# ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection

**Devamanyu Hazarika**[1], **Soujanya Poria**[2] , **Rada Mihalcea**[3],
**Erik Cambria**[2] and **Roger Zimmermann**[1]
[1] School of Computing, National University of Singapore
[2] School of Computer Science and Engineering, NTU, Singapore
[3] Computer Science & Engineering, University of Michigan, Ann Arbor
{hazarika|rogerz}@comp.nus.edu.sg, {sporia|cambria}@ntu.edu.sg,
mihalcea@umich.edu

## Abstract

Emotion recognition in conversations is crucial for building empathetic machines. Current work in this domain do not explicitly consider the inter-personal influences that thrive in the emotional dynamics of dialogues. To this end, we propose Interactive COnversational memory Network (ICON), a multimodal emotion detection framework that extracts multimodal features from conversational videos and hierarchically models the *self*- and *inter-speaker* emotional influences into global memories. Such memories generate contextual summaries which aid in predicting the emotional orientation of utterance-videos. Our model outperforms state-of-the-art networks on multiple classification and regression tasks in two benchmark datasets.

## 1 Introduction

Emotions play an important role in our daily life. A long-standing goal of AI has been to create affective agents that can detect and comprehend emotions. Research in affective computing has mainly focused on understanding affect (emotions and sentiment) in monologues. However, with increasing interactions of humans with machines, researchers now aim at building agents that can seamlessly analyze affective content in conversations. This can help in creating empathetic dialogue systems, thus improving the overall human-computer interaction experience (Young et al., 2018).

Analyzing emotional dynamics in conversations, however, poses complex challenges. This is due to the presence of intricate dependencies between the affective states of speakers participating in the dialogue. In this paper, we address the problem of emotion recognition in conversational videos. We specifically focus on dyadic conversations where two entities participate in a dialogue.

We propose Interactive COnversational memory Network (ICON), a multimodal network for identifying emotions in utterance-videos. Here, utterances are units of speech bounded by breaths or pauses of the speaker. Emotional dynamics in conversations consist of two important properties: *self* and *inter-personal dependencies* (Morris and Keltner, 2000). Self-dependencies, also known as *emotional inertia*, deal with the aspect of emotional influence that speakers have on themselves during conversations (Kuppens et al., 2010). On the other hand, inter-personal dependencies relate to the emotional influences that the counterparts induce into a speaker. Conversely, during the course of a dialogue, speakers also tend to mirror their counterparts to build rapport (Navarretta et al., 2016).

Figure 1 demonstrates a sample conversation from the dataset involving both self and inter-personal dependencies. While most conversational frameworks only focus on self dependencies, ICON leverages both such dependencies to generate affective summaries of conversations. First, it extracts multimodal features from all utterance-videos. Next, given a test utterance to be classified, ICON considers the preceding utterances of both speakers falling within a context-window and models their self-emotional influences using local gated recurrent units (GRUs).

Furthermore, to incorporate inter-speaker influences, a global representation is generated using a GRU that intakes output of the local GRUs. For each instance in the context-window, the output of this global GRU is stored as a memory cell. These memories are then subjected to multiple read/write cycles that include attention mechanism for generating contextual summaries of the conversational history. At each iteration, the representation of the test utterance is improved with this summary representation and finally used for prediction.


*Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604
Brussels, Belgium, October 31 - November 4, 2018. 

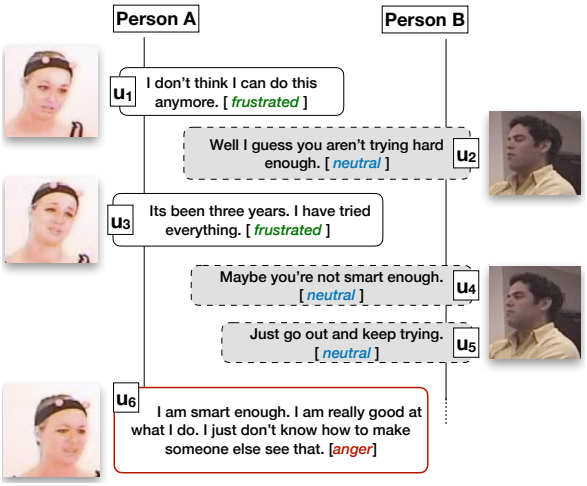


Figure 1: An abridged dialogue from the IEMOCAP dataset. $P_a$ is frustrated over her long term unemployment and seeks encouragement $(u_1, u_3)$. $P_b$, however, is pre-occupied and replies sarcastically $(u_4)$. This enrages $P_a$ to appropriate an angry response $(u_6)$. In this dialogue, *emotional inertia* is evident in $P_b$ who does not deviate from his nonchalant behavior. $P_a$, however, gets emotionally influenced by her counterpart. This influence is content-based, not label-based.

The contributions of this paper are as follows:

- We propose ICON, a novel model for emotion recognition that incorporates self and inter-speaker influences in a dialogue. Memory networks are used to model contextual summaries for prediction.
- We introduce a multimodal approach that provides comprehensive features from modalities such as language, visual, and audio in utterance-videos.
- ICON can be considered as a generic framework for conversational modeling that can be extended to multi-party conversations.
- Experiments on two benchmark datasets show that ICON significantly outperforms existing models on multiple discrete and continuous emotional categories.

The remainder of the paper is organized as follows: Section 2 presents related works; Section 3 formalizes the problem statement and Section 4 describes our proposed approach; Section 5 provides details on experimental setup; Section 6 reports the results and related analysis; finally, Section 7 concludes the paper.

## 2 Related Works

Emotion recognition is an interdisciplinary field of research with contributions from psychology, cognitive science, machine learning, natural language processing, and others (Picard, 2010).

