Intelligent Data Analysis 25 (2021) 1031–1045
DOI 10.3233/IDA-205183
IOS Press

Multimodal emotion recognition with hierarchical memory networks

Helang Lai[a,b], Keke Wu[c,*] and Lingli Li[a]
[a]*Guangdong Justice Police Vocational College, Guangzhou, Guangdong, China*
[b]*School of Computer Science, South China Normal University, Guangzhou, Guangdong, China*
[c]*Shenzhen Institute of Information Technology, Shenzhen, Guangdong, China*

Abstract. Emotion recognition in conversations is crucial as there is an urgent need to improve the overall experience of human-computer interactions. A promising improvement in this field is to develop a model that can effectively extract adequate contexts of a test utterance. We introduce a novel model, termed hierarchical memory networks (HMN), to address the issues of recognizing utterance level emotions. HMN divides the contexts into different aspects and employs different step lengths to represent the weights of these aspects. To model the self dependencies, HMN takes independent local memory networks to model these aspects. Further, to capture the interpersonal dependencies, HMN employs global memory networks to integrate the local outputs into global storages. Such storages can generate contextual summaries and help to find the emotional dependent utterance that is most relevant to the test utterance. With an attention-based multi-hops scheme, these storages are then merged with the test utterance using an addition operation in the iterations. Experiments on the IEMOCAP dataset show our model outperforms the compared methods with accuracy improvement.

Keywords: Dyadic conversations, emotion recognition, multimodal, memory network, GRUs

1. Introduction

With the rapid development of Artificial Intelligence (AI) and the increasing interactions of humans with machines, multimodal emotion recognition (MER) in conversations has been attracting great attention from the research community. Argueta et al. [1] employed an unsupervised pattern extraction to implement multilingual emotion classifier, Zhang et al. [2] proposed a hierarchical emotion structure to classify emotions, and Vu et al. [3] researched the facial expression recognition task. The potential applications of MER have been involving in many important and challenging tasks such as counseling, dialogue generation, public opinion mining, financial forecasting, intelligent systems, and user behavior understanding over chat history and social media threads on YouTube, Facebook, Twitter, and so on [4,5]. One of the important applications is to create empathetic dialogue systems with emotional understanding [6]. Previous research has considered dialogues as an essential basis of capturing the emotional dynamics [7–9]. However, as the expressiveness of emotion varies widely from person to person, analyzing and identifying emotional dynamics in conversations pose enormous challenges. There are complex dependencies between the affective states of speakers participating in the dialogue [10]. In

*Corresponding author: Keke Wu, Shenzhen Institute of Information Technology, Shenzhen, Guangdong, China. E-mail: wukk@sziit.edu.cn.

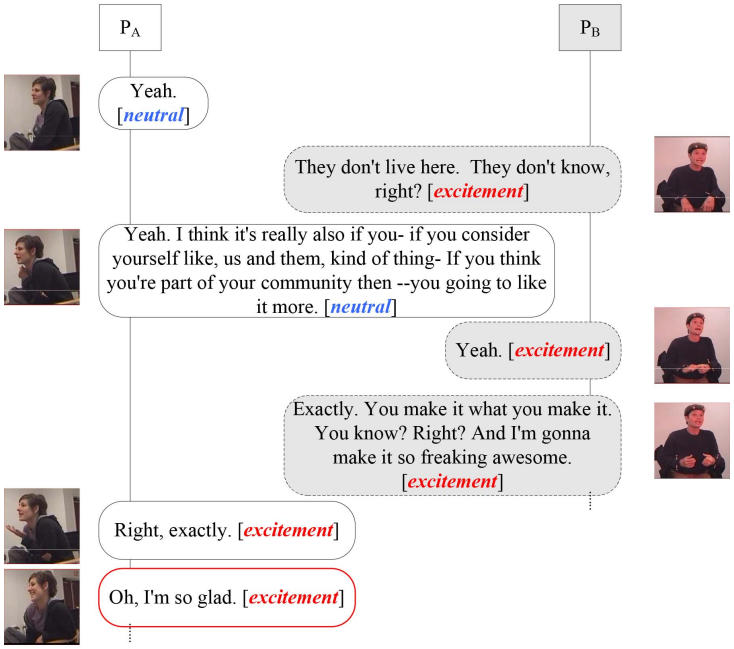

Fig. 1. An abridged dialogue from the IEMOCAP dataset. In this dialogue, P_A's emotion changes are influenced by the behavior of P_B.

this paper, we cope with the challenges of emotion recognition in conversational videos. Specifically, we focus on utterance level (generally, a unit of speech bounded by breathes or pauses of the speaker) MER in dyadic conversations that is a form of a dialogue between two entities.

We propose a novel model termed hierarchical memory networks (HMN). The novel of our model is to distinguish the contexts of the test utterance and integrate different aspects of the contexts with hierarchical memory networks. The emotional dynamics in a dyadic conversation is known to consist of two main factors: self and inter-personal dependencies [11,12]. Self-dependencies also termed emotion inertia, reflect the degree to which a person's feelings carry over from one moment to another [13], or in other words, involve the processes of emotional influence that speakers have on themselves during conversations [14]. On the other hand, inter-personal dependencies refer to the emotional influences that the other speakers induce into a specific speaker [4], as speakers may tend to follow the emotions of their counterparts during the course of a dialogue [15].

Figure 1 presents a conversational example with these two traits of both self and inter-personal dependencies. Although several existing works have capitalized on these two factors [4,5,10], they ignore that different aspects of contexts of the test utterance in a conversation may have different weights. In this study, we propose to differentiate the contexts. In other words, we consider different aspects of the contexts should have different weights relevant to the test utterance. We mark these aspects as *ownHis*, *otherHis*, *ownFut*, and *otherFut* and use different step lengths to represent the implicit weights of these different aspects. First, HMN can generate different aspects of contexts of u_t by collecting corresponding utterances within different step lengths while classifying a particular utterance u_t. Then, HMN integrates these aspects with hierarchical memory networks. To model self-dependencies, these aspects are independently modeled into local storage cells using local memory networks (GRU^L cells).

