Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos

Devamanyu Hazarika
School of Computing,
National University of Singapore
hazarika@comp.nus.edu.sg

Soujanya Poria
Artificial Intelligence Initiative,
A*STAR, Singapore
sporia@ihpc.a-star.edu.sg

Amir Zadeh
Language Technologies
Institute, CMU, USA
abagherz@cs.cmu.edu

Erik Cambria
School of Computer Science and
Engineering, NTU, Singapore
cambria@ntu.edu.sg

Louis-Philippe Morency
Language Technologies
Institute, CMU, USA
morency@cs.cmu.edu

Roger Zimmermann
School of Computing,
National University of Singapore
rogerz@comp.nus.edu.sg

Abstract

Emotion recognition in conversations is crucial for the development of empathetic machines. Present methods mostly ignore the role of inter-speaker dependency relations while classifying emotions in conversations. In this paper, we address recognizing utterance-level emotions in dyadic conversational videos. We propose a deep neural framework, termed conversational memory network, which leverages contextual information from the conversation history. The framework takes a multimodal approach comprising audio, visual and textual features with gated recurrent units to model past utterances of each speaker into memories. Such memories are then merged using attention-based hops to capture inter-speaker dependencies. Experiments show an accuracy improvement of $3-4\%$ over the state of the art.

1 Introduction

Development of machines with emotional intelligence has been a long-standing goal of AI. With the increasing infusion of interactive systems in our lives, the need for empathetic machines with emotional understanding is paramount. Previous research in affective computing has looked at dialogues as an essential basis to learn emotional dynamics (Sidnell and Stivers, 2012; Poria et al., 2017a; Zhou et al., 2017).

Since the advent of Web 2.0, dialogue videos have proliferated across the internet through platforms like movies, webinars, and video chats. Emotion detection from such resources can benefit numerous fields like counseling (De Choudhury et al., 2013), public opinion mining (Cambria et al., 2017), financial forecasting (Xing et al., 2018), and intelligent systems such as smart homes and chatbots (Young et al., 2018).

In this paper, we analyze emotion detection in videos of dyadic conversations. A dyadic conversation is a form of a dialogue between two entities. We propose a conversational memory network (CMN), which uses a multimodal approach for emotion detection in utterances (a unit of speech bound by breathes or pauses) of such conversational videos.

Emotional dynamics in a conversation is known to be driven by two prime factors: self and inter-speaker emotional influence (Morris and Keltner, 2000; Liu and Maitlis, 2014). Self-influence relates to the concept of *emotional inertia*, i.e., the degree to which a person's feelings carry over from one moment to another (Koval and Kuppens, 2012). Inter-speaker emotional influence is another trait where the other person acts as an influencer in the speaker's emotional state. Conversely, speakers also tend to mirror emotions of their counterparts (Navarretta et al., 2016). Figure 1 provides an example from the dataset showing the presence of these two traits in a dialogue.

Existing works in the literature do not capitalize on these two factors. Context-free systems infer emotions based only on the current utterance in the conversation (Bertero et al., 2016). Whereas, state-of-the-art context-based networks like Poria et al., 2017b, use long short-term memory (LSTM) networks to model speaker-based context that suffers from incapability of long-range summarization and unweighted influence from context, leading to model bias.

Our proposed CMN incorporates these factors by using emotional context information present in the conversation history. It improves speaker-based emotion modeling by using memory networks which are efficient in capturing long-term

Proceedings of NAACL-HLT 2018, pages 2122–2132
New Orleans, Louisiana, June 1 - 6, 2018.

dependencies and summarizing task-specific details using attention models (Weston et al., 2014; Graves et al., 2014; Young et al., 2017).

Specifically, the memory cells of CMN are continuous vectors that store the context information found in the utterance histories. CMN also models interplay of these memories to capture inter-speaker dependencies.

CMN first extracts multimodal features (audio, visual, and text) for all utterances in a video. In order to detect the emotion of a particular utterance, say u_i, it gathers its histories by collecting previous utterances within a context window. Separate histories are created for both speakers. These histories are then modeled into memory cells using gated recurrent units (GRUs).

After that, CMN reads both the speaker's memories and employs attention mechanism on them, in order to find the most useful historical utterances to classify u_i. The memories are then merged with u_i using an addition operation weighted by the attention scores. This is done to model inter-speaker influences and dynamics. The whole cycle is repeated for multiple hops and finally, this merged representation of utterance u_i is used to classify its emotion category.

The contributions of this paper can be summarized as follows:

1. We propose an architecture, termed CMN, for emotion detection in a dyadic conversation that considers utterance histories of both the speaker to model emotional dynamics. The architecture is extensible to multi-speaker conversations in formats such as textual dialogues or conversational videos.

2. When applied to videos, we adopt a multimodal approach to extract diverse features from utterances. It also makes our model robust to missing information.

3. CMN provides a significant increase in accuracy of 3 – 4% over previous state-of-the-art networks. One variant called CMN$_{self}$ which does not consider the inter-speaker relation in emotion detection also outperforms the state of the art by a significant margin.

The remainder of the paper is organized as follows: Section 2 provides a brief literature review; Section 3 formalizes the problem statement; Section 4 describes the proposed method in detail; experimental results are covered in Section 5; finally, Section 6 provides concluding remarks.

