Information Fusion 88 (2022) 296–304

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Deep Emotional Arousal Network for Multimodal Sentiment Analysis and Emotion Recognition

Feng Zhang [a], Xi-Cheng Li [a], Chee Peng Lim [b], Qiang Hua [a], Chun-Ru Dong [a,*], Jun-Hai Zhai [a]

[a] Hebei Key Laboratory of Machine Learning and Computational Intelligence College of Mathematics and Information Science, Hebei University, Baoding, P.R. China
[b] Institute for Intelligent Systems Research and Innovation, Deakin University, Australia

ARTICLE INFO

Keywords:
Deep Emotional Arousal Network
Multimodal Fusion Strategy
Multimodal Sentiment Analysis
Multimodal Emotion Recognition

ABSTRACT

Multimodal sentiment analysis and emotion recognition has become an increasingly popular research area, where the biggest challenge is to efficiently fuse the input information from different modality. The recent success is largely credited to the attention-based models, e.g., transformer and its variants. However, the attention-based mechanism often neglects the coherency of human emotion due to its parallel structure. Inspired by the emotional arousal model in cognitive science, a Deep Emotional Arousal Network (DEAN) that is capable of simulating the emotional coherence is proposed in this paper, which incorporates the time dependence into the parallel structure of transformer. The proposed DEAN model consists of three components, i.e., a cross-modal transformer is devised to simulate the functions of perception analysis system of humans; a multimodal BiLSTM system is developed to imitate the cognitive comparator, and a multimodal gating block is introduced to mimic the activation mechanism in human emotional arousal model. We perform extensive comparison and ablation studies on three benchmarks for multimodal sentiment analysis and emotion recognition. The empirical results indicate that DEAN achieves state-of-the-art performance, and useful insights are derived from the results.

1. Introduction

Human emotions are controlled by neuronal circuits, which collect emotional information and generate emotional behaviors through physiological arousal [1]. In most communication scenario, people usually need to extend language-based sentiment analysis and emotion recognition to multimodal settings [2,3].

Transformer (Vaswani et al., 2017) and its variants ([4]; Lan et al., 2019) have become increasingly popular for multimodal sentiment analysis and emotion recognition recently [5,6]. These attention-based methods can model the global dependency between every two utterances directly and can be implemented in a parallel structure as well. Therefore, it is powerful to deal with the dependency in sequence with a large scale by reducing the constraint of sequential computation. However, three main challenges remain:

(a) the attention-based fusion strategy cannot model the coherence of emotions due to its parallel structure. In general, current human emotion is usually affected by past emotional memory. As shown in Fig. 1, the positive emotions of narrator are getting more and more explicit over time from successive fragment;

(b) the existing attention-based models usually neglect the distinction of different modalities by simple concatenation ([5], 2020), while the experimental results of previous study [7] indicate that different modalities have different influences on classification results, and language modality often has greater influence than visual and audio.

More importantly, the accuracy of most multimodal models depends heavily on the fusion strategy. Therefore, instead of putting great efforts to the fusion strategy, it is essential to explore an integral framework by simulating human communication with multimodal inputs.

Inspired by the emotional arousal model in psychology [8], a Deep Emotional Arousal Network (DEAN) is proposed in this paper to deal with the aforementioned challenges. DEAN consists of three components: a Cross-modal Transformer, a Multimodal BiLSTM System and a Multimodal Gating Block. These components are designed to simulate the functions of perception analysis system, cognitive comparator, and activation mechanism in the psychological emotional arousal in humans, respectively (see Fig. 2).