Furthermore, to incorporate inter-personal dependencies, two global representations are generated using two global memory networks (GRU^G cells). These two representations are corresponding to historical

contexts and future contexts. Specifically, one global memory network deals with historical contexts by integrating the outputs of the local memory networks of both *ownHis* and *otherHis* aspects. The other global memory network processes future contexts by integrating the outputs of the local memory networks of both *ownFut* and *otherFut* aspects. The outputs of these two global memory networks are conveyed to iterative memory networks (GRU^M cells), followed by a multi-hops scheme that includes an attention mechanism. The scheme can help to generate the contextual summaries of both the historical and future contexts in the conversation. At each hop in the multi-hops scheme, the representation of the test utterance can be improved with these contextual summaries. After H hops, the updated representation of utterance u_t is used to classify its emotion category. The experiments show the effect of our model on capturing self and inter-personal dependencies.

The contributions of this paper can be summarized as follows:

- We propose a novel model HMN for emotion recognition. The model represents the weights of the different aspects of contexts using different step lengths and integrates these aspects with hierarchical memory networks. For MER of utterance level in a dyadic conversation, our model is effective in capturing emotional dynamics and boosting accuracy.
- HMN captures the self-dependencies using four separate local memory networks, incorporates inter-personal dependencies using two global memory networks, and conducts attention-based multi-hops using two iterative memory networks. Besides, to obtain comprehensive features, all utterances of conversational videos are represented with a multimodal approach that can resist the interference of noise information.
- Experiments on benchmark dataset IEMOCAP show that HMN achieves competitive performance in comparison with the baselines.

This paper is further structured as follows: Section 2 discusses the related work in this area of research; Section 3 describes our proposed method in detail; Section 4 provides an experimental setup and reports the result; Section 5 presents the discussion and analysis; Section 6 summarizes the conclusion and further research directions.

2. Related work

Over the years, emotion recognition as an interdisciplinary field of research has obtained impetus from researchers across various areas such as social psychology, cognitive science, natural language processing, machine learning, and so on [16]. Ekman [17] conducted an initial finding between emotion and facial expressions. Datcu and Rothkrantz [18,19] combined acoustic information with visual cues in emotion modeling. Alm et al. [20] addressed the emotion recognition by adding the text-based information, developed in the later work of Strapparava and Mihalcea [21]. Current research in emotion recognition is mainly from a multimodal learning perspective [8,22]. A large number of previous works, such as [23–28], have relied on multimodal fusion techniques that research emotion recognition from a multimodal perspective. The fusion of modalities has exhibited excellent performances in affect recognition systems [23,27,29,30], therefore stimulating to use of multimodal fusion. For processing context-sensitive recognition, our proposed method also represents all the utterances in the videos with a multimodal fusion approach.

Convolutional neural network (CNN) and recurrent neural network (RNN) are two widely used neural networks, such as Liu et al. [31] proposed an Attention-Gated CNN for sentence classification, and Lu et al. [32] utilized RNN for topic discovery. These applications also have been instrumental in the progress

of the emotion recognition problem. Ebrahimi et al. [33] focused on a hybrid CNN-RNN architecture for facial expression analysis. Poria et al. [29] successfully used RNN-based deep networks for MER, followed by other researchers [26,28,34]. However, RNN is known to have difficulty in performing memorization. Memory networks can efficiently capture long-term dependencies, thus solving this problem [35–37]. As a storage mechanism and their storage cells being continuous vectors, memory networks have been successfully applied in multiple research problems, including speech recognition [36], question-answering [35,38–40], commonsense reasoning [41], and machine translation [42,43]. In the emotional analysis, Kar et al. [44] employed memory networks to classify the perception of emotional instances conveyed in facial expressions as well as to localize sources. Zadeh et al. [34] presented a method of memory-based sequential learning for multi-view signals. Hazarika et al. used two distinct memory networks to independently model context for each speaker [4], and employed local and global memory networks to hierarchically model the self and inter-speaker emotional influences into memories [10]. Further, memory networks are often together with the attention modules and multi-hops mechanism to improve the performance. In this study, we utilize both memory networks and attention-based multi-hops to build our model.

Emotions play a pivotal role in shaping conversational interactions [45]. According to Poria et al. [9, 46,47], a conversational emotion recognition system can be used to generate appropriate responses by analyzing the emotions. However, understanding emotional dynamics profoundly is a challenge for machines. Several significant works have argued that emotional dynamics can be looked upon as an interactive phenomenon, rather than being within-person and one-directional [48,49]. For capturing these emotional dynamics, Yang et al. investigate the patterns of emotional transition properties [50], and Xiaolan et al. employ finite state machines to model transitions [51] by observing stimuli and personality properties. Besides, contexts play a significant role in emotional recognition. As a conversation is a temporal event, the associated emotions of the participants' utterances generally depend on their conversational contexts. In other words, the contexts act as a set of parameters that can influence a person to speak an utterance with an emotion [46]. Sun et al. [52] consider that contextual information of both the surrounding environment and the human body can provide extra clues to recognize emotions accurately. Metallinou et al. [53] point out that long-term temporal context is beneficial for emotion recognition systems that encounter a variety of emotional manifestations. Unlike the above methods, our model first distinguishes the contexts of the test utterance and then integrates these aspects with hierarchical memory networks.

Our work is a follow-up of previous research [4] that used separate memory networks for both speakers participating in a dyadic conversation. For the improvement, we employ four different step lengths based local memory networks to capture the self dependencies and two global memory networks to model the inter-personal dependencies. Through adopting such a scheme, HMN can effectively capture the contexts surrounding the test utterance.

3. Methodology

In this section, we discuss our HMN model behind solving the MER problem.

3.1. Task definition

Let there be a set of asynchronous exchange of utterances between two persons P_a and P_b over time in a dyadic conversation. We aim to predict the emotion labels (*happiness*, *sadness*, *neutral*, and *anger*)

of utterances. With T utterances, $U = \{u_1, u_2, \ldots, u_T\}$ denotes a totally ordered set based on temporal occurrence, where u_t is the t^{th} utterance at time step $t \in [1, T]$. The utterance representation u_t is obtained using the method of feature extraction described in Section 3.2.