Initial research in this area primarily involved visual and audio processing (Ekman, 1993; Datcu and Rothkrantz, 2008). The role of text in emotional analysis became evident with later research such as Alm et al. (2005); Strapparava and Mihalcea (2010). Current research in this domain is mainly performed from a multimodal learning perspective (Poria et al., 2017a; Baltrušaitis et al., 2018). Numerous previous approaches have relied on fusion techniques that leverage multiple modalities for affect recognition (Soleymani et al., 2012; Zadeh et al., 2017; Chen et al., 2017; Tzirakis et al., 2017; Zadeh et al., 2018b).

Understanding conversations is crucial for machines to replicate human language and discourse. Emotions play an important role in shaping such social interactions (Ruusuvuori, 2013). Richards et al. (2003) attribute emotional dynamics to be an interactive phenomena, rather than being within-person. We utilize this trait in the design of our model that accommodates inter-personal dynamics. Being a temporal event, context also plays an important role in conversational analysis. Poria et al. (2017b) use contextual information from neighboring utterances of the same speaker to predict emotions. However, there is no provision to model interactive influences. Work by Yang et al., 2011; Xiaolan et al., 2013 stresses the study of patterns for emotion transitions. In contrast, we posit the use of utterance content to model context with multimodal features.

In the literature, memory networks have been successfully applied in many areas, including 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 others. In emotional analysis, Zadeh et al. (2018a) propose a memory-based sequential learning for multi-view signals. Although we utilize memory networks, our work is different as we use memories to encode whole utterances. Also, each memory cell in our network is processed using GRUs to capture temporal dependencies. This technique deviates from the traditional use of embedding matrices to encode information into memory cells.

ICON builds on our previous research (Hazarika et al., 2018) that used separate memory networks for both interlocutors participating in a dyadic conversation. In contrast, ICON adopts an interactive scheme that actively models inter-speaker emotional dynamics with fewer trainable parameters.

## 3 Problem Setting

Let us define a conversation $U$ to be a set of asynchronous exchange of utterances between two persons $P_a$ and $P_b$ over time. With $T$ utterances, $U = \{u_1, u_2, ..., u_T\}$ is a totally ordered set which can be arranged as a sequence $(u_1, ..., u_T)$ based on temporal occurrence. Here, each utterance $u_i$ is spoken by either $P_a$ or $P_b$. Furthermore, for each $\lambda \in \{a, b\}$, $U_\lambda$ denotes person $P_\lambda$'s individual utterances in $U$, i.e., $U_\lambda = \{u_i \mid u_i \in U \text{ and } u_i \text{ spoken by } P_\lambda, \ \forall i \in [1, |U|]\}$. This provides two sets of utterances for both the respective speakers, such that $U = U_a \cup U_b$.

Our aim is to identify the emotions of utterances in conversational videos. At each time step $t \in [1, T]$ of video $U$, our model is provided with the utterance spoken at that time, i.e. $u_t$, and tasked to predict its emotion. Moreover, we also utilize the previous utterances within $U$ spoken by both persons. Considering a context-window of size $K$, the preceding utterances of $P_a$ and $P_b$ (starting with the most recent) within this context-window can be represented by $H_a$ and $H_b$, respectively. Formally, for each $\lambda \in \{a, b\}$, $H_\lambda$ is created as,

$$H_\lambda = \{u_i \mid i \in [t-K, t-1] \textit{ and } u_i \in U_\lambda\} \quad (1)$$

$$\textit{and} \quad |H_a| + |H_b| \leq K \quad (2)$$

Table 1 provides a sample conversation with a context-window of size $K = 5$.

| | |
|---|---|
| $U$ | $\{ u_1^a, u_2^a, u_3^b, u_4^a, u_5^a, u_6^b \}$ |
| $U_a, U_b$ | $\{ u_1, u_2, u_4, u_5 \}, \{ u_3, u_6 \}$ |
| test utterance | $u_7^a$ |
| $H_a, H_b$ | $\{ u_2, u_4, u_5 \}, \{ u_3, u_6 \}$ |

Table 1: Sample conversation $U$ with test utterance $u_7$. Context-window $K = 5$. Here, $u_i^\lambda = i^{th}$ utterance by $P_\lambda$.

## 4 Methodology

ICON has been designed as a generic framework for affective modeling of conversations. Its computations can be categorized as a sequence of four successive modules: Multimodal Feature Extraction, Self-Influence Module, Dynamic Global-Influence Module, and Multi-hop Memory. Figure 2 illustrates the overall model.

### 4.1 Multimodal Feature Extraction

ICON adopts a multimodal framework and performs feature extraction from three modalities, i.e., language (transcripts), audio and visual.

These features are extracted for each utterance in the conversation and their concatenated vectors serve as the utterance representations. The motivation of this setup derives from previous works that demonstrate the effectiveness of multimodal features in creating rich feature representations (D'mello and Kory, 2015). These features provide complementary information from heterogeneous sources which helps to accumulate comprehensive features. Its need is particularly pronounced in videos as they are often plagued with noisy signals and missing-information within individual modalities (e.g., facial occlusion, loud background music, imperfect transcriptions).

#### 4.1.1 Textual Features

We employ a convolutional neural network (CNN) to extract textual features from the transcript of each utterance. CNNs are capable of learning abstract semantic representations of a sentence based on its words and n-grams (Kalchbrenner et al., 2014). For our purpose, we utilize a simple CNN with a single convolutional layer followed by max-pooling (Kim, 2014). The input to this network consists of pre-trained word embeddings extracted from the 300-dimensional FastText embeddings (Bojanowski et al., 2016). The convolution layer consists of three filters with sizes $f_t^1, f_t^2, f_t^3$ with $f_{out}$ feature maps each. We perform 1D convolutions using these filters followed by max-pooling on its output. The pooled features are finally projected onto a dense layer with dimension $d_t$ and its activations are used as the textual representation $\boldsymbol{t_u} \in \mathbb{R}^{d_t}$.