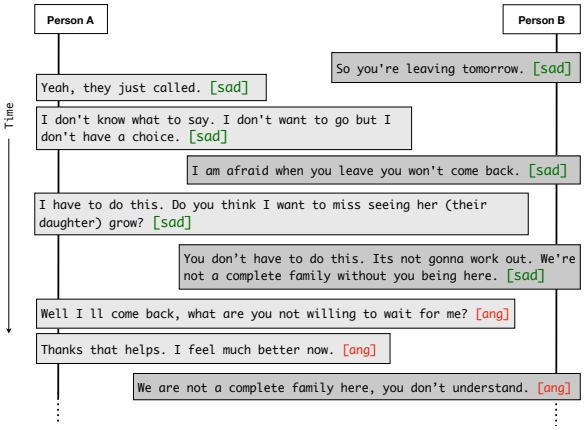


Figure 1: An abridged dialogue from the dataset. Person A (wife) is leaving B (husband) for a work assignment. Initially both A and B are emotionally driven by their own *emotional inertia*. In the end, *emotional influence* can be seen when B, despite being sad, reacts angrily to A's angry statement.

2 Related Works

Over the years, emotion recognition as an area of research has seen contributions from researchers across varied fields like signal processing, machine learning, cognitive and social psychology, natural language processing, etc. (Picard, 2010). Ekman, 1993, provided initial findings that related facial expressions as universal indicators of emotions. Datcu and Rothkrantz, 2008, 2011, showed the importance of acoustic cues in affect modeling.

A large section of researchers approaches emotion recognition from a multimodal learning perspective. Hence, many works used visual and audio features together for detecting affect (Busso et al., 2004; Castellano et al., 2008; Ranganathan et al., 2016). An in-depth review of the literature in these systems is provided by D'mello and Kory, 2015. Our work, which performs context-sensitive recognition (Wöllmer et al., 2010) uses three modalities: audio, visual and text. Recently, this combination of modalities has provided the best performance in affect recognition systems (Poria et al., 2017b; Wang et al., 2017; Tzirakis et al., 2017), thus motivating the use of a multimodal approach.

Previous works have focused on conversations as a resourceful event for emotion analysis. Rususvuori, 2013, provides an in-depth analysis on how emotions affect social interactions and conversations. In fact, significant works have attributed emotional dynamics as an interactive phe-

nomenon, rather than being within-person and one-directional (Richards et al., 2003; Hareli and Rafaeli, 2008). Such emotional dynamics are modeled by observing transition properties. Yang et al., 2011, study patterns for emotion transitions and show the evidence of emotional inertia. Xiaolan et al., 2013, use finite state machines to model transitions using stimuli and personality characteristics. Our work also tries to model emotional transitions using multimodal features. Unlike these works, however, we use memory networks to achieve the same.

The use of memory networks have been instrumental in the progress of multiple research problems, e.g., question-answering (Weston et al., 2014; Sukhbaatar et al., 2015; Kumar et al., 2016), machine translation (Bahdanau et al., 2014), speech recognition (Graves et al., 2014), and commonsense reasoning (Cambria et al., 2018). The repeated read and write to their memory cells is often coupled with attention modules, thus allowing it to filter only relevant memories.

Our model is loosely inspired from Sukhbaatar et al., 2015. Unlike their model, which directly encodes sentences into memories, we perform temporal sequence processing on our utterance histories using GRUs. We also extend their architecture to handle two speakers while keeping the possibility to add more. Finally, our model is different in the fact that we use multimodal features for input and processing.

3 Task Definition

Our goal is to infer the emotion of utterances present in a dyadic conversation. Let us define a dyadic conversation to be an asynchronous exchange of utterances between two persons P_a and P_b. Both the speakers speak a sequence of utterances U_a and U_b, respectively. Here, $U_\lambda = (s^1_\lambda, s^2_\lambda, ..., s^{l_\lambda}_\lambda)$ is ordered temporally, where s^i_λ is the i^{th} utterance by P_λ and l_λ is the total number of utterances spoken by person $P_\lambda,\ \lambda \in \{a, b\}$. Overall, the utterances by both speakers can be linearly ordered based on temporal occurrence as $(u_1, u_2, ...u_{l_a+l_b})$, where, $u_j \in U_a$ or U_b.

Our model takes as input an utterance u_i whose emotion category (Section 5.1) needs to be classified. To get its history, preceding K utterances of each person are separately collected as $hist_a$ and $hist_b$. Here, K serves as the length of the context window for history of u_i. Thus, for $\lambda \in \{a, b\}$:

$$hist_\lambda = \{u_j \mid u_j \in U_\lambda,\ j < i\}\ ,\ \mid hist_\lambda \mid \leq K \quad (1)$$

$hist_\lambda$ is also ordered temporally. At the beginning of the conversation, histories would have lesser than K utterances, i.e., $\mid hist_\lambda \mid < K$.

In the remaining sections, for brevity, we explain the processes using a subscript λ which can instantiate to either a or b, i.e., $\lambda \in \{a, b\}$.

4 Approach

We start by detailing the multimodal feature extraction scheme for all utterances followed by the mechanism to model emotional context using memory networks.

4.1 Multimodal Feature Extraction

The first phase of CMN is to extract multimodal features of all utterances in the conversations. The dyadic conversations are present in the form of videos. Each utterance of a particular conversation is thus a small segment of the full video. For each utterance, we extract features for the modes: audio, visual and text. The process of feature extraction for each mode is described below.