For the aim of emotion classification, we utilize not only the test utterance at time t (i.e. u_t), but its surrounding contextual information. This contextual information includes four aspects (respectively marked as *ownHis*, *otherHis*, *ownFut*, and *otherFut*). Each aspect separately collects the utterances within its corresponding step length. Specially, we can distinguish individual utterances in U. Let $\lambda \in \{\textit{ownHis}, \textit{otherHis}, \textit{ownFut}, \textit{otherFut}\}$, there is $U_\lambda = \{u_i | u_i \in U$ and $u_i (i \neq t)$ belongs to the aspect of $\lambda, \forall i \in [1, |U|]$ and $i \neq t\}$. In other words, each utterance $u_i (i \neq t)$ is among one of these four aspects. Hence, we have $U = U_{\textit{ownHis}} \bigcup U_{\textit{otherHis}} \bigcup U_{\textit{ownFut}} \bigcup U_{\textit{otherFut}} \bigcup u_t$. Further, we use different step lengths, i.e., $N_{\textit{ownHis}}$, $N_{\textit{otherHis}}$, $N_{\textit{ownFut}}$, and $N_{\textit{otherFut}}$, to process these different sets ($U_{\textit{ownHis}}$, $U_{\textit{otherHis}}$, $U_{\textit{ownFut}}$, $U_{\textit{otherFut}}$). Within these step lengths, four sets can be respectively constrained to $U_{\textit{ownHis}}^{N_{\textit{ownHis}}}$, $U_{\textit{otherHis}}^{N_{\textit{otherHis}}}$, $U_{\textit{ownFut}}^{N_{\textit{ownFut}}}$, and $U_{\textit{otherFut}}^{N_{\textit{otherFut}}}$.

We take a scheme of calculation direction that is starting from the oldest one for historical contexts and starting from the farthest one for future contexts. Thus, if we are processing the preceding utterances, $\lambda_1 \in \{\textit{ownHis}, \textit{otherHis}\}$, $U_{\lambda_1}^{N_{\lambda_1}}$ can be created as,

$$U_{\lambda_1}^{N_{\lambda_1}} = \{u_i | i \in [t - N_{\lambda_1}, t - 1] \textit{ and } u_i \in U_{\lambda_1}\} \tag{1}$$

and if we are processing the upcoming utterances, $\lambda_2 \in \{\textit{ownFut}, \textit{otherFut}\}$, $U_{\lambda_2}^{N_{\lambda_2}}$ can be created as,

$$U_{\lambda_2}^{N_{\lambda_2}} = \{u_i | i \in [t + 1, t + N_{\lambda_2}] \textit{ and } u_i \in U_{\lambda_2}\} \tag{2}$$

Besides, at the beginning or the ending of the conversation, the number of utterances of each aspect would have lesser than its corresponding step length. Hence we have the inequality as,

$$|U_\lambda^{N_\lambda}| \leqslant N_\lambda \tag{3}$$

Where, $\lambda \in \{\textit{ownHis}, \textit{otherHis}, \textit{ownFut}, \textit{otherFut}\}$.

In the remaining sections, we explain our model using a subscript λ for brevity. This λ can instantiate to *ownHis*, *otherHis*, *ownFut* or *otherFut*.

Table 1 demonstrates a sample conversation with different step lengths, $N_{\textit{ownHis}} = 10$, $N_{\textit{otherHis}} = 7$, $N_{\textit{ownFut}} = 6$ and $N_{\textit{otherFut}} = 3$. u_i is the i^{th} utterance in U.

3.2. Multimodal feature data

We employ the identical feature data that is downloaded from the public website[1] such that our model can remain consistent with the prior research. The extraction procedures of feature data are briefly described below [4].

A simple CNN with one convolutional layer is employed to extract the textual features from the transcript of each utterance. Max-pooling [54] is operated on the output feature maps, followed by rectified linear unit (ReLU) activation. These activations are concatenated and fed to a fully connected layer, which is regarded as the textual utterance representation t_u [4,10].

The open-sourced software open SMILE [55] can use to extract the audio features. Specifically, the IS13_ComParE[2] config file can provide 6373 features for each utterance. These features after the

[1] https://github.com/senticnet/.

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

Table 1
Sample conversation U with test utterance $u_t = u_{16}$

Set symbol	Set elements
U	$\{u_1, u_2, u_3, \ldots, u_{23}, u_{25}, u_{26}\}$
u_t	u_{16}
U_{ownHis}	$\{u_2, u_4, u_6, u_8, u_{10}, u_{12}, u_{14}\}$
$U_{ownHis}^{N_{ownHis}}$	$\{u_6, u_8, u_{10}, u_{12}, u_{14}\}$
$U_{otherHis}$	$\{u_1, u_3, u_5, u_7, u_9, u_{11}, u_{13}, u_{15}\}$
$U_{otherHis}^{N_{otherHis}}$	$\{u_9, u_{11}, u_{13}, u_{15}\}$
U_{ownFut}	$\{u_{18}, u_{20}, u_{22}, u_{24}, u_{26}\}$
$U_{ownFut}^{N_{ownFut}}$	$\{u_{18}, u_{20}, u_{22}\}$
$U_{otherFut}$	$\{u_{17}, u_{19}, u_{21}, u_{23}, u_{25}\}$
$U_{otherFut}^{N_{otherFut}}$	$\{u_{17}, u_{19}\}$

standardization are sent to a fully-connected layer to reduce the dimension, which generates the audio feature vector a_u.

A 3D-CNN [56] is employed to capture the details of facial expressions and visual surroundings from the utterance video. Then the outputs are sent to max-pooling followed by a fully connected layer, whose activations of this layer form the video representation v_u.

We obtain the final representation of an utterance by concatenating the above three multimodal features, such as $u = [t_u, a_u, v_u]$. With this multimodal representation, all utterances in a conversation are generated. Readers can refer to the literature [4,10] to obtain detailed information.