The Cross-modal Transformer explores the spatial interaction between modalities by employing an improved self-attention mechanism

* Corresponding author.
E-mail address: 35076322@qq.com (C.-R. Dong).

https://doi.org/10.1016/j.inffus.2022.07.006
Received 25 May 2021; Received in revised form 20 December 2021; Accepted 19 July 2022
Available online 20 July 2022

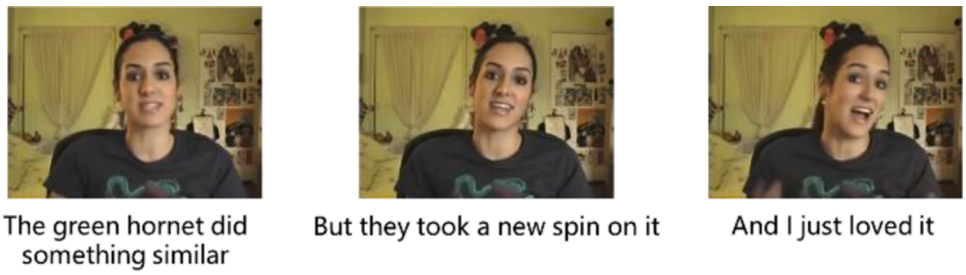

Fig. 1. Continuous video clips in MOSI dataset

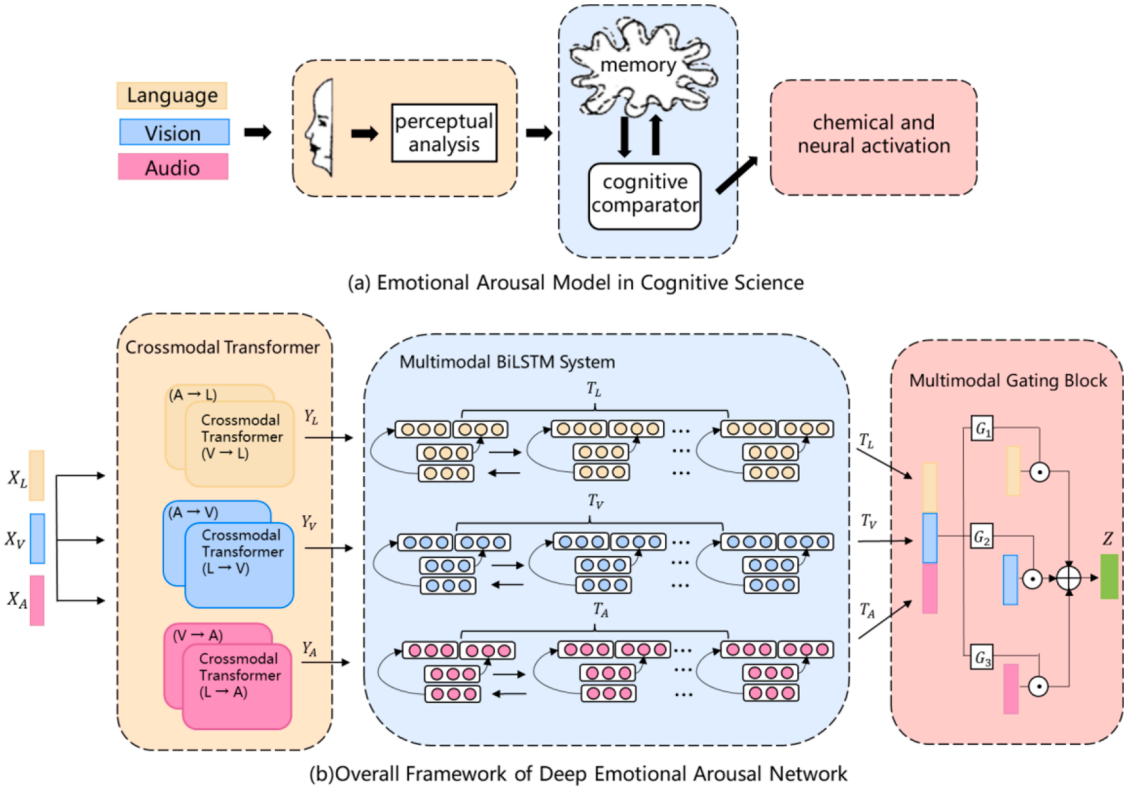

Fig. 2. (a) The psychological emotional arousal model consists of three subsystems: a perceptual analysis system, a cognitive comparator, and an activation mechanism. (b) Deep Emotional Arousal Network (DEAN) is constitutive of three modules: a Multimodal Transformer, a Multimodal BiLSTM System and a Multimodal Gating Block, which are used to mimic the function of the subsystems in the psychological emotional arousal model in (a).

with cross talking-heads attention. The Multimodal BiLSTM System models the coherence of emotions by exploiting a Bidirectional LSTM (Long Short-Term Memory) network, which enables DEAN to capture the temporal interaction among modalities. The Multimodal Gating Block implicitly performs the fusion of multimodal information by adaptively controlling the output of the gated systems. DEAN tries to provide an integral framework and an alternative idea of guiding the learning system along a human-like path that leads to the progressive acquisition of complex understanding of human emotions.

