g_{t-1} \in \mathbb{R}^{D_g} \), \( D_P \) is party state size, and \( \oplus \) represents concatenation.

3.4 Joint State (Joint GRU)

The joint state, \( q_{s(u_t),t} \), of a party is the output of the 
emotional dynamics in a conversation is an inter-

Figure 1: In this dialogue, \( P_A \)’s emotion changes are influ-
enced by the behavior of \( P_B \).
Party State (Party GRU) DialogueRNN keeps track of the state of individual speakers using fixed size vectors $q_1, q_2, \ldots, q_M$ through out the conversation. These states are representative of the speakers’ state in the conversation, relevant to emotion classification. We update these states based on the current (at time $t$) role of a participant in the conversation, which is either speaker or listener, and the incoming utterance $u_t$. These state vectors are initialized with null vectors for all the participants. The main purpose of this module is to ensure that the model is aware of the speaker of each utterance and handle it accordingly.

Speaker Update (Speaker GRU): Speaker usually frames the response based on the context, which is the preceding utterances in the conversation. Hence, we capture context $c_t$ relevant to the utterance $u_t$ as follows:

$$\alpha = \text{softmax}(u_t^T W_{\alpha}(g_1, g_2, \ldots, g_{t-1})), \quad (2)$$

$$\text{softmax}(x) = \left[e^{x_1}/\sum_i e^{x_i}, e^{x_2}/\sum_i e^{x_i}, \ldots\right], \quad (3)$$

$$c_t = \alpha[g_1, g_2, \ldots, g_{t-1}]^T, \quad (4)$$

where $g_1, g_2, \ldots, g_{t-1}$ are preceding $t-1$ global states ($g_t \in \mathbb{R}^{D_g}$), $W_{\alpha} \in \mathbb{R}^{D_m \times D_g}$, $\alpha^T \in \mathbb{R}^{D_g}$, and $c_t \in \mathbb{R}^{D_g}$. In Eq. (2), we calculate attention scores $\alpha$ over the previous global states representative of the previous utterances. This assigns higher attention scores to the utterances emotionally relevant to $u_t$. Finally, in Eq. (4) the context vector $c_t$ is calculated by pooling the previous global states with $\alpha$.

Now, we employ a GRU cell $GRUP$ to update the current speaker state $q_{s(u_t), t-1}$ to the new state $q_{s(u_t), t}$ based on incoming utterance $u_t$ and context $c_t$ using GRU cell $GRUP$ of output size $D_P$

$$q_{s(u_t), t} = GRUP(q_{s(u_t), t-1}, (u_t \oplus c_t)), \quad (5)$$

where $W_{P,h}^{[r,z,c]} \in \mathbb{R}^{D_P \times D_P}$, $W_{P,x}^{[r,z,c]} \in \mathbb{R}^{D_P \times (D_m + D_g)}$, $b_{P}^{[r,z,c]} \in \mathbb{R}^{D_P}$, and $q_{s(u_t), t}, q_{s(u_t), t-1} \in \mathbb{R}^{D_p}$. This encodes the information on the current utterance along with its context from the global GRU into the speaker’s state $q_{s(u_t)}$, which helps in emotion classification down the line.

Listener Update: Listener state models the listeners’ change of state due to the speaker’s utterance. We tried two listener state update mechanisms:

- Simply keep the state of the listener unchanged, that is
  $$\forall i \neq s(u_t), q_{i,t} = q_{i,t-1}. \quad (6)$$

- Employ another GRU cell $GRUL$ to update the listener state based on listener visual cues (facial expression) $v_{i,t}$ and its context $c_t$, as
  $$\forall i \neq s(u_t), q_{i,t} = GRUL(q_{i,t-1}, (v_{i,t} \oplus c_t)), \quad (7)$$

where $v_{i,t} \in \mathbb{R}^{D_V}$, $W_{L,h}^{[r,z,c]} \in \mathbb{R}^{D_P \times D_P}$, $W_{L,x}^{[r,z,c]} \in \mathbb{R}^{D_P \times (D_V + D_g)}$, and $b_{L}^{[r,z,c]} \in \mathbb{R}^{D_P}$. Listener visual features of party $i$ at time $t$ $v_{i,t}$ are extracted using the model introduced by Arriaga, Valdenegro-Toro, and Pföger (2017), pretrained on FER2013 dataset, where feature size $D_V = 7$.