4.1.1 Textual Features Extraction

We extract features from the transcript of an utterance video using convolutional neural networks (CNNs). CNNs are effective in learning high level abstract representations of sentences from constituting words or n-grams (Kalchbrenner et al., 2014). To get our sentence representation, we use a simple CNN with one convolutional layer followed by max-pooling (Kim, 2014; Poria et al., 2016).

Specifically, the convolution layer consists filters of sizes $3, 4$ and 5 with 50 feature maps each. Max-pooling is employed on these feature maps with a pooling window of size 2. Finally, a fully connected layer is used with 100 neurons. The activations of this layer form our sentence representation t_u.

4.1.2 Audio Feature Extraction

To extract audio features we use openSMILE (Eyben et al., 2010). It is an open-source software which provides high dimensional audio vectors. These vectors comprise of features like loudness, Mel-spectra, MFCC, pitch, etc. Audio features play a significant role in providing information on the emotional state of a speaker (Song et al., 2004).

In fact, the literature shows that there exists a high correlation between many statistical measures of speech with speakers' emotion. For example, high pitch and fast speaking rate often denote anger while sadness associates low standard deviation of pitch and slow speech rate (Dellaert et al., 1996; Amir, 1998). In this work, we use the *IS13_ComParE*[1] config file which extracts a total of 6373 features for each utterance video. Z-standardization is performed for voice normalization and dimension of the audio vector is reduced to 100 using a fully-connected neural layer. This provides the final audio feature vector a_u.

4.1.3 Visual Feature Extraction

Facial expressions and visual surrounding provide rich emotional indicators. We use a 3D-CNN to capture these details from the utterance video. Apart from the benefits of extracting relevant features from each image frame, 3D-CNN also extracts spatiotemporal features across frames (Tran et al., 2015). This leads to the identification of emotional expressions like a smile or frown.

The working of a 3D-CNN is identical to its 2D counterpart with an input being a video v of dimension: $(3, f, h, w)$. Here, 3 represents the RGB channels and f, h, w are the number of frames, height, and width of each frame, respectively. For the convolution operation, a 3D filter f_l of dimension $(f_m, 3, f_d, f_h, f_w)$ is used where, $f_{[m/d/h/f]}$ represents number of feature maps, depth, height and width of the filter, respectively. Max-pooling is applied to the output of this convolution across a 3D sliding window of dimension (m_p, m_p, m_p).

In our model, we use 128 feature maps for 3D filters of size 5. For pooling, we set m_p to be 3 whose output is fed to a fully connected layer with 100 neurons. All the values are decided using hyperparameter tuning (see Section 5). For the input utterance, the activations of this layer form the video representation v_u.

Fusion: We perform feature level fusion to map the individual modalities to a joint space. This is done through a simple feature concatenation. Thus, the extracted features t_u, a_u and v_u are joined to form the utterance representation $u = [t_u; a_u; v_u]$ of dimension $d_{in} = 300$. This multimodal representation is generated for all utterances in a conversation.

[1] http://audeering.com/technology/opensmile

Literature consists of numerous fusion techniques for multimodal data (Atrey et al., 2010; Zadeh et al., 2017; Poria et al., 2017c). Exploring these on CMN, however, is beyond the scope of this paper and left as a future work.

4.2 Conversational Memory Network

For classifying the emotion of an utterance u_i, its corresponding histories ($hist_a$ and $hist_b$) are taken. Each history $hist_\lambda$ contains the preceding K utterances by person P_λ (see Section 3). Here, both u_i and utterances in the histories are represented using their multimodal feature vectors of dimension $\mathcal{R}^{d_{in}}$ (Figure 2).

The histories are first modeled into memory cells using GRUs. This provides the memories with context information summarized by the GRU. We call this step as memory representation. Following cognitive evidence of self-emotional dynamics, we model separate memory cells for each person. Thus, identical but separate computations are performed on both histories. From these memories, content relevant to utterance u_i is then filtered out using attention mechanism over multiple input/output hops. At each hop, both memories are accumulated and merged with u_i to model inter-speaker emotional dynamics. First, we describe our model as a single layer memory network which runs one hop operation on the memories.

4.2.1 Single Layer

Here, we explain the representation scheme of the memories for both histories and the input/output operations on them along with attention mechanism. The memory representation for each history is generated using a GRU for modeling emotion transitions. First, we define the GRU cell.

Gated Recurrent Unit: GRUs are a gating mechanism in recurrent neural networks introduced by (Cho et al., 2014). Similar to an LSTM (Hochreiter and Schmidhuber, 1997), GRU provides a simpler computation with similar performance. At any timestep t, it utilizes two gates r_t (*reset gate*) and z_t (*update gate*) to control the combination criteria with current input utterance u_t and previous hidden state s_{t-1}.