#### 4.1.2 Audio Features

Audio plays a significant role in determining the emotional states of a speaker (De Silva and Ng, 2000; Song et al., 2004). To extract audio features, we first format the audio of each utterance-video as a 16-bit PCM WAV file and use the open-sourced software openSMILE (Eyben et al., 2010). This tool provides high dimensional vectors for audio files that summarizes important statistical descriptors such as loudness, pitch, Mel-spectra, MFCC, etc. Specifically, we use the *IS13_ComParE*[1] extractor which provides 6373 features for each utterance. The features are then normalized using Min-Max scaling followed by L2-based feature selection. This selection provides low-dimensional audio features $\boldsymbol{a_u} \in \mathbb{R}^{d_a}$ of dimensions $d_a$.

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

#### 4.1.3 Visual Features

Visual indicators such as facial expressions are key to understand emotions. In our work, we use a deep 3D-CNN to model spatiotemporal features of each utterance video (Tran et al., 2015). 3D-CNN helps to understand emotional concepts such as smiling or frowning that are often spread across multiple frames of a video with no predefined spatial location. The input to this network is a video with dimensions $(c, h, w, f)$, where $c$ is the number of channels, $h, w$ are the height and width of each frame, with a total of $f$ frames per video.

The network contains three blocks of convolution where each block contains two convolutional layers followed by max-pooling. For the convolution, 3D filters are employed having dimensions $(f_{out}, f_{in}, f_h, f_w, f_d)$, where, $f_{[out/in/h/w/d]}$ represents the number of feature maps, input channels, height, width, and depth of the filter, respectively. After a non-linear reLU activation (LeCun et al., 2015), max-pooling is performed using a sliding window of dimensions $(m_p, m_p, m_p)$. For an input utterance video, the final features of the third convolutional block is mapped onto a dense layer of dimension $d_v$ whose activations are used as the visual features $\boldsymbol{v_u} \in \mathbb{R}^{d_v}$.

#### 4.1.4 Fusion

We generate the final representation of an utterance $u$ by concatenating all three multimodal features:

$$\boldsymbol{u} = tanh((W^f[\boldsymbol{t_u}; \boldsymbol{a_u}; \boldsymbol{v_u}]) + \boldsymbol{b}^f) \quad (3)$$

Concatenation is one of the most common fusion methods (Shwartz et al., 2016). Its simplicity also allows us to emphasize the contribution of the remaining components of ICON.

### 4.2 SIM: Self-Influence Module

Given a test utterance $u_t$ to be classified, this module independently processes the histories of both speakers. SIM consists of two GRUs, $GRU^s_a$ and $GRU^s_b$, for $H_a$ and $H_b$, respectively. For each $\lambda \in \{a, b\}$, $GRU^s_\lambda$ attempts to model the *emotional inertia* of speaker $P_\lambda$ which represents the emotional dependency of a speaker with their own previous states. In particular, for each historical utterance $\boldsymbol{u}_{i<t} \in H_\lambda$, an internal memory state $\boldsymbol{h}^{(j)}_\lambda$ is computed by $GRU^s_\lambda$ conditioned on utterance $\boldsymbol{u}_i$ and previous memory state $\boldsymbol{h}^{(j-1)}_\lambda$. This can be abbreviated as $\boldsymbol{h}^{(j)}_\lambda = GRU^s_\lambda(\boldsymbol{u}_i, \boldsymbol{h}^{(j-1)}_\lambda)$.

**Gated Recurrent Unit:** GRUs are gated recurrent cells introduced by Cho et al. (2014). At time step $j$, GRU computes hidden state $s_j \in \mathbb{R}^{d_{em}}$ by calculating two gates, $r_j$ (*reset gate*) and $z_j$ (*update gate*) with $j^{th}$ input $\boldsymbol{x}_j$ and previous state $\boldsymbol{s}_{j-1}$. The computations are:

$$\begin{aligned}
\boldsymbol{z}_j &= \sigma(V^z \boldsymbol{x}_j + W^z \boldsymbol{s}_{j-1} + \boldsymbol{b}^z) \\
\boldsymbol{r}_j &= \sigma(V^r \boldsymbol{x}_j + W^r \boldsymbol{s}_{j-1} + \boldsymbol{b}^r) \\
\boldsymbol{v}_j &= tanh(V^h \boldsymbol{x}_j + W^h(\boldsymbol{s}_{j-1} \otimes \boldsymbol{r}_j) + \boldsymbol{b}^h) \\
\boldsymbol{s}_j &= (1 - \boldsymbol{z}_j) \otimes \boldsymbol{v}_j + \boldsymbol{z}_j \otimes \boldsymbol{s}_{j-1}
\end{aligned}$$

In this work, input $\boldsymbol{x}_j = \boldsymbol{u}_i$ and $\boldsymbol{s}_j = \boldsymbol{h}^{(j)}_\lambda$. SIM computes both sequences $H^*_a \in \mathbb{R}^{d_{em} \times |H_a|}$ and $H^*_b \in \mathbb{R}^{d_{em} \times |H_b|}$ using the respective GRUs,

$$H^*_\lambda = [\boldsymbol{h}^{(j)}_\lambda]^{|H_\lambda|}_{j=1} = GRU^s_\lambda(H_\lambda)\ ,\ \lambda \in \{a, b\} \quad (4)$$

### 4.3 DGIM: Dynamic Global Influence Module

Emotions are not only regarded as internal-psychological phenomena but also interpreted and processed communicatively through social interactions (Fiehler, 2002). Conversations exemplify such a scenario where inter-personal emotional influence persists. Theories in cognitive science also suggest the existence of *emotional contagion* that causes humans to mirror their counterpart's gesture, posture and emotional state (Chartrand and Bargh, 1999; Navarretta et al., 2016). Additionally, these interactions occur dynamically through the discourse of a dialogue.