3.3. Hierarchical memory networks

In our model, we propose to differentiate the contexts of the test utterance. The surrounding contexts of a test utterance u_t should have different weights in different aspects. We mark these different aspects as *ownHis*, *otherHis*, *ownFut* and *otherFut*, respectively. Each aspect contains its corresponding utterances within its corresponding step length. With different step lengths in these different aspects, HMN attempts to capture the different weights of different aspects relevant to the test utterance. Furthermore, we separately model these different aspects. Although these four aspects are not necessarily independent, separate modeling can help to improve the accuracy performance, which is validated by the experiment results.

The overall process of HMN is as follows. We model different aspects of the contexts into local storage cells by employing separate local memory networks (*GRU*^{L} cells). For *ownHis* and *otherHis* aspects, the outputs of their corresponding local memory networks are sorted based on time occurrence order. The sorted sequence is sent to the global memory network (*GRU*^{G} cells), which is to incorporate the inter-personal dependencies of historical contexts. The output of the global memory network is conveyed to the attention block to perform an attention mechanism, resulting in weighted contextual information relevant to the test utterance. The weighted information is merged with the test utterance using an addition operation. Besides, the output of the global memory network is the input of the storage read/write module that conducts a multi-hops scheme. In the iterations, the iterative memory network (*GRU*^{M} cells) produces new storage cells. For *ownFut* and *otherFut* aspects, the outputs of their corresponding local memory network (*GRU*^{L} cells) are sorted in reverse chronological order. The rest of the process is the same as *ownHis* and *otherHis* aspects. After H hops, the final emotion representation of u_t is used to predict the classification of emotions. Figure 2 demonstrates the architecture of our model.

In the following, we first define the GRU cell; Then, present the computations of Local dependencies and Global dependencies, followed by Multiple hops storage and Final prediction.

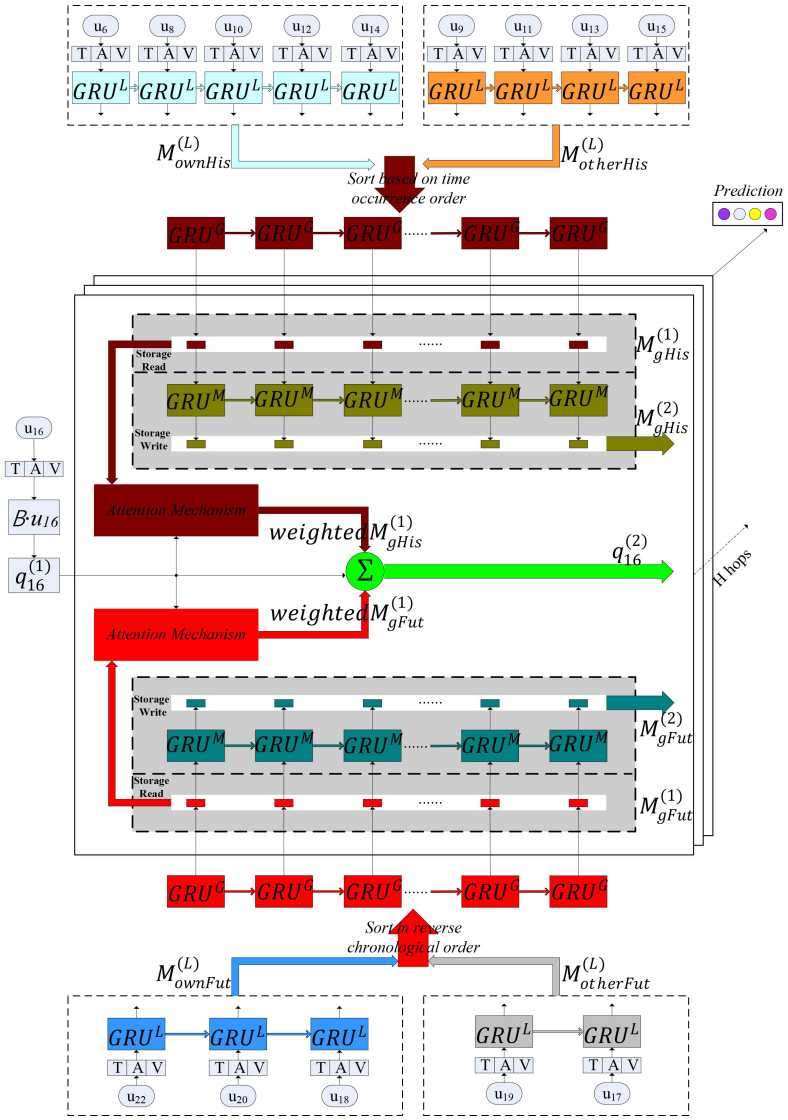

Fig. 2. Illustration of HMN. Input conversation is as presented in Table 1.

3.3.1. Gated recurrent unit

The GRU is introduced by Cho et al. [57]. It is a simpler gating mechanism with similar computation performance with respect to the LSTM [58] in RNN. At any time step i of a temporal sequence, the GRU uses two gates, r_i (reset gate) and z_i (update gate), to control the combination criteria with the i^{th} input u_i and previous state s_{i-1}. The computations are:

$$z_i = \sigma(V^z \cdot u_i + W^z \cdot s_{i-1} + b^z) \tag{4}$$

$$r_i = \sigma(V^r \cdot u_i + W^r \cdot s_{i-1} + b^r) \tag{5}$$

$$h_i = tanh(V^h \cdot u_i + W^h \cdot (s_{i-1} \otimes r_i) + b^h) \tag{6}$$

$$s_i = (1 - z_i) \otimes h_i + z_i \otimes s_{i-1} \tag{7}$$

In the above equations, V and W are parameter matrices, and b is a parameter vector. $\otimes$ represents element-wise multiplication.

3.3.2. Local dependencies

The module is to model self dependencies. As the contexts of the test utterance are divided into four aspects, we employ four separate local memory networks (*GRU*^{L} cells) to model these different aspects into local storage cells. The computations of these aspects are as follows.