The new state s_t is computed as:

$$z_t = \sigma(V^z.u_t + W^z.s_{t-1} + b^z) \quad (2)$$

$$r_t = \sigma(V^r.u_t + W^r.s_{t-1} + b^r) \quad (3)$$

$$h_t = tanh(V^h.u_t + W^h.(s_{t-1} \otimes r_t) + b^h) \quad (4)$$

$$s_t = (1 - z_t) \otimes h_t + z_t \otimes s_{t-1} \quad (5)$$

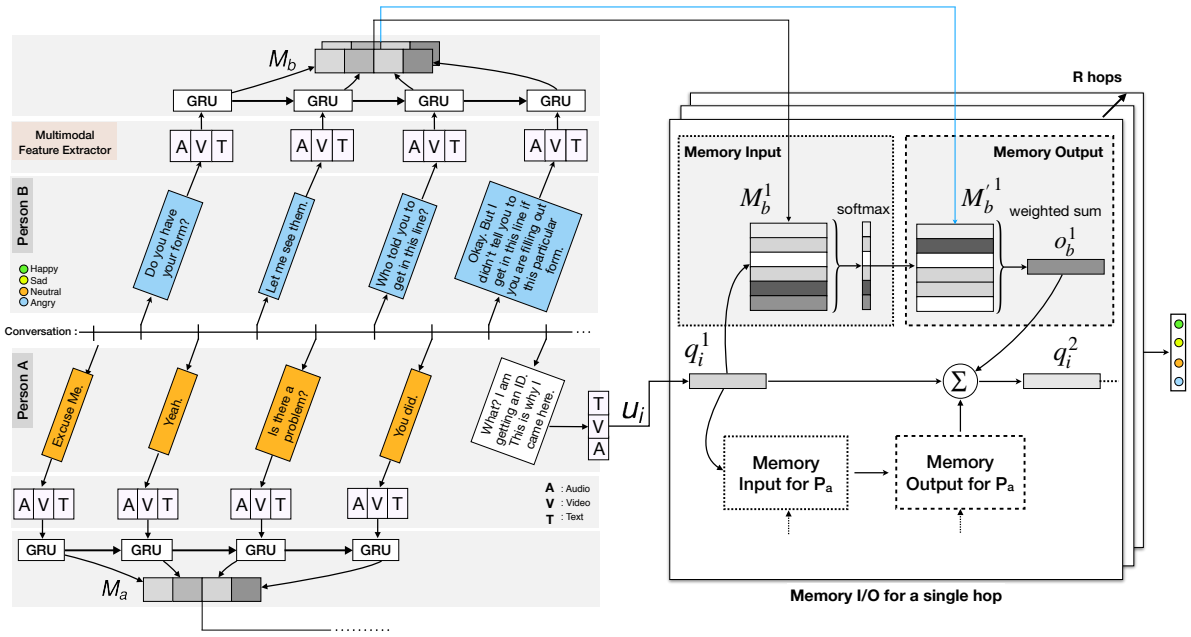


Figure 2: Overall architecture of proposed model: CMN. First, multimodal representations are extracted for each utterance and then the previous $K = 4$ utterances by both persons are used to model the histories using GRUs. For each person, $R+1$ different GRUs are used to represent $M_\lambda^{(r)}$ for all R hops. Then, attention based filtering using multiple memory hops is performed. Finally, Person A's utterance u_i is classified to predict its emotion category.

Here, V, W and b are parameter matrices and vector and $\otimes$ represents element-wise multiplication. The above equations can be summarized as: $s_t = GRU_\lambda(s_{t-1}, u_t)$.

Memory Representation: For each $\lambda \in \{a, b\}$, a memory representation $M_\lambda = [m_\lambda^1, ..., m_\lambda^K]$ for $hist_\lambda$ is generated using a GRU. To grasp the temporal context, the K utterances in $hist_\lambda$ are framed as a sequence (starting from the oldest one) and fed to the GRU$_\lambda$. At each timestep $t \in [1, K]$, the GRU$_\lambda$'s internal state s_t (equation 5) forms the t^{th} memory cell m_λ^t of memory representation M_λ.

Memory Input: This step takes the memory representation M_λ and performs an attention mechanism on it, resulting in an attention vector $p_\lambda \in \mathcal{R}^K$. First, the current utterance u_i is embedded into a vector q_i of dimension $\mathcal{R}^d$ using a projection matrix $B \in \mathcal{R}^{d \times d_{in}}$. To find the relevance of each memory m_λ^t's context with q_i, a match between both is computed.
We do this by taking an inner product as follows:

$$q_i = B.u_i \tag{6}$$

$$p_\lambda^t = softmax(q_i^T.m_\lambda^t) \tag{7}$$

Here, $softmax(x_i) = e^{x_i} / \sum_j e^{x_j}$ and attention vector $p_\lambda = \{p_\lambda^t\}$ is a probability distribution over the input memories $M_\lambda = \{m_\lambda^t\}$ for $t \in [1, K]$.

Memory Output: First a new set of memories are created using another GRU$_\lambda^{'}$ to get new memory representation $M_\lambda^{'} = \{(m_\lambda^t)^{'}\}$. An output representation $o_\lambda \in \mathcal{R}^d$ is then generated using the weighted sum of attention vector p_λ and new memory $M_\lambda^{'}$ as follows:

$$o_\lambda = \sum_t p_\lambda^t.(m_\lambda^t)^{'} = M_\lambda^{'}.p_\lambda \tag{8}$$

Thus, the output representation o_λ contains weighted contextual summary accumulated from the memory.

Final Prediction: To generate the predictions for the current utterance u_i, we combine the output representations of both persons: o_a and o_b with u_i's representation q_i and perform an affine transformation using matrix W_o. Softmax is applied to this final vector to get the emotion predictions,

$$\hat{y} = softmax(W_o.(q_i + o_a + o_b)) \tag{9}$$

Categorical cross-entropy is used as the loss:

$$Loss = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{i,j} \; log_2(\hat{y}_{i,j}) \tag{10}$$

Here, N denotes total utterances across all videos and C is the number of emotion categories. y_i is the one-hot vector ground truth of i^{th} utterance from the training set and $\hat{y}_{i,j}$ is its predicted probability of belonging to class j.