While modeling the contextual history, we incorporate such properties using a dynamic influence module. This module maintains a global representation of the conversation and updates it recurrently at each time step of the $K$-length conversation history. For any $k \in [1, K]$, the global state is updated using a GRU operation on the previous state $\boldsymbol{s}_{k-1}$ and current speaker $P_\lambda$'s SIM memory $\boldsymbol{h}^{(j)}_\lambda$ for the corresponding spoken utterance $\boldsymbol{u}_{(t-K+k-1)}$, i.e., $\boldsymbol{h}^{(j)}_\lambda = GRU^s_\lambda(\boldsymbol{u}_{(t-K+k-1)})$.

Formally, DGIM consists of a GRU network, $GRU^g$, where the $k^{th}$ global state $\boldsymbol{s}_k$ is computed as:

$$\boldsymbol{s}_k = \begin{cases} GRU^g(\boldsymbol{h}^{(j)}_a, \boldsymbol{s}_{k-1}), & \text{if } \boldsymbol{u}_{(t-K+k-1)} \in H_a \\ GRU^g(\boldsymbol{h}^{(j)}_b, \boldsymbol{s}_{k-1}), & \text{if } \boldsymbol{u}_{(t-K+k-1)} \in H_b \end{cases} \quad (5)$$

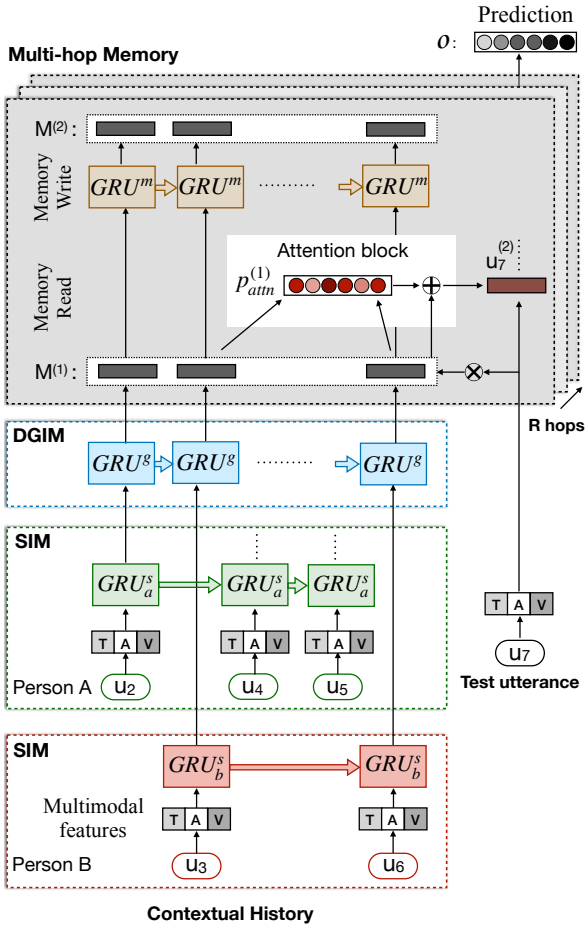


Figure 2: Illustration of ICON. Input conversation is as presented in Table 1.

### 4.4 Multi-hop Memory

The overall operation of the $GRU^g$ produces a sequence of memories $M = [\boldsymbol{s}_1, ..., \boldsymbol{s}_K] \in \mathbb{R}^{d_{em} \times K}$. These memories incorporate dynamic influences from each of the $K$ utterances spoken in the history. They serve as a contextual memory bank from which selective person-specific information can be incorporated into test utterance $u_t$ to get discriminative features. To achieve this, a series of $R$ memory read/write cycles are performed that are coupled with soft attention for refinement of $u_t$ into a context-aware representation.

The need for multiple hops is inspired by recent works on memory networks (Kumar et al., 2016; Weston et al., 2014), which suggests the importance of multiple read/write iterations for performing transitive inference. Multiple hops also help in improving the focus of attention heads which might miss essential memories in a single hop. At the $r^{th}$ hop, the computations are as follows:

- **Memory Read:** An attention mechanism is used to read the memories from $r^{th}$ memory bank $M^{(r)}$ (Weston et al., 2014). First, each memory $\boldsymbol{m}_{\boldsymbol{k}}^{(r)} \in M^{(r)}$ is matched with test utterance $\boldsymbol{u}_t^{(r)}$ (initially, $\boldsymbol{u}_t^{(1)} = \boldsymbol{u}_t$ and $M^{(1)} = M$).