The simpler first approach turns out to be sufficient, since the second approach yields very similar result while increasing number of parameters. This is due to the fact that a listener becomes relevant to the conversation only when he/she speaks. In other words, a silent party has no influence in a conversation. Now, when a party speaks, we update his/her state $q_i$ with context $c_t$ which contains relevant information on all the preceding utterances, rendering explicit listener state update unnecessary. This is shown in Table 2.
Emotion Representation (Emotion GRU) We infer the emotionally relevant representation $e_t$ of utterance $u_t$ from the speaker’s state $q_{s(u_t),t}$ and the emotion representation of the previous utterance $e_{t-1}$. Since context is important to the emotion of the incoming utterance $u_t$, $e_{t-1}$ feeds fine-tuned emotionally relevant contextual information from other the party states $q_{s(e_{t-1}),t}$ into the emotion representation $e_t$. This establishes a connection between the speaker state and the other party states. Hence, we model $e_t$ with a GRU cell ($GRU_E$) with output size $D_E$ as

$$e_t = GRU_E(e_{t-1}, q_{s(u_t),t}),$$

(8)

where $D_E$ is the size of emotion representation vector, $e_{t,t-1} \in \mathbb{R}^{D_E}$, $W_{E,h}^{(r,z,c)} \in \mathbb{R}^{D_E \times D_E}$, $W_{E,x}^{(r,z,c)} \in \mathbb{R}^{D_E \times D_P}$, and $b_E^{(r,z,c)} \in \mathbb{R}^{D_E}$.

Since speaker state gets information from global states, which serve as speaker-specific utterance representation, one may claim that this way the model already has access to the information on other parties. However, as shown in the ablation study (Section 5.6) emotion GRU helps to improve the performance by directly linking states of preceding parties. Further, we believe that speaker and global GRUs ($GRU_p$, $GRU_G$) jointly act similar to an encoder, whereas emotion GRU serves as a decoder.

Emotion Classification We use a two-layer perceptron with a final softmax layer to calculate $c = 6$ emotion-class probabilities from emotion representation $e_t$ of utterance $u_t$ and then we pick the most likely emotion class:

$$l_t = ReLU(W_l e_t + b_l),$$

(9)

$$P_t = \text{softmax}(W_{smax} l_t + b_{smax}),$$

(10)

$$y_t = \text{argmax}(P_t[i]),$$

(11)

where $W_l \in \mathbb{R}^{D_l \times D_E}$, $b_l \in \mathbb{R}^{D_l}$, $W_{smax} \in \mathbb{R}^{c \times D_l}$, $b_{smax} \in \mathbb{R}^c$, $P_t \in \mathbb{R}^c$, and $y_t$ is the predicted label for utterance $u_t$.

Training We use categorical cross-entropy along with L2-regularization as the measure of loss ($L$) during training:

$$L = -\frac{1}{\sum_{s=1}^N \sum_{i=1}^c \log \frac{P_{i,s}(y_{i,s}) + \lambda \| \theta \|_2}{2}}.$$ (12)

where $N$ is the number of samples/dialogues, $c(i)$ is the number of utterances in sample $i$, $P_{i,j}$ is the probability distribution of emotion labels for utterance $j$ of dialogue $i$, $y_{i,j}$ is the expected class label of utterance $j$ of dialogue $i$, $\lambda$ is the L2-regularizer weight, and $\theta$ is the set of trainable parameters where $\theta = \{W_\alpha, W_\beta, W_{\gamma}, W_\delta, b_\alpha, b_\beta, b_\gamma, b_\delta, b_\epsilon\}.$

We used stochastic gradient descent based Adam (Kingma and Ba 2014) optimizer to train our network. Hyperparameters are optimized using grid search (values are added to the supplementary material).

3.4 DialogueRNN Variants

We use DialogueRNN (Section 3.3) as the basis for the following models:

**DialogueRNN + Listener State Update (DialogueRNN):** This variant updates the listener state based on the resulting speaker state $q_{s(u_t),t}$, as described in Eq. (7).

**Bidirectional DialogueRNN (BiDialogueRNN):** Bidirectional DialogueRNN is analogous to bidirectional RNNs, where two different RNNs are used for forward and backward passes of the input sequence. Outputs from the RNNs are concatenated in sequence level. Similarly, in BiDialogueRNN, the final emotion representation contains information from both past and future utterances in the dialogue through forward and backward DialogueRNNs respectively, which provides better context for emotion classification.