4.2.2 Multiple Layers

Many recent works on memory networks adopt a multiple hop scheme in their network. This repeated input and output cycle on the memories along with a soft attention module, leads to a refined representation of the memories (Sukhbaatar et al., 2015; Kumar et al., 2016). Motivated by these works, we extend our model to perform R hops on the memories. This is done by stacking the single hop layers (Section 4.2.1) as follows:

- At a particular hop r, the output memory of the previous hop $M_\lambda^{'(r-1)}$ is used as the input memory of the current hop $M_\lambda^{(r)}$. Output memory of current r^{th} hop is generated using a new $GRU_\lambda^{(r)}$. This constraint of sharing parameters adjacently between layers is added for reduction in total parameters and ease of training.
- At every hop, the query utterance u_i's representation q_i is updated as:

$$q_i^{(r+1)} = q_i^{(r)} + o_a^{(r)} + o_b^{(r)} \quad (11)$$

  $o_\lambda^{(r)}$ is calculated as per equation 8 using $M_\lambda^{'(r)}$.
- After R hops, the final prediction is done using equation 9 as: $\hat{y} = softmax(W_o.(q_i^{(R+1)}))$. Algorithm 1 summarizes the overall CMN network.

Algorithm 1 Conversational Memory Network

procedure CMN($u_i, hist_a, hist_b, K, R$) ▷ predict the emotion of u_i
  $q_i^{(1)} \leftarrow B.u_i$
  $M_\lambda^{'(0)} \leftarrow GRU_\lambda^{(0)}(hist_\lambda)$
  for `r in [1,R]` **do** ▷ Multi-hop memory I/O
    $M_\lambda^{(r)} \leftarrow M_\lambda^{'(r-1)}$
    $M_\lambda^{'(r)} \leftarrow GRU_\lambda^{(r)}(hist_\lambda)$
    $p_\lambda \leftarrow softmax(\ (q_i^r)^T.M_\lambda^{(r)}\ )$ ▷ Memory in
    $o_\lambda^{(r)} \leftarrow M_\lambda^{'(r)}.p_\lambda$ ▷ Memory out
    $q_i^{(r+1)} \leftarrow q_i^{(r)} + o_a^{(r)} + o_b^{(r)}$ ▷ Query update
  return $\hat{y} \leftarrow softmax(W_o.(q_i^{(R+1)}))$ ▷ Prediction

5 Experiments

5.1 Dataset

We perform experiments on the IEMOCAP dataset[2] (Busso et al., 2008). It is a multimodal database of 10 speakers (5 male and 5 female) involved in two-way dyadic conversations. A pair of speakers is given multiple conversation scenarios which are grouped in a single session. All the conversations are segmented into utterances. Each utterance is annotated using the following emotion categories: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. However, in our experiments, we consider the first four categories. This is done to compare our method with state-of-the-art frameworks (Poria et al., 2017b; Rozgic et al., 2012). The dataset provides rich video and audio samples for all the utterances along with transcriptions.

Apart from these emotional states, we also investigate the valence and arousal degrees of each utterance. IEMOCAP provides labels for both these attributes on a 5-point Likert scale. Following Aldeneh et al., 2017, we convert the attributes into 3 categories, namely, *low* ($\leq$ 2), *medium* ($>$ 2 and $<$ 4) and *high* ($\geq$ 4). The dataset configuration for the experiments is obtained from Poria et al. (2017b). The first 8 speakers (Session 1 - 4) compose the training fold while the last session is used as the testing fold. Overall, the training and testing set comprises of 4290 utterances (120 conversational videos) and 1208 utterances (31 conversational videos), respectively. There is no speaker overlap in the training and testing set to make the model person-independent.

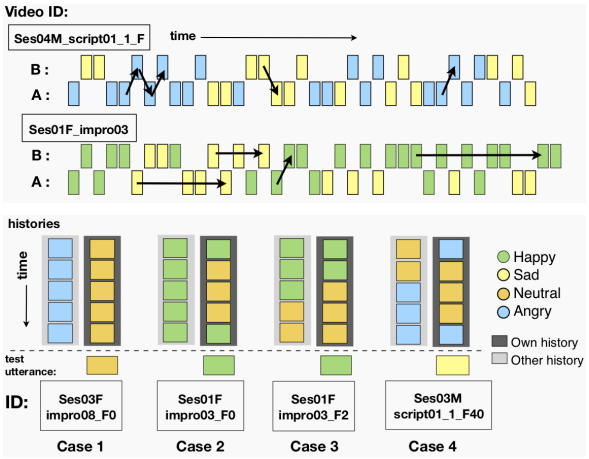


Figure 3: Each block represents an utterance and the blocks are ordered as per temporal occurrence. Color scheme identifies their corresponding emotions. The arrows denote emotional influence directions.

5.2 Emotional Influence Patterns

In this section, we perform dataset exploration to check the existence of emotional influences. Figure 3a) presents the emotion sequence of two videos

[2] http://sail.usc.edu/iemocap/

sampled from the dataset. Both videos show the presence of self and inter-speaker emotional influences. Visual exploration of videos from the dataset reveal significant existence of such instances in the conversations. To provide quantitative evidence of the emotional influence patterns, we curate a non exhaustive list of possible cases of influence. For all utterances in the dataset, we sample their histories by setting $K = 5$, i.e., five previous utterances (as per availability) from both speakers.