This matching generates an attention vector $\boldsymbol{p}_{attn}^{(r)} \in \mathcal{R}^K$ whose $k^{th}$ normalized score represents the relevance of $k^{th}$ memory cell with respect to the test utterance. Inner product is used for the matching as follows:

$$\boldsymbol{p}_{attn}^{(r)} = softmax(\ (M^{(r)})^T \boldsymbol{u}_t^{(r)}\ ) \quad (6)$$

Where, $softmax(x_i) = e^{x_i} / \sum_j e^{x_j}$. These scores are then used to find a weighted representation of the memories as

$$\boldsymbol{m}^{(r)} = \sum_{k=1}^{K} (\boldsymbol{p}_{attn}^{(r)})^k . (\boldsymbol{m}_k) = M^{(r)} \boldsymbol{p}_{attn}^{(r)} \quad (7)$$

This vector denotes the summary of the context that is person-specific and based on the test utterance. Finally, the representation of the test utterance is updated by consolidating itself with the weighted memory $\boldsymbol{m}$ as:

$$\boldsymbol{u}_t^{(r+1)} = tanh(\boldsymbol{m}^{(r)} + \boldsymbol{u}_t^{(r)}) \quad (8)$$

- **Memory Write:** After the read operation at each hop, memories are updated for the next hop. For this purpose, a GRU network, $GRU^m$, takes the $r^{th}$ memory cells $M^{(r)}$ as input and reprocesses this sequence to generate memories $M^{(r+1)}$, i.e., $M^{(r+1)} = GRU^m(M^{(r)})$. Across all hops, this write operation can be viewed as that of a stacked recurrent neural network (RNN) where each level (or hop) improves the representational output of the RNN. The parameters of $GRU^m$ are shared across all hops.

**Final Prediction:** We use the $(R+1)^{th}$ test utterance vector $\boldsymbol{u}_t^{(R+1)}$ and get the final prediction vector through its affine transformation,

$$\boldsymbol{o} = softmax(W^o \boldsymbol{u}_t^{(R+1)} + \boldsymbol{b}^o) \quad (9)$$

For classification, dimensions of vector $\boldsymbol{o}$ is the number of classes $C$, i.e., $\boldsymbol{o} \in \mathbb{R}^C$ and *categorical cross-entropy* loss is used as the cost measure for training. For regression, $\boldsymbol{o}$ is a scalar (without softmax normalization) whose scores are used to calculate the *mean squared error* cost metric.

| Dataset | Fold | No. of Utterances | No. of Videos | Avg. history length |
|---|---|---|---|---|
| IEMOCAP | train/val | 5810 | 120 | 36.54 |
| | test | 1623 | 31 | 39.00 |
| SEMAINE | train/val | 4368 | 63 | 43.61 |
| | test | 1430 | 32 | 45.61 |

*val = validation set.

Table 2: Summary of datasets. Note: Avg. history length represents the expected number of historical utterances available for any utterance in the dataset.

## 5 Experiments

### 5.1 Datasets

We perform experiments on two benchmark datasets in dialogue-based emotion detection: *IEMOCAP*[2] (Busso et al., 2008) and *SEMAINE*[3] (McKeown et al., 2012).

**IEMOCAP** is a database consisting of videos of dyadic conversations between pairs of 10 speakers. Grouped into five sessions, each pair is assigned with diverse scenarios for dialogues. Videos are segmented into utterances with annotations of fine-grained emotion categories. We consider six such categories for the classification task: *anger, happiness, sadness, neutral, excitement,* and *frustration*. The training set is curated using the first 8 speakers from session 1-4 while session 5 is used for testing.

**SEMAINE** is a video database of human-agent interactions. Here, users interact with characters whose responses are based on users' emotional state. Specifically, we utilize the AVEC 2012's *fully continuous sub-challenge* (Schuller et al., 2012) that requires predictions of four continuous affective dimensions: arousal, expectancy, power, and valence. The gold annotations are available for every 0.2 seconds in each video (Nicolle et al., 2012). However, to align with our problem statement, we approximate the utterance-level annotation as the mean of the continuous values within the spoken utterance. The sub-challenge provides standard training and testing splits which has been summarized in Table 2.

### 5.2 Training Details

20% of the training set is used as validation set for hyper-parameter tuning. We use the Adam optimizer (Kingma and Ba, 2014) for training the parameters starting with an initial learning rate of 0.001. Termination of the training-phase is decided by early-stopping with a patience of 10 epochs. The network is subjected to regularization in the form of Dropout (Srivastava et al., 2014) and Gradient-clipping for a norm of 40. Finally, the best hyper-parameters are decided using a grid-search. Their values are summarized in Table 3.

| | | |
|---|---|---|
| $(f_t^1, f_t^2, f_t^3) = (3, 4, 5)$ | $f_{[h,w,d]} = 3$ | $f_{out} = 64$ |
| $d_{[t,a]} = 100$ | $d_v = 512$ | $d_{em} = 100$ |
| $K = 40$ | $R = 3$ | |

Table 3: Hyper-parameter values for the best model.

For multimodal feature extraction, we explore different designs for the employed CNNs. For text, we find the single layer CNN to perform at par with deeper variants. For visual features, however, a deeper CNN provides better representations. We also find that contextually conditioned features perform better than context-less features. Thus, in our experiments, we extract video-level contextual features for utterances from each modality using the network proposed by Poria et al. 2017b. These modified features are then used to form the multimodal utterance representations using equation 3.

### 5.3 Baselines

We compare our proposed model with multiple state-of-the-art networks in multimodal utterance-level emotion detection.

- *memnet* (Sukhbaatar et al., 2015) is an end-to-end memory network. For comparison, we modify our network to adopt their embedding-based memory-encoding in the multi-hop stage.
- *cLSTM*[4] (Poria et al., 2017b) classifies utterances using neighboring utterances (of same speaker) as context. LSTM is used for this purpose.
- *TFN*[5] (Zadeh et al., 2017) models intra- and inter-modality dynamics by explicitly aggregating uni-, bi- and trimodal interactions. Unlike cLSTM, contextual utterances are not considered.
- *MFN* (Zadeh et al., 2018a) performs multi-view learning by using Delta-memory Attention Network, a fusion mechanism to learn cross-view interactions. Similar to TFN, the modeling is performed within utterances.
- *CMN* (Hazarika et al., 2018) models separate contexts for both speaker and listener to an utterance. These contexts are stored as memories and combined with test utterance using attention mechanism.