For *ownHis* aspect, the utterances in $U_{ownHis}^{N_{ownHis}}$ are framed as a sequence (starting from the oldest one) and fed to the local memory network (*GRU*^{L} cells for *ownHis* aspect). The output is marked as $M_{ownHis}^{(L)}$. For *otherHis* aspect, the utterances in $U_{otherHis}^{N_{otherHis}}$ are framed as a sequence (starting from the oldest one) and fed to the local memory network (*GRU*^{L} cells for *otherHis* aspect). The output is marked as $M_{otherHis}^{(L)}$. For *ownFut* aspect, the utterances in $U_{ownFut}^{N_{ownFut}}$ are framed as a sequence (starting from the farthest one) and fed to the local memory network (*GRU*^{L} cells for *ownFut* aspect). The output is marked as $M_{ownFut}^{(L)}$. For *otherFut* aspect, the utterances in $U_{otherFut}^{N_{otherFut}}$ are framed as a sequence (starting from the farthest one) and fed to the local memory network (*GRU*^{L} cells for *otherFut* aspect). The output is marked as $M_{otherFut}^{(L)}$.

3.3.3. Global dependencies

The module is to capture inter-personal dependencies. We use two different global memory networks (*GRU*^{G} cells) to incorporate the inter-personal dependencies of both historical contexts and future contexts. For *ownHis* and *otherHis* aspects of historical contexts, the outputs of their corresponding local memory networks are sorted based on time occurrence order. The sorted sequence is sent to global memory network (*GRU*^{G} cells for historical contexts), which is to incorporate the inter-personal dependencies of historical contexts. The output of this global memory network is marked as $M_{gHis}^{(1)}$. For *ownFut* and *otherFut* aspects of future contexts, the outputs of their corresponding local memory networks are sorted in reverse chronological order. The sorted sequence is sent to global memory network (*GRU*^{G} cells for future contexts), which is to incorporate the inter-personal dependencies of future contexts. The output of this global memory network is marked as $M_{gFut}^{(1)}$.

3.3.4. Multi-hop storage

Several recent works [35,40] consider that multiple read/write iterations are important for performing transitive inference. Multiple hops can help in improving the focus of attention heads that may miss essential memories in a single hop. Inspired by this, we conduct a series of H storage read/write cycles that are combined with an attention mechanism.

At the h^{th} hop, to find the relevance of each storage cell with the test utterance u_t, an attention mechanism is respectively performed on the storages $M_{gHis}^{(h)}$ and $M_{gFut}^{(h)}$. The calculations are:

$$W_{gHis}^{(h)} = \textit{softmax}((M_{gHis}^{(h)})^T \cdot q_t^{(h)}) \tag{8}$$

$$W_{gFut}^{(h)} = \textit{softmax}((M_{gFut}^{(h)})^T \cdot q_t^{(h)}) \tag{9}$$

Where, $\textit{softmax}(x_j) = e^{x_j} / \sum_k e^{x_k}$. Initially, $q_t^{(1)} = B \cdot u_t$ (e.g., $u_t = u_{16}$, $q_t^{(1)} = q_{16}^{(1)}$ in Fig. 2).

The attention vector $W_{gHis}^{(h)}$ is a probability distribution over the storage $M_{gHis}^{(h)}$. The attention vector $W_{gFut}^{(h)}$ is a probability distribution over the storage $M_{gFut}^{(h)}$. The j^{th} normalized score of $W_{gHis}^{(h)}$ or $W_{gFut}^{(h)}$ can indicate the relevance of j^{th} storage cell with respect to the test utterance. These storages and their attention vectors are then used to find weighted storage representations with an inner product operation as follow:

$$weightedM_{gHis}^{(h)} = M_{gHis}^{(h)} \cdot W_{gHis}^{(h)} \tag{10}$$

$$weightedM_{gFut}^{(h)} = M_{gFut}^{(h)} \cdot W_{gFut}^{(h)} \tag{11}$$

Each representation contains a weighted contextual summary that is accumulated from the corresponding storage. We add these two weighted storage representations $weightedM_{gHis}^{(h)}$ and $weightedM_{gFut}^{(h)}$ to test utterance $q_t^{(h)}$ as:

$$q_t^{(h+1)} = q_t^{(h)} + weightedM_{gHis}^{(h)} + weightedM_{gFut}^{(h)} \tag{12}$$

Futher, we employ the two iterative memory networks (GRU^M cells), to update the storages for the next hop. These two networks take the h^{th} storages $weightedM_{gHis}^{(h)}$ and $weightedM_{gFut}^{(h)}$ as inputs respectively. The computations are:

$$weightedM_{gHis}^{(h+1)} = GRU^M(weightedM_{gHis}^{(h)}) \tag{13}$$

$$weightedM_{gFut}^{(h+1)} = GRU^M(weightedM_{gFut}^{(h)}) \tag{14}$$

After H hops, the final representation $q_t^{(H+1)}$ is sent to the final prediction.

3.3.5. Final prediction

In this step, a softmax function is used to calculate emotion-class probabilities from the final emotion representation $q_t^{(H+1)}$ of utterance u_t using the equation as:

$$P_t = softmax(W_{smax} \cdot (q_t^{(H+1)}) + b_{smax}) \tag{15}$$

Then, the predicted label for utterance u_t is picked as:

$$\hat{y} = \arg\max_a(P_t[a]) \tag{16}$$

For classification training, categorical cross-entropy loss along with L2 regularization is used as the cost measure:

$$Loss = \frac{-1}{\sum_{k=1}^{N} c(k)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} (y_{i,j}) \log_2 P_{i,j} + \delta \parallel \theta \parallel_2 \tag{17}$$

Here, N is the total number of dialogues, $c(i)$ is the number of utterances in the dialogue i, $P_{i,j}$ is the probability distribution of emotion labels for utterance j of dialogue i, $y_{i,j}$ is the expected class label of utterance j of dialogue i, δ is the L2-regularizer weight, and θ is the set of trainable parameters.

Algorithm 1 summarizes the overall HMN.

4. Experiments

4.1. Dataset details

We perform experiments on dataset: IEMOCAP [59].[3] This dataset is regarded as a dialogue-based emotion detection benchmark and provides rich multimodal samples for all the utterances. There are

[3] https://sail.usc.edu/iemocap/.