**DialogueRNN + attention (DialogueRNN+Att):** For each emotion representation $e_t$, attention is applied over all surrounding emotion representations in the dialogue by matching them with $e_t$ (Eqs. (13) and (14)). This provides context from the relevant (based on attention score) future and preceding utterances.

**Bidirectional DialogueRNN + Emotional attention (BiDialogueRNN+Att):** For each emotion representation $e_t$ of BiDialogueRNN, attention is applied over all the emotion representations in the dialogue to capture context from the other utterances in dialogue:

$$\beta_t = \text{softmax}(e_t W_\beta [e_1, e_2, \ldots, e_N]),$$

(13)

$$\tilde{e}_t = \beta_t [e_1, e_2, \ldots, e_N]^T,$$

(14)

where $e_t \in \mathbb{R}^{2D_E}$, $W_\beta \in \mathbb{R}^{2D_E \times 2D_E}$, $\tilde{e}_t \in \mathbb{R}^{2D_E}$, and $\beta_t \in \mathbb{R}^N$. Further, $\tilde{e}_t$ are fed to a two-layer perceptron for emotion classification, as in Eqs. (9) to (11).

4 Experimental Setting

4.1 Datasets Used

We use two emotion detection datasets IEMOCAP (Busso et al. 2008) and AVEC (Schuller et al. 2012) to evaluate DialogueRNN. We partition both datasets into train and test sets with roughly 80/20 ratio such that the partitions do not share any speaker. Table 1 shows the distribution of train and test samples for both dataset.

| Dataset          | Partition | Utterance Count | Dialogue Count |
|------------------|-----------|-----------------|----------------|
| IEMOCAP          | train + val | 5810             | 120            |
|                  | test      | 1623             | 31             |
| AVEC             | train + val | 4308             | 63             |
|                  | test      | 1430             | 32             |

Table 1: Dataset split ((train + val) / test ≈ 80%/20%).

**IEMOCAP:** IEMOCAP (Busso et al. 2008) dataset contains videos of two-way conversations of ten unique speakers, where only the first eight speakers from session one to four belong to the train-set. Each video contains a single dyadic dialogue, segmented into utterances. The utterances are annotated with one of six emotion labels, which are happy, sad, neutral, angry, excited, and frustrated.
4.2 Baselines and State of the Art

For a comprehensive evaluation of DialogueRNN, we compare our model with the following baseline methods:

**c-LSTM (Poria et al. 2017):** Bireational LSTM (Hochreiter and Schmidhuber 1997) is used to capture the context from the surrounding utterances to generate context-aware utterance representation. However, this model does not differentiate among the speakers.

**c-LSTM+Att (Poria et al. 2017):** In this variant attention is applied to the c-LSTM output at each timestamp by following Eqs. (13) and (14). This provides better context to the final utterance representation.

**TFN (Zadeh et al. 2017):** This is specific to multimodal scenario. Tensor outer product is used to capture intermodality and intra-modality interactions. This model does not capture context from surrounding utterances.

**MFN (Zadeh et al. 2018a):** Specific to multimodal scenario, this model utilizes multi-view learning by modeling view-specific and cross-view interactions. Similar to TFN, this model does not use contextual information.

**CNN (Kim 2014):** This is identical to our textual feature extractor network (Section 3.2) and it does not use contextual information from the surrounding utterances.

**Memnet (Sukhbaatar et al. 2015):** As described in Hazarika et al. (2018), the current utterance is fed to a memory network, where the memories correspond to preceding utterances. The output from the memory network is used as the final utterance representation for emotion classification.

**CMN (Hazarika et al. 2018):** This state-of-the-art method models utterance context from dialogue history using two distinct GRUs for two speakers. Finally, utterance representation is obtained by feeding the current utterance as query to two distinct memory networks for both speakers.

4.3 Modalities

We evaluated our model primarily on textual modality. However, to substantiate efficacy of our model in multimodal scenario, we also experimented with multimodal features.