[2] http://sail.usc.edu/iemocap/
[3] http://sspnet.eu/avec2012/
[4] http://github.com/senticnet/contextual-sentiment-analysis
[5] http://github.com/A2Zadeh/TensorFusionNetwork

| Models | IEMOCAP: Emotion Categories | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Happy* | | *Sad* | | *Neutral* | | *Angry* | | *Excited* | | *Frustrated* | | ***Avg.*** | |
| | acc. | F1 | acc. | F1 | acc. | F1 | acc. | F1 | acc. | F1 | acc. | F1 | acc. | F1 |
| memnet | 24.4 | 33.0 | 60.4 | 69.3 | 56.8 | 55.0 | 67.1 | 66.1 | 65.2 | 62.3 | 68.4 | 63.0 | 59.9 | 59.5 |
| cLSTM | 25.5 | **35.6** | 58.6 | 69.2 | 56.5 | 53.5 | 70.0 | 66.3 | 58.8 | 61.1 | 67.4 | 62.4 | 59.8 | 59.0 |
| TFN | 23.2 | 33.7 | 58.0 | 68.6 | 56.6 | 55.1 | 69.1 | 64.2 | 63.1 | 62.4 | 65.5 | 61.2 | 58.8 | 58.5 |
| MFN | 24.0 | 34.1 | 65.6 | 70.5 | 55.5 | 52.1 | **72.3**$^{\dagger}$ | 66.8 | 64.3 | 62.1 | 67.9 | 62.5 | 60.1 | 59.9 |
| CMN | **25.7** | 32.6 | 66.5 | 72.9 | 53.9 | 56.2 | 67.6 | 64.6 | 69.9 | 67.9 | 71.7 | 63.1 | 61.9 | 61.4 |
| ICON | 23.6 | 32.8 | **70.6**$^{\dagger}$ | **74.4**$^{\dagger}$ | **59.9** | **60.6**$^{\dagger}$ | 68.2 | **68.2** | **72.2**$^{\dagger}$ | **68.4** | **71.9** | **66.2**$^{\dagger}$ | **64.0**$^{\dagger}$ | **63.5**$^{\dagger}$ |

Table 4: Performance of ICON on the IEMOCAP dataset. $^{\dagger}$ represents statistical significance over state-of-the-art scores under the paired-t test ($p < 0.05$).

| Models | SEMAINE | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | DV | | DA | | DP | | DE | |
| | MAE | r | MAE | r | MAE | r | MAE | r |
| memnet | .20 | .16 | .21 | .24 | .21 | .23 | 8.97 | .05 |
| cLSTM | **.18** | .14 | .21 | .23 | .20 | .25 | 8.90 | -.04 |
| TFN | .21 | .01 | .22 | .10 | .21 | .12 | 9.19 | .12 |
| MFN | .19 | .14 | .20 | .25 | **.18** | .26 | 8.60 | .15 |
| CMN | **.18** | .23 | .20 | .30 | **.18** | .26 | 8.89 | -.02 |
| ICON | **.18** | **.24** | **.19** | **.31** | **.18** | **.27** | **8.45** | **0.24** |

Table 5: Performance on the SEMAINE dataset. Note: MAE = Mean Absolute Error, r = Pearson's correlation coefficient, DV = Valence, DA = Activation/Arousal, DP = Power, DE = Anticipation/Expectation.

| Modality | IEMOCAP | | SEMAINE | | | |
|---|---|---|---|---|---|---|
| | Emotions | | DV | DA | DP | DE |
| | acc. | F1 | r | r | r | r |
| T | 58.3 | 57.9 | .237 | .297 | .260 | .225 |
| A | 50.7 | 50.9 | .021 | .082 | .250 | .035 |
| V | 41.2 | 39.8 | .001 | .068 | .251 | .001 |
| A+V | 52.0 | 51.2 | .031 | .122 | .283 | .050 |
| T+A | 63.8 | 63.2 | .237 | **.310** | .272 | **.242** |
| T+V | 61.4 | 61.2 | .238 | .293 | .268 | .239 |
| T+A+V | **64.0** | **63.5** | **.243** | **.312** | **.279** | **.244** |

Table 6: Comparison of the performance of ICON on both IEMOCAP and SEMAINE considering different modality combinations. Note: T=Text, A=Audio, V=Video

## 6 Results

Tables 4 and 5 present the results on the IEMOCAP and SEMAINE testing sets, respectively. In Table 4, we evaluate the mean classification performance using Weighted Accuracy (*acc.*) and F1-Score (*F1*) on the discrete emotion categories. ICON performs better than the compared models with significant performance increase in emotions (~2.1% *acc.*). For each emotion, ICON outperforms all the compared models except for *happiness* emotion. However, its performance is still at par with *cLSTM* without a significant gap. Also, ICON manages to correctly identify the relatively similar *excitement* emotion by a large margin.

In Table 5, evaluations of the four continuous labels from SEMAINE are performed using Mean Absolute Error (*MAE*) and Pearson's Correlation Coefficient (*r*). In all the labels, ICON attains improved performance over its counterparts, suggesting the efficacy of its context-modeling scheme.

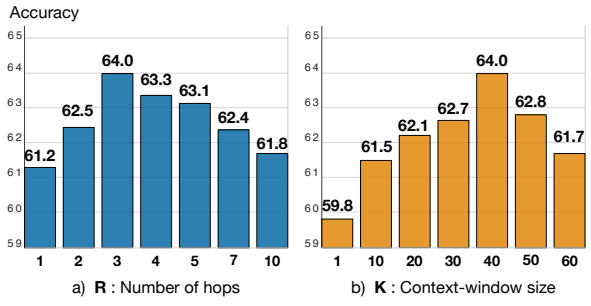


Figure 3: Trends in the performance of ICON on IEMOCAP dataset with varying R (hops) and K (Context-window size).