Algorithm 1 Hierarchical Memory Networks

Input: u_t, $U_\lambda^{N_\lambda}$, H
Output: P_t

$q_t^{(1)} = B \cdot u_t$
for λ in [ownHis, otherHis, ownFut, otherFut] **do**
  $M_\lambda^{(L)} = GRU^L(U_\lambda^{N_\lambda})$
end for
$M_{ownHis}^{(L)}$ and $M_{otherHis}^{(L)}$ are sorted into a sequence *seqHis* based on time occurrence order
$M_{ownFut}^{(L)}$ and $M_{otherFut}^{(L)}$ are sorted into a sequence *seqFut* in reverse chronological order
$M_{gHis}^{(1)} = GRU^G(seqHis)$
$M_{gFut}^{(1)} = GRU^G(seqFut)$
for h in [1,H+1] **do**
  $W_{gHis}^{(h)} = softmax((M_{gHis}^{(h)})^T \cdot q_t^{(h)})$
  $weightedM_{gHis}^{(h)} = M_{gHis}^{(h)} \cdot W_{gHis}^{(h)}$
  $M_{gHis}^{(h+1)} = GRU^M(M_{gHis}^{(h)})$
  $W_{gFut}^{(h)} = softmax((M_{gFut}^{(h)})^T \cdot q_t^{(h)})$
  $weightedM_{gFut}^{(h)} = M_{gFut}^{(h)} \cdot W_{gFut}^{(h)}$
  $M_{gFut}^{(h+1)} = GRU^M(M_{gFut}^{(h)})$
  $q_t^{(h+1)} = q_t^{(h)} + weightedM_{gHis}^{(h)} + weightedM_{gFut}^{(h)}$
end for
return $P_t = softmax(W_{smax} \cdot (q_t^{(H+1)}) + b_{smax})$

12 hours of dyadic conversational videos, which are grouped into five sessions. These sessions are split into 5 minutes of interaction between pairs of 10 unique speakers (5 male and 5 female), with each pair assigned to diverse multiple conversation scenarios for dialogues. All the conversations are segmented into spoken utterances, with each utterance being tagged by at least 3 annotators with one of the labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. In this study, we consider the first four categories with majority agreement (i.e., at least two out of three annotators labeled the same emotion) to compare with the baseline models. Readers can refer to the literature [59,60] for obtaining a more detailed description of the dataset.

4.2. Training details

As each individual has a unique way of expressing emotions, it is necessary to perform person independent experiments to test our method. The IEMOCAP dataset is partitioned into train and test sets with roughly 80/20 ratio such that the partitions do not share any speaker. For hyper-parameters tuning, 20% of the train set is further used as a validation set. For optimization of the parameters, the Adam [61] optimizer is used to train our network. Through monitoring the validation loss, the training phase stops with the patience of 10 epochs. The Dropout [62] and Gradient-clipping for a norm of 40 are used to regular the network. Hyper-parameters are decided using random search [63]. Finally, step length of N_{ownHis}, $N_{otherHis}$, N_{ownFut} and $N_{otherFut}$ are respectively set to be 28, 17, 13, and 6. The number of hops H is fixed at 3. The embedding dimension size R^d is set as 50.

4.3. Baselines

We compare HMN with the following baseline methods.

– *Tripathi et al.* [64] exploits the effectiveness of neural networks to perform MER on the IEMOCAP

Table 2
Comparison of HMN in terms of accuracy with the baseline methods for IEMOCAP dataset

Models	Accuracy
Tripathi et al. [64]	71.04
bc-LSTM [29]	76.1
CHFusion [23]	76.8
CMN [4]	77.6
HMN	**79.64**

dataset using data from speech, text, and motions captured from facial expressions, rotation, and hand movements. Their neural networks are adept at estimating complex functions that depend on a large number and diverse source of input data.

- *bc-LSTM* [29] has two unidirectional LSTMs stacked together with opposite directions that are used to model the contexts from the neighboring utterances into the utterance representation, making each utterance can get information from utterances occurring before and after itself in the video.
- *CHFusion* [23] presents a hierarchical feature fusion strategy that first fuses the modalities two in two and only then fuses all three modalities. In their method, RNN, specifically GRU, is alternately used to extract context-aware utterance features in the processes of bimodal combination and trimodal combination.
- *CMN* [4] uses two distinct GRUs for two speakers to capture the separate contexts from dialogue histories. These contexts are stored as memories after performing an attention mechanism. These memories are merged with the test utterance using an addition operation. Further, the multiple hops operation is used to improve the merged representation.

However, bc-LSTM, Tripathi et al., and CHFusion do not distinguish the speakers in the conversations; CMN does not consider the weights of different aspects of contexts.

4.4. Results

For the performance comparison, we perform experiments on the IEMOCAP test set. Table 2 summarizes the results. The results of other methods in the table are all from the literature. We use Accuracy to evaluate classification performance. As shown in Table 2, our model HMN outperforms the compared models with significant classification performance increase in Accuracy ranging from 2.0% to 8.6%. It suggests that the HMN model is more capable of capturing the contextual information from surrounding utterances using our proposed modeling scheme.

5. Discussion and analysis

5.1. Hyperparameters

To avoid a combinatorial explosion, we consider a random search to set the values of N_{ownHis}, $N_{otherHis}$, N_{ownFut} and $N_{otherFut}$. The experiments verify our hypothesis that the surrounding contexts of a test utterance u_t should be differentiated. Based on the experimental results, we set the value of $N_{ownHis} = 28$, $N_{otherHis} = 17$, $N_{ownFut} = 13$, and $N_{otherFut} = 6$. When fixing these values, we concern the performance trends of HMN on the IEMOCAP dataset with two hyperparameters, H (number of hops), and R^d (embedding dimension size). Table 3 provides a summary of the performance trend of our model for

Table 3
Performance trends of HMN with different number of hops

H	Accuracy	F1
1	77.94	77.86
2	75.72	75.62
3	**79.64**	**79.64**
4	74.97	74.85
5	78.26	78.22
6	70.52	69.99
7	61.82	58.14
8	69.25	68.71
9	63.20	63.10
10	60.66	60.72
11	52.70	48.23
12	53.55	49.39
13	49.20	44.96
14	57.16	53.79
15	54.08	49.42
16	41.25	36.28