5 Results and Discussion

We compare DialogueRNN and its variants with the baselines for textual data in Table 2. As expected, on average DialogueRNN outperforms all the baseline methods, including the state-of-the-art CMN, on both of the datasets.
### Table 2: Comparison with the baseline methods for trimodal (T+V+A) scenario. BiDialogueRNN+att\textsubscript{AF} = BiDialogueRNN+att in multimodal setting.

| Methods          | IEMOCAP     | AVEC       |
|------------------|-------------|------------|
|                  | Valence     | Arousal    | Expectancy | Power     |
|                  | (r)         | (r)        | (r)        | (r)       |
| TSN              | 36.8        | 0.03       | 0.10       | 0.12      | 0.12      |
| MFN              | 53.5        | 0.14       | 0.23       | 0.25      | -0.04     |
| c-LSTM           | 58.3        | 0.14       | 0.20       | 0.26      | -0.02     |
| CMN              | 58.5        | 0.23       | 0.30       | 0.26      | -0.02     |
| BiDialogueRNN\textsubscript{AF}         | 62.9        | 0.57       | 0.60       | 0.37      | 0.41      |
| BiDialogueRNN\textsubscript{AF}+att\textsubscript{AF} | 62.9        | 0.57       | 0.60       | 0.37      | 0.41      |

BiDialogueRNN. The improvement over BiDialogueRNN is 1.23% f1-score for IEMOCAP and consistently lower MAE and higher r in AVEC.

**BiDialogueRNN+Att:** Since this setting generates the final emotion representation by attending over the emotion representation from BiDialogueRNN, we expect better performance than both BiDialogueRNN and DialogueRNN+Att. This is confirmed in Table 2, where this setting performs the best in general than any other methods discussed on both datasets. This setting yields 6.62% higher f1-score on average than the state-of-the-art CMN and 2.86% higher f1-score than vanilla DialogueRNN for IEMOCAP dataset. For AVEC dataset also, this setting gives the best performance across all the four attributes.

### 5.3 Multimodal Setting

As both IEMOCAP and AVEC dataset contain multimodal information, we have evaluated DialogueRNN on multimodal features as used and provided by Hazarika et al. (2018). We use concatenation of the unimodal features as a fusion method by following Hazarika et al. (2018), since fusion mechanism is not a focus of this paper. Now, as we can see in Table 3, DialogueRNN significantly outperforms the strong baselines and state-of-the-art method CMN.

### 5.4 Case Studies

**Dependency on preceding utterances (DialogueRNN)**

One of the crucial components of DialogueRNN is its attention module over the outputs of global GRU (GRU\textsubscript{G}). Figure 3b shows the α attention vector (Eq. (2)) over the history of a given test utterance compared with the attention vector from the CMN model. The attention of our model is more focused compared with CMN: the latter gives diluted attention scores leading to misclassifications. We observe this trend of focused-attention across cases and posit that it can be interpreted as a confidence indicator. Further in this example, the test utterance by \( P_A \) (turn 44) comprises of a change in emotion from neutral to frustrated. DialogueRNN anticipates this correctly by attending to turn 41 and 42 that are spoken by \( P_A \) and \( P_B \), respectively. These two utterances provide self and inter-party influences that trigger the emotional shift. CMN, however, fails to capture such dependencies and wrongly predicts neutral emotion.

**Dependency on future utterances (BiDialogueRNN+Att)**

Fig. 3a visualizes the \( \beta \) (Eq. (13)) attention over the emotion representations \( e_t \) for a segment of a conversation between a couple. In the discussion, the woman (\( P_A \)) is initially at a neutral state, whereas the man (\( P_B \)) is angry throughout. The figure reveals that the emotional attention of the woman is localized to the duration of her neutral state (turns 1-16 approximately). For example, in the dialogue, turns 5, 7, and 8 strongly attend to turn 8. Interestingly, turn 5 attends to both past (turn 3) and future (turn 8) utterances. Similar trend across other utterances establish inter-dependence between emotional states of future and past utterances. The beneficial consideration of future utterances through GRU\textsubscript{G} is also apparent through turns 6, 9, 7, and 11. These utterances focus on the distant future (turn 49, 50) where the man is at an enraged state, thus capturing emotional correlations across time. Although, turn 6 is misclassified by our model, it still manages to infer a related emotional state (anger) against the correct state (frustrated). We analyze more of this trend in section 5.5.