**Hyperparameters:** We plot the performance trends of ICON on the IEMOCAP dataset concerning the two main hyperparameters, R (number of hops) and K (context-window size). For R, the performance initially improves showing the importance of multiple hops in the memories. However, with a further increase, the hopping recurrence deepens and causes the *vanishing gradient* problem. This leads to decrease in performance. The best performance is obtained at $R = 3$. For K, similar trends are observed where performance improvement is seen by increasing the number of historical utterances. The best results are obtained for $K = 40$ which also aligns with the average number of historical utterances in the dataset (Table 2). Further increase in context does not provide relevant information and rather leads to performance degradation due to model confusion.

**Multimodality:** We investigate the importance of multimodal features for our task. Table 6 presents the results for different combinations of modes used by ICON on IEMOCAP. As seen, the trimodal network provides the best performance which is preceded by the bimodal variants. Among

| ICON variants | | | | IEMOCAP Emotions | | SEMAINE DV | SEMAINE DA | SEMAINE DP | SEMAINE DE |
|---|---|---|---|---|---|---|---|---|---|
| | history | DGIM | hop | acc. | F1 | r | r | r | r |
| 1. | - | - | - | 58.0 | 57.6 | .10 | .14 | .10 | .01 |
| 2. | self | ✓ | - | 60.7 | 60.2 | .17 | .23 | .15 | .13 |
| 3. | dual | ✓ | - | 61.2 | 60.7 | .19 | .24 | .19 | .20 |
| 4. | self | - | ✓ | 61.9 | 61.3 | .19 | .23 | .22 | .20 |
| 5. | dual | - | ✓ | 63.1 | 62.4 | .21 | .25 | .26 | .22 |
| 6. | self | ✓ | ✓ | 62.2 | 61.7 | .20 | .28 | .21 | .22 |
| 7. | dual | ✓ | ✓ | **64.0** | **63.5** | **.24** | **.31** | **.27** | **.24** |

Table 7: Ablation study for components of ICON.

unimodals, language modality performs the best, reaffirming its significance in multimodal systems. Interestingly, the audio and visual modality, on their own, do not provide good performance, but when used with text, complementary data is shared to improve overall performance.

### 6.1 Ablation Study

To check the importance of the modules present in ICON, we perform an ablation study where we remove constituent components and evaluate the model's performance. Table 7 provides the results on this study. In the first variant, none of the histories and the associated context-modeling is used. This provides the worst relative performance.

**Self vs Dual History:** We evaluate the scenarios where only self-history of the speaker is considered (variants 2, 4, and 6). Compared to the dual-history variants (variants 3, 5, and 7), these models provide lesser performance. Reasons involve the provision of partial information from the conversational histories. Similar trends can be seen for the *cLSTM* model in Table 4 which works in the same regime.

**DGIM vs no-DGIM:** Variants 4 and 5 do not contain the DGIM. In variant 5, separate memory banks are created for both histories ($M_a = H_a^*$ and $M_b = H_b^*$). Memory hops are also separately performed without parameter sharing. Absence of DGIM prevents the storage of dynamic influences between speakers at each historical time step and leads to performance deterioration.

**Multi-hop vs No-hop:** Variants 2 and 3 represent cases where multi-hop is omitted, i.e., $R = 1$. Performance for them are poorer than variants having multi-hop mechanism (variants 4-7). Also, removal of multi-hop leads to worse performance than the removal of DGIM. This suggests that multi-hop is more crucial than the latter. However, best performance is achieved by variant 6 which contains all the proposed modules in its pipeline.

### 6.2 Dependency on distant history

For all the test utterances of IEMOCAP correctly classified by ICON, we analyze the global memories receiving the highest attention. First, we divide the conversational history (context-length $K = 40$) into three regions: *long*, *short*, and *medium*. Figure 4 provides a summary of how much the model attends each of these regions. The short region (labeled green) covering 10 utterances, corresponds to conversational history just preceding the test utterance. Utterances which occur more than 30 time steps behind the current test utterance are considered part of the long region (labeled red). Remaining utterances in between fall on the medium region (labeled blue).

The distribution of top-valued attention scores across the histories reveal interesting insights. Most of the correctly classified instances focus on the immediate or short history. In other words, 63% of the time, at least one of the top-5 attention value belongs to a memory in the short-history range. A significant share is also present for distant history (22%). This result indicates the presence of long-term emotional dependencies and the need to consider histories far away from the current test utterance.

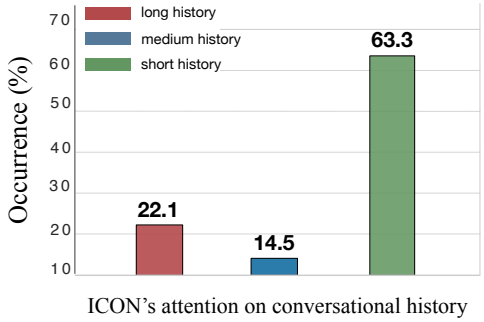


Figure 4: Distribution of top-attention by ICON on correctly classified instances in the testing set.