Table 4
Performance trends of HMN with different embedding dimension sizes

R^d	Accuracy	F1
30	73.17	72.43
40	74.87	74.62
50	**79.64**	**79.64**
60	78.79	78.60
70	75.40	75.20
80	78.47	78.48

Table 5
Comparison of the performance of HMN on IEMOCAP in different modalities

Modality	Accuracy	F1
Unimodal text (T)	75.50	75.32
Unimodal audio (A)	64.58	64.60
Unimodal visual (V)	44.43	44.13
Trimodal (T+A+V)	**79.64**	**79.64**

different values of the hyperparameter H. With H increasing, the performance initially improves because more weighted storages are added to the test utterance. However, with hopping recurrence deepens, the total parameters also increase, and the overfitting happens. We obtain the best performance at $H = 3$ in our experiments. In Table 4, the similar trends are observed where the performance initially improves by increasing the embedding dimension size R^d. However, with a further increase of this embedding dimension size, the performance decrease for saturation. Finally, the best performance is obtained at $R^d = 50$ in our experiments.

5.2. Importance of the modalities

As we know, the multimodal analysis outperforms the unimodal analysis, which has been already

established in the literature [65,66]. We also observe the same trend in our experiments. The textual modality is the strongest individual modality and performs best among others, which is aligned with our expectations. Although other modalities contribute to improving the performance of multimodal classifiers, the contribution is little compared to the textual modality. All above is reaffirming the significance of textual modality in multimodal systems. Table 5 summarizes the comparison of the performance of HMN on IEMOCAP in different modalities.

6. Conclusion

In this work, we have developed HMN for MER in dyadic conversational videos. Our HMN is to distinguish the contexts of a test utterance and uses different step lengths to represent the weights of different aspects of contexts. Besides, HMN leverages hierarchical memory networks to integrate these aspects so that it can capture the self and inter-personal dependencies. With the attention mechanism and multiple hops mechanism applied to HMN, adequate contexts can be generated, thus improving the representation of the test utterance and helping to boost accuracy performance. Our experimental results confirm the performance of HMN with such a modeling scheme. In the future, we plan to explore the setting of the values of the step lengths in different aspects and to improve the quality of multimodal feature data.

Acknowledgments

This work was supported in part by the Guangdong Natural Science Foundation under Grant 2018A030313746, in part by the Youth Innovative Talents Project in Guangdong Universities (2020KQNCX186), in part by the Special Innovation Project in Guangdong Universities (2020KTSCX273), and in part by the Fourth College Level Project of Guangdong Justice Police Vocational College (2020YB16).

References

[1] C. Argueta, F.H. Calderon and Y.-S. Chen, Multilingual emotion classifier using unsupervised pattern extraction from microblog data, *Intelligent Data Analysis* **20**(6) (2016), 1477–1502.
[2] F. Zhang, H. Xu and X. Bai, On the need of hierarchical emotion classification: detecting the implicit feature using constrained topic model, *Intelligent Data Analysis* **21**(6) (2017), 1393–1406.
[3] V.-V. Vu, H.-Q. Do, V.-T. Dang and N.-T. Do, An efficient density-based clustering with side information and active learning: a case study for facial expression recognition task, *Intelligent Data Analysis* **23**(1) (2019), 227–240.
[4] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency and R. Zimmermann, 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, Vol. 1 (Long Papers)*, 2018, pp. 2122–2132.
[5] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh and E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 6818–6825.
[6] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas and M. Huang, Augmenting end-to-end dialogue systems with commonsense knowledge, in: *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[7] J. Sidnell and T. Stivers, *The handbook of conversation analysis*, Vol. 121, John Wiley & Sons, 2012.
[8] S. Poria, E. Cambria, R. Bajpai and A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, *Information Fusion* **37** (2017), 98–125.

[9] H. Zhou, M. Huang, T. Zhang, X. Zhu and B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[10] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria and R. Zimmermann, ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.
[11] M.W. Morris and D. Keltner, How emotions work: the social functions of emotional expression in negotiations, *Research in Organizational Behavior* **22** (2000), 1–50.
[12] F. Liu and S. Maitlis, Emotional dynamics and strategizing processes: a study of strategic conversations in top team meetings, *Journal of Management Studies* **51**(2) (2014), 202–234.
[13] P. Koval and P. Kuppens, Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia, *Emotion* **12**(2) (2012), 256.
[14] P. Kuppens, N.B. Allen and L.B. Sheeber, Emotional inertia and psychological maladjustment, *Psychological Science* **21**(7) (2010), 984–991.
[15] C. Navarretta, Mirroring facial expressions and emotions in dyadic conversations, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 469–474.
[16] R.W. Picard, Affective computing: from laughter to IEEE, *IEEE Transactions on Affective Computing* **1**(1) (2010), 11–17.
[17] P. Ekman, Facial expression and emotion, *American Psychologist* **48**(4) (1993), 384.
[18] D. Datcu and L.J.M. Rothkrantz, *Semantic audio-visual data fusion for automatic emotion recognition*, 2008.
[19] D. Datcu and L.J. Rothkrantz, Emotion recognition using bimodal data fusion, in: *Proceedings of the 12th International Conference on Computer Systems and Technologies*, ACM, 2011, pp. 122–128.
[20] C.O. Alm, D. Roth and R. Sproat, 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*, Association for Computational Linguistics, 2005, pp. 579–586.
[21] C. Strapparava and R. Mihalcea, Annotating and identifying emotions in text, in: *Intelligent Information Access*, Springer, 2010, pp. 21–38.
[22] T. Baltrušaitis, C. Ahuja and L.-P. Morency, Multimodal machine learning: a survey and taxonomy, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2) (2018), 423–443.
[23] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria and S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, *Knowledge-Based Systems* **161** (2018), 124–133.
[24] M. Soleymani, M. Pantic and T. Pun, Multimodal emotion recognition in response to videos, *IEEE Transactions on Affective Computing* **3**(2) (2011), 211–223.
[25] A. Zadeh, M. Chen, S. Poria, E. Cambria and L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, *arXiv preprint arXiv:1707.07250*, 2017.
[26] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh and L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ACM, 2017, pp. 163–171.
[27] S. Poria, E. Cambria, G. Winterstein and G.-B. Huang, Sentic patterns: Dependency-based rules for concept-level sentiment analysis, *Knowledge-Based Systems* **69** (2014), 45–63.
[28] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria and L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[29] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh and L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, 2017, pp. 873–883.
[30] E. Owusu, Y. Zhan and Q.R. Mao, An SVM-AdaBoost facial expression recognition system, *Applied Intelligence* **40**(3) (2014), 536–545.
[31] Y. Liu, L. Ji, R. Huang, T. Ming, C. Gao and J. Zhang, An attention-gated convolutional neural network for sentence classification, *Intelligent Data Analysis* **23**(5) (2019), 1091–1107.
[32] H.-Y. Lu, N. Kang, Y. Li, Q.-Y. Zhan, J.-Y. Xie and C.-J. Wang, Utilizing Recurrent Neural Network for topic discovery in short text scenarios, *Intelligent Data Analysis* **23**(2) (2019), 259–277.
[33] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic and C. Pal, Recurrent neural networks for emotion recognition in video, in: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 467–474.
[34] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria and L.-P. Morency, Memory fusion network for multi-view sequential learning, in: *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[35] J. Weston, S. Chopra and A. Bordes, Memory networks, *arXiv preprint arXiv:1410.3916*, 2014.
[36] A. Graves, G. Wayne and I. Danihelka, Neural turing machines, *arXiv preprint arXiv:1410.5401*, 2014.
[37] T. Young, D. Hazarika, S. Poria and E. Cambria, Recent trends in deep learning based natural language processing, *Ieee Computational Intelligen Ce Magazine* **13**(3) (2018), 55–75.