Cases 1 and 2 (Figure 3) represent scenarios when the emotion of current utterance is influenced by self or the other person respectively. In case 3, the utterance has relevant content in the histories that do not precede immediately. An effective attention mechanism provides the capability to capture this pattern. Finally case 4 presents the situation when the utterance is independent of the history. Such situations are indicative from the content of the utterance which often deviates from the previous topic of discussion or introduces a new information. Table 1 presents a statistical summary of these cases present in the dataset. From the table it can be seen that a large section of the dataset demonstrate these influence patterns. This provides motivation to explicitly model these patterns. We thus hypothesize that models that are able to capture these cases would have superior emotion inference capabilities.

This passive exploration is a label-based analysis which is performed as a sanity check. Needless to say, existence of some false positive patterns at the label level is imminent. On the other hand, our model CMN is content-based which enables it to mine intricate patterns from the utterance histories.

5.2.1 Training Details

We use 10% of the training set as a held-out validation set for hyperparameter tuning. To optimize the parameters, we use Stochastic Gradient Descent (SGD) optimizer, starting with an initial learning rate (lr) of 0.01. An annealing approach halves the lr every 20 epochs and termination is decided using an early-stop measure with a patience of 12 by monitoring the validation loss. Gradient clipping is used for regularization with a norm set to 40. Hyperparameters are decided using a Random Search (Bergstra and Bengio, 2012). Based on validation performance, context window length K is set to be 40 and the number of hops R is fixed at 3 hops. If K previous utterances are unavailable, then null utterances are added at the beginning of the history sequence. The dimension size of the memory cells d is set as 50.

	Case 1	*Case 2*	*Case 3*	*Case 4*
Percentage	63.77	40.44	30.97	16.24

Table 1: Percentage of occurrence of different cases in the dataset as mentioned in Section 5.2. All cases are analyzed with $K = 5$. Utterances whose history has atleast 3 similar emotion labels in either own history or the history of the other person, is counted in case 1 or 2, respectively. Case 3 is considered when the utterance's emotion is found in atleast 3 utterances which occur before the second past-utterance of each history. Case 4 is considered when no history has the emotion label of the current utterance.

5.2.2 Baselines

We compare CMN with the following baselines:

SVM-ensemble: A strong context-free benchmark model which uses similar multimodal approach on an ensemble of trees. Each node represents binary support vector machines (SVM) (Rozgic et al., 2012).

bc-LSTM: A bi-directional LSTM equipped with hierarchical fusion, proposed by Poria et al., 2017b. It is the present state-of-the-art method. The model uses context features from unimodal LSTMs and its concatenation is fed to a final LSTM for classification. For fair comparison in an end-to-end learning paradigm, we remove the penultimate SVM of this model. The model doesn't accommodate inter-speaker dependencies.

Memn2n: The original memory network as proposed by Sukhbaatar et al., 2015. Contrasting to CMN, the model generates the memory representations for each historical utterance using an embedding matrix B as used in equation 7, without sequential modeling. Thus for utterance u_i, both memories are created as M_λ using $\{m^t_\lambda = B.u_t \mid u_t \in hist_\lambda \text{ and } t \in [1, K]\}$ for $\lambda \in \{a, b\}$.

CMN$_{Self}$: In this baseline, we use only self history for classifying emotion of utterance u_i. Thus, if u_i is spoken by person P_a, then only $hist_a$ is considered. Clearly, this variant is also incapable of modeling inter-speaker dependencies.

CMN$_{NA}$: Single layer variant of the CMN with no attention module. Thus, its output o_λ (equation 8) is generated using a uniform probability distribution p_λ, i.e., $\{p^t_\lambda = \frac{1}{K}\}^K_{t=1}$.

5.3 Results

Table 2 presents the performances of CMN and its variants along with the state-of-the-art mod-

Models	hops	history	Emotion Categories					Valence		Arousal	
			Happiness	*Sadness*	*Neutral*	*Anger*	*WAA*	*WAA*	*UAR*	*WAA*	*UAR*
SVM-ensemble[1]	-	single	72.40	61.90	58.10	73.10	69.50	-	-	-	-
bc-LSTM[2]	-	single	74.21	76.50	66.31	75.68	74.31	64.3	62.3	70.1	45.0
Memn2n	1	dual	72.36	76.16	66.93	80.23	74.17	-	-	-	-
	3	dual	75.03	76.36	66.45	81.59	75.08	65.3	64.0	71.5	45.6
CMN_{Self}	3	single	77.14†	76.99	66.99	87.26†	76.54†	65.5	64.0	72.1	47.1
CMN_{NA}	1	dual	74.33	76.93	66.49	86.29†	75.77	65.6	64.2	71.6	46.3
CMN	3	dual	**81.75**†	77.73	67.32	**89.88**†	**77.62**†	**66.1**	**64.3**	**72.2**	**47.6**

[1](Rozgic et al., 2012), [2](Poria et al., 2017b). †: significantly better than bc-LSTM[1]

Table 2: Comparison of CMN and its variants with state-of-the-art models (Section 5.2.2). All results use multimodal features. We report scores using weighted accuracy (WAA) and unweighted recall (UAR). UAR is a popular metric that is used when dealing with imbalanced classes (Rosenberg, 2012). Results are an average of 10 runs with varied weight initializations. We assert significance when $p < 0.05$ under McNemar's test.

els. CMN succeeds over both neural (Poria et al., 2017b) and SVM-based (Rozgic et al., 2012) methods by 3.3% and 8.12%, respectively. Improvement in performance is seen for all emotions over the ensemble-SVM based method. A similar trend is seen with bc-LSTM (Poria et al., 2017b), where our model does explicitly well for the active emotions *happiness* and *anger*. This trend suggests that CMN is capable of capturing inter-speaker emotional influences which are often seen in the presence of such active emotions.