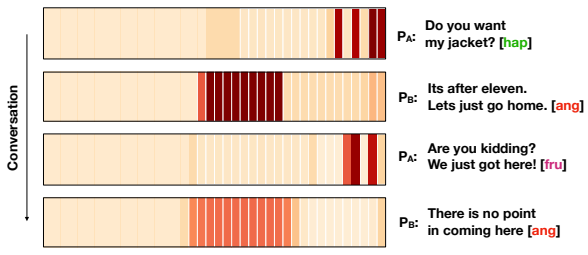


Figure 5: As a conversation develops, different speakers induce different affective bias which reflects in the memory selection for generation of the summaries.

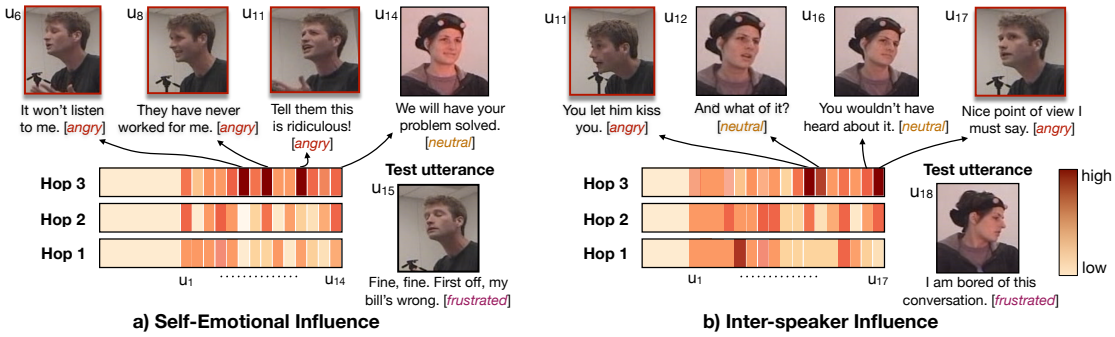


Figure 6: Case studies for emotional influence. 20 memories in the history which are nearest to test utterance, i.e. $k \in [21, 40]$ are visualized from the trained ICON.

### 6.3 Dynamic Modeling of Global Memories:

ICON holds the capability to model dynamic interactions between speakers. The memories by its DGIM (§4.3) are used to create summaries conditioned on the test utterance. Consequently, these summaries contain characteristics that are specific to the affective state of the current speaker (of the test utterance).

Figure 5 presents a sample slice of conversation from the dataset. As seen, summary selection for Person A varies from Person B. Such differences arise due to person-specific characteristics and unique affective interpretations of the conversation. Apart from the inter-speaker variance, the emotional state of a speaker also varies across turns.

### 6.4 Case Studies

To understand ICON's behavior while processing the global memories through multi-hop, we manually explore the utterances in the testing set of IEMOCAP. Figure 6 presents two cases which provide traces of self and inter-personal emotional influences and were correctly classified by ICON. Both the figures show the trend where multiple hops gradually improve the focus of attention mechanism on relevant memories.

In Figure 6a, person $P_a$ registers a complaint to an operator $P_b$. Throughout the dialogue, $P_a$ maintains an angry demeanor while $P_b$ remains calm and neutral ($u_{14}$). While classifying utterance $u_{15}$, ICON focuses more on the histories uttered by $P_a$ ($u_6$, $u_8$, and $u_{11}$). This demonstrates ICON's ability to model self-emotional influences. It should be noted that emotion of $P_a$ here also depends on the utterances of $P_b$ but compared to self-utterances, this dependency is much less. Figure 6b presents another scenario where a couple argue over an alleged affair.

A man ($P_b$) is angry over this fact and questions his partner ($P_a$) asking for details. The woman tries to behave unperturbed by providing neutral responses ($u_{12}$, $u_{16}$) but is eventually affected by $P_b$'s continuous anger and expresses a frustrated response ($u_{18}$). These characteristics are captured by the attention mechanism applied on the global memories (generated by DGIM), which finds contextual information from histories that are relevant to the test utterance $u_{18}$. This example displays the role of inter-speaker influences and how ICON processes such dependencies.

## 7 Conclusion

In this paper, we presented ICON, a multimodal framework for emotion detection in conversations. ICON capitalizes on modeling contextual information that incorporates self and inter-speaker influences. We accomplish this by using an RNN-based memory network with multi-hop attention modeling. Experiments show that ICON outperforms state-of-the-art models on multiple benchmark datasets. Extensive evaluations and case studies demonstrate the effectiveness of our proposed model. Additionally, the ability to visualize the attentions brings a sense of interpretability to the model, as it allows us to investigate which utterances in the conversational history provide important emotional cues for the current emotional state of the speaker.

In the future, we plan to test ICON on other relevant dialogue-based applications and also use it for empathetic dialogue generation.

## Acknowledgments

This research has been supported in part by Singapore's Ministry of Education (MOE) Academic Research Fund Tier 1, grant number T1 251RES1713.

## References


Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 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*, pages 579–586. Association for Computational Linguistics.

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

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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.

Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

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*.

Liyanage C De Silva and Pei Chi Ng. 2000. Bimodal emotion recognition. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 332–335. IEEE.

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 *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Reinhard Fiehler. 2002. How to do emotions with words: Emotionality in conversations. *The verbal communication of emotions*, pages 79–106.

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

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 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)*, volume 1, pages 2122–2132.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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 *International Conference on Machine Learning*, pages 1378–1387.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7):984–991.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

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*.

Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. 2012. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 501–508. ACM.

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.

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.

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

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. ACM.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. *arXiv preprint arXiv:1603.06076*.

Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Carlo Strapparava and Rada Mihalcea. 2010. Annotating and identifying emotions in text. In *Intelligent Information Access*, pages 21–38. Springer.

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 Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

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

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.

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 dialogue systems with commonsense knowledge. In *AAAI*, pages 4970–4977.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *AAAI*, pages 5634–5641.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *AAAI*, pages 5642–5649.