**Dependency on distant context**

For all correct predictions in the IEMOCAP test set in Fig. 3d we summarize the distribution over the relative distance between test utterance and 2\textsuperscript{nd} highest attended utterance – either in the history or future – in the conversation. This reveals a decreasing trend with the highest dependence being within the local context. However, a significant portion of the test utterances (∼18%), attend to utterances that are 20 to 40 turns away from themselves, which highlights the important role of long-term emotional dependencies. Such cases primarily occur in conversations that maintain a specific affective tone, the local context. Although, a significant portion of test utterances are distant dependencies, they do not incur frequent emotional shifts.
Figure 3: (a) Illustration of the $\beta$ attention over emotion representations $e_t$; (b) Comparison of attention scores over utterance history of CMN and DialogueRNN ($\alpha$ attention). (c) An example of long-term dependency among utterances. (d) Histogram of $\Delta t = \text{distance between the target utterance and its context utterance based on } \beta$ attention scores.

As expected, following Table 4, party state stands very important, as without its presence the performance falls by 4.33%. We suspect that party state helps in extracting useful contextual information relevant to parties’ emotion.

Emotion GRU is also impactful, but less than party state, as its absence causes performance to fall by only 2.51%. We believe the reason to be the lack of context flow from the other parties’ states through the emotion representation of the preceding utterances.

Table 4: Ablated DialogueRNN for IEMOCAP dataset.

| Party State | Emotion GRU | F1   |
|-------------|-------------|------|
| -           | +           | 55.56|
| +           | -           | 57.38|
| +           | +           | 59.89|

6 Conclusion

We have presented an RNN-based neural architecture for emotion detection in a conversation. In contrast to the state-of-the-art method, CMN, our method treats each incoming utterance taking into account the characteristics of the speaker, which gives finer context to the utterance. Our model outperforms the current state-of-the-art on two distinct datasets in both textual and multimodal settings. Our method is designed to be scalable to multi-party setting with more than two speakers, which we plan to explore in future work.
References

Alm, C. O.; Roth, D.; and Sproat, R. 2005. Emotions from text: machine learning for text-based emotion prediction. In Proceedings of the conference on human language technology and empirical methods in natural language processing, 579–586. Association for Computational Linguistics.

Arriaga, O.; Valdenegro-Toro, M.; and Plöger, P. 2017. Real-time convolutional neural networks for emotion and gender classification. CoRR abs/1710.07557.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

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

Chen, M.; Wang, S.; Liang, P. P.; Baltrušaitis, T.; Zadeh, A.; and Morency, L.-P. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, 163–171. ACM.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. CoRR abs/1412.3555.

Datcu, D., and Rothkrantz, L. 2008. Semantic audio-visual data fusion for automatic emotion recognition. Euromedia'2008.

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

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, 1459–1462. ACM.

Graves, A.; Wayne, G.; and Danihelka, I. 2014. Neural turing machines. arXiv preprint arXiv:1410.5401.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018. 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), 2122–2132. New Orleans, Louisiana: Association for Computational Linguistics.

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

Kim, Y. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kingma, D. P., and Ba, J. 2014. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In International Conference on Machine Learning, 1378–1387.

McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; and Schroder, M. 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.

Picard, R. W. 2010. Affective computing: From laughter to tears. IEEE Transactions on Affective Computing 1(1):11–17.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 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), 873–883. Vancouver, Canada: Association for Computational Linguistics.

Richards, J. M.; Butler, E. A.; and Gross, J. J. 2003. Emotion regulation in romantic relationships: The cognitive consequences of concealing feelings. Journal of Social and Personal Relationships 20(5):599–620.

Rusuuvuori, J. 2013. Emotion, affect and conversation. The handbook of conversation analysis 330–349.

Schuller, B.; Valser, M.; Eyben, F.; Cowie, R.; and Pantic, M. 2012. AVEC 2012: The Continuous Audio/Visual Emotion Challenge. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12, 449–456. New York, NY, USA: ACM.

Strapparava, C., and Mihalcea, R. 2010. Annotating and identifying emotions in text. In Intelligent Information Access. Springer. 21–38.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-to-end Memory Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, 2440–2448. Cambridge, MA, USA: MIT Press.

Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; and Narayanan, S. S. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In INTERSPEECH 2010.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 1103–1114. Copenhagen, Denmark: Association for Computational Linguistics.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory Fusion Network for Multi-view Sequential Learning. In AAAI Conference on Artificial Intelligence, 5634–5641.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018b. Multi-attention recurrent network for human communication comprehension. In AAAI Conference on Artificial Intelligence, 5642–5649.