The importance of sequential processing of the histories using a recurrent neural network (in our case, a GRU) is evidenced by the poorer performance of Memn2n with respect to CMN. This suggests that gathering contexts temporally through sequential processing is indeed a superior method over non-temporal memory representations. CMN_{self} which uses only single history channel also provides lesser performance when compared to CMN. This signifies the role of inter-speaker influences that often moderate the emotions of the current utterance. Overall, predictions on valence and arousal levels also show similar results which reinforce our hypothesis of CMN's ability to model emotional dynamics.

Hyperparameters: Figure 4 provides a summary of the performance trend of our model for different values of the hyperparameters K (context window length) and Q (number of hops). In the first graph, as K increases, more past-utterances are provided to the model as memories. The performance maintains a positive correlation with K. This trend supplements our intuition that the historical context acts as an essential resource to model emotional dynamics. Given enough history, the performance saturates. The second graph shows that multiple hops on the histories indeed lead to an improvement in performance. The attention-based filtering in each hop provides a refined context representation of the histories. Models with hops in the range of 3 – 10 outperform the single layer variant. However, each added hop contributes a new set of parameters for memory representation, leading to an increase in total parameters of the model and making it susceptible to overfitting. This effect is evidenced in the figure where higher hops lead to a dip in performance.

Multimodality: Table 3 summarizes the performance of unimodal and multimodal variants of the baselines along with CMN. As seen in the table, text modality performes best out of the three. This is in contrast to Rozgic et al. 2012 where audio provides the best performance. A possible reason for

Models	*unimodal audio*	*unimodal visual*	*unimodal text*	*trimodal*
SVM-ensemble	60.8	51.5	48.5	69.5‡
bc-LSTM	62.2	56.1	72.5	74.3‡
Memn2n	63.0	61.8	72.6	75.0‡
CMN_{self}	63.1	62.5	73.0	76.5‡
CMN_{NA}	62.4	60.9	74.1	75.7
CMN	**65.3**	**64.2**	**74.2**	**77.6**‡

‡: significantly better than unimodals ($p < 0.05$)

Table 3: Comparison of CMN to all the baselines in different modalities. Weighted accuracy is used as the metric.

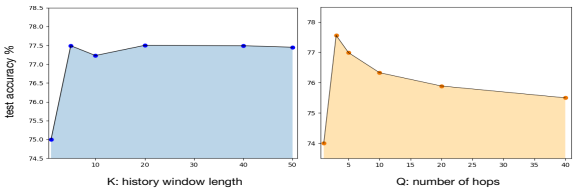
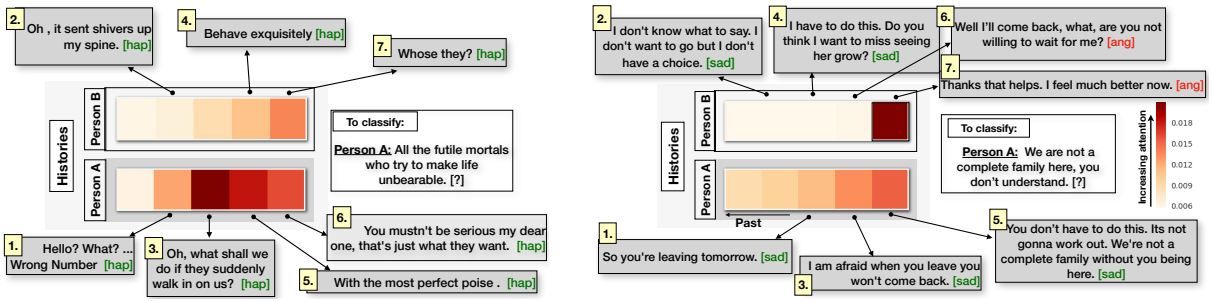


Figure 4: Performance trends of our model with different values of K (history length) and Q (number of hops). While K is varied, Q is set to be 3. Similarly, $K = 20$ when Q varies.



(a) Correct label: *happiness* (b) Correct label: *anger*

Figure 5: Average attention vectors across 3 hops for both memories for a given test utterance.

this shift is the improved representational scheme of the textual modality. Text tends to have lesser noisy signals as opposed to audio-visual sources, thus providing better features in the joint representation. Overall, multimodal systems outperform the unimodal variants justifying the design of CMN as a multimodal system.

Table 3 also showcases the superiority of CMN and its variants over bc-LSTM. The proposed model achieves better performance over the state of the art in all the unimodal and multimodal segments. This asserts the importance of the memory-network framework and its ability to effectively store context information.