[38] D. Kundu and D.P. Mandal, Formulation of a hybrid expertise retrieval system in community question answering services, *Applied Intelligence* **49**(2) (2019), 463–477.
[39] C. Fu, User correlation model for question recommendation in community question answering, *Applied Intelligence* **50**(2) (2020), 634–645.
[40] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus and R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: *International Conference on Machine Learning*, 2016, pp. 1378–1387.
[41] S. Poria, A. Gelbukh, E. Cambria, A. Hussain and G.-B. Huang, EmoSenticSpace: A novel framework for affective common-sense reasoning, *Knowledge-Based Systems* **69** (2014), 108–123.
[42] T. Daybelge and I. Cicekli, A ranking method for example based machine translation results by learning from user feedback, *Applied Intelligence* **35**(2) (2011), 296–321.
[43] J. Sangeetha and S. Jothilakshmi, Speech translation system for english to dravidian languages, *Applied Intelligence* **46**(3) (2017), 534–550.
[44] R. Kar, A. Konar, A. Chakraborty, B.S. Bhattacharya and A.K. Nagar, EEG source localization by memory network analysis of subjects engaged in perceiving emotions from facial expressions, in: *2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2015, pp. 1–8.
[45] J. Ruusuvuori, 16 Emotion, Affect and Conversation, *The handbook of conversation analysis*, 2013, 330.
[46] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria and R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, *arXiv preprint arXiv:1810.02508*, 2018.
[47] H. Rashkin, E.M. Smith, M. Li and Y.-L. Boureau, I know the feeling: Learning to converse with empathy, 2018.
[48] J.M. Richards, E.A. Butler and J.J. Gross, Emotion regulation in romantic relationships: The cognitive consequences of concealing feelings, *Journal of Social and Personal Relationships* **20**(5) (2003), 599–620.
[49] S. Hareli and A. Rafaeli, Emotion cycles: on the social influence of emotion in organizations, *Research in Organizational Behavior* **28** (2008), 35–59.
[50] L. Yang, f.L. Hong and W. GUO, Textbased emotion transformation analysis, *Computer Engineering & Science* **9** (2011), 026.
[51] P. Xiaolan, X. Lun, L. Xin and W. Zhiliang, Emotional state transition model based on stimulus and personality characteristics, *China Communications* **10**(6) (2013), 146–155.
[52] M.-C. Sun, S.-H. Hsu, M.-C. Yang and J.-H. Chien, Context-aware cascade attention-based RNN for video emotion recognition, in: *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, IEEE, 2018, pp. 1–6.
[53] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller and S. Narayanan, Context-sensitive learning for enhanced audiovisual emotion classification, *IEEE Transactions on Affective Computing* **3**(2) (2012), 184–198.
[54] Y. Kim, Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882*, 2014.
[55] F. Eyben, M. Wöllmer and B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: *Proceedings of the 18th ACM International Conference on Multimedia*, ACM, 2010, pp. 1459–1462.
[56] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.
[57] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, *arXiv preprint arXiv:1406.1078*, 2014.
[58] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation* **9**(8) (1997), 1735–1780.
[59] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee and S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, *Language Resources and Evaluation* **42**(4) (2008), 335.
[60] E. Cambria, D. Hazarika, S. Poria, A. Hussain and R. Subramanyam, Benchmarking multimodal sentiment analysis, in: *International Conference on Computational Linguistics and Intelligent Text Processing*, Springer, 2017, pp. 166–179.
[61] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, 2014.
[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research* **15**(1) (2014), 1929–1958.
[63] J. Bergstra and Y. Bengio, Random search for hyper-parameter optimization, *Journal of Machine Learning Research* **13**(Feb) (2012), 281–305.
[64] S. Tripathi, S. Tripathi and H. Beigi, Multi-modal emotion recognition on iemocap dataset using deep learning, *arXiv preprint arXiv:1804.05788*, 2018.
[65] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: *2016 IEEE 16th International Conference on Data Mining (ICDM)*, IEEE, 2016, pp. 439–448.
[66] V. Pérez-Rosas, R. Mihalcea and L.-P. Morency, Utterance-level multimodal sentiment analysis, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 973–982.