Role of Attention: Attention module plays a vital role in memory refinement. This is also observed in Table 2, where CMN$_{NA}$ provides inferior performance over CMN. With the uniform weight, all the memory cells in both memories M_a and M_b equally contribute to the output representation. This incorporates irrelevant information from the perspective of emotional context.

Case Study: We perform qualitative visualization of the attention module by applying it on the testing set. Figure 5a represents a conversation where both the speakers are in an excited and jolly mood. Person A, in particular, drives the dialogue with less influence from Person B. To classify the test utterance of A, the attention module of CMN successfully focuses on the utterances $1, 3, 5$ which had triggered the speaker's positive mood in the video. This shows CMN's capacity to model speaker-based emotions. Also, at the textual level, utterances 3 and 6 do not seem to depict a happy mood. However, audio and visual sources provide contrasting evidence which helps CMN to correctly model them as utterances spoken with happiness. This shows the advantage of a multimodal system.

In Figure 5b we reiterate through the dialogue presented in Figure 1. As shown, Person A converses in a sad mood (utterances $1, 3, 5$ in Fig 5b), bounded by the grief of his wife's departure. But when he expresses his inhibitions, his wife B reacts in an angry and sarcastic manner (utterance 7). This ignites an emotional shift for A who then replies angrily. In this example, CMN is able to focus on utterance 7 spoken by B to anticipate A's test utterance to be an angry statement, thus showing its ability to model inter-speaker influences. However, there are cases where our model fails, e.g., in the absence of historical utterances as this forces attention to focus on null memories.

6 Conclusion

In this paper, we presented a deep neural framework that identifies emotions for utterances in dyadic conversational videos. Our results suggest that leveraging context information from utterance histories and representing them as memories indeed helps to better recognize emotions. Performing speaker-specific modeling and considering inter-speaker influences also helps in capturing emotional dynamics.

This work also showed the importance of attention mechanism in filtering relevant contextual information from utterance histories and, hence, paved the path to the development of more efficient and human-like dialogue systems.

Acknowledgement

This research was supported in part by the National Natural Science Foundation of China under Grant no. 61472266 and by the National University of Singapore (Suzhou) Research Institute, 377 Lin Quan Street, Suzhou Industrial Park, Jiang Su, People's Republic of China, 215123.

References

Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Pooling acoustic and lexical features for the prediction of valence.

Noam Amir. 1998. Towards an automatic classification of emotions in speech. *ICSLP*, pages 699–702.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *EMNLP*, pages 1042–1047.

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.

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *ICMI*, pages 205–211.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *AAAI*.

Ginevra Castellano, Loic Kessous, and George Caridakis. 2008. Emotion recognition through multiple modalities: face, body gesture, speech. *Affect and emotion in human-computer interaction*, pages 92–103.

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*.

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

Dragoş Datcu and Léon JM Rothkrantz. 2011. Emotion recognition using bimodal data fusion. In *International Conference on Computer Systems and Technologies*, pages 122–128. ACM.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Frank Dellaert, Thomas Polzin, and Alex Waibel. 1996. Recognizing emotion in speech. In *ICSLP*, volume 3, pages 1970–1973.

Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43.

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 *International Conference on Multimedia*, pages 1459–1462. ACM.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Shlomo Hareli and Anat Rafaeli. 2008. Emotion cycles: On the social influence of emotion in organizations. *Research in organizational behavior*, 28:35–59.

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

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL 2014*, volume 1, pages 655–665.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP 2014*, pages 1746–1751.

Peter Koval and Peter Kuppens. 2012. Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion*, 12(2):256.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, pages 1378–1387.

Feng Liu and Sally Maitlis. 2014. Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings. *Journal of Management Studies*, 51(2):202–234.

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.

Costanza Navarretta, K Choukri, T Declerck, S Goggi, M Grobelnik, and B Maegaard. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *LREC*.

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

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *ACL 2017*, volume 1, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017c. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *ICDM 2017*, pages 1033–1038. IEEE.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *COLING 2016*, pages 1601–1612.

Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *WACV*, pages 1–9. IEEE.

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

Andrew Rosenberg. 2012. Classifying skewed data: Importance weighting to optimize average recall. In *INTERSPEECH 2012*.

Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *APSIPA ASC*, pages 1–4.

Johanna Ruusuvuori. 2013. Emotion, affect and conversation. *The handbook of conversation analysis*, pages 330–349.

Jack Sidnell and Tanya Stivers. 2012. *The handbook of conversation analysis*, volume 121. John Wiley & Sons.

Mingli Song, Jiajun Bu, Chun Chen, and Nan Li. 2004. Audio-visual based emotion recognition-a new approach. In *CVPR*, volume 2.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *NIPS 2015*, pages 2440–2448.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV 2015*, pages 4489–4497.

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 JSTSP*, 11(8):1301–1309.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *ICME*, pages 949–954.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth S Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *INTERSPEECH 2010*.

Peng Xiaolan, Xie Lun, Liu Xin, and Wang Zhiliang. 2013. Emotional state transition model based on stimulus and personality characteristics. *China Communications*, 10(6):146–155.

Frank Xing, Erik Cambria, and Roy Welsch. 2018. Natural language based financial forecasting: A survey. *Artificial Intelligence Review*.

Liang Yang, Hong-fei LIN, and Wei GUO. 2011. Text-based emotion transformation analysis. *Computer Engineering & Science*, 9:026.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialog systems with commonsense knowledge. In *AAAI*.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2017. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP 2017*, pages 1103–1114.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv:1704.01074*.