For evaluation, we conduct extensive experiments and ablation studies using the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets for multimodal sentiment analysis and emotion recognition. The experimental results indicate that DEAN achieves state-of-the-art performances on these benchmark datasets.

2. Related Studies

The purpose of multimodal sentiment analysis and emotion recognition is to predict the sentiment or emotion label of each multimodal input. The vital challenge lies in the fusion strategies of multimodal inputs, which can be achieved in model-agnostic or model-based methods. The model-agnostic fusion methods include early, late and hybrid fusion, without involving specific classifiers or regression models. On the contrary, model-based methods address multimodal fusion problem with fusion model construction.

Model-agnostic fusion methods: The model-agnostic fusion methods can be divided into early, late and hybrid fusion strategies according to the way multimodal inputs are fused. The early fusion strategy, also called feature-level fusion, usually relies on generic models to learn the representative features and then simply integrates the extracted features by concatenation or weighted combination. Due to the powerful representation capability of deep learning, the recent proposed early fusion strategies usually employ Convolutional Neural Networks (CNNs) [9–11], LSTM models [12,13] or Recurrent Neural Networks (RNNs) [14,15] for feature extraction.

However, these early fusion strategies are ineffective in tackling the intra-modality dynamics. In addition, deep learning-based early fusion strategies tend to suffer from the overfitting problem due to their

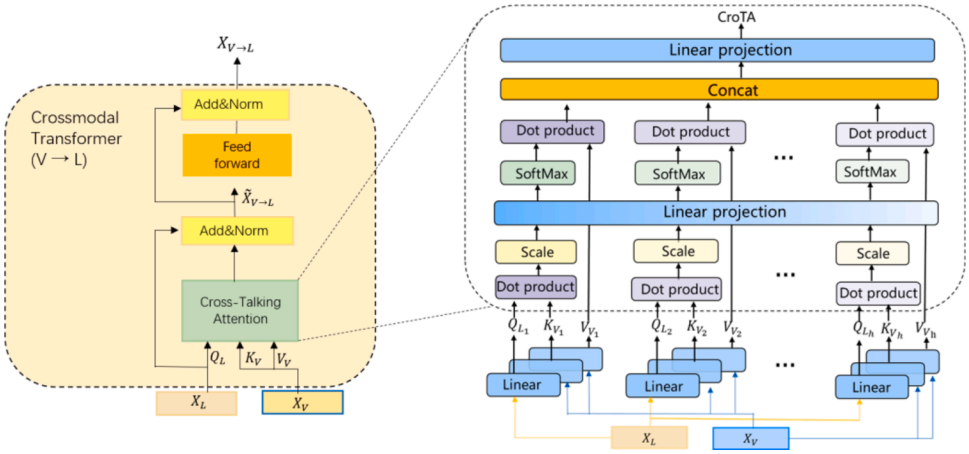

Fig. 3. Cross-modal Transformer of $V \rightarrow L$

massive network structures. The late fusion strategies independently devise one classifier for each modality and aggregate the outputs of each classifier by averaging, weighted sum or voting [2,16]. Hybrid fusion strategies [17], which combine the advantages of early fusion and later fusion based on each unimodal prediction, normally outperform early or later fusion counterparts [18]. Nevertheless, all these fusion strategies lack the ability to model the inter-modality dynamics since the dynamics behind fusion strategies are far more complex than a decision vote. Therefore, fusion strategies remain a major challenge for sentiment analysis and emotion recognition.

Model-based fusion methods: Earlier examples of model-based fusion methods include Multiple Kernel Learning (Gönen and Ethem, 2011), Bilinear Fusion [19] and Graphical Models [20]. Recent model-based fusion methods mainly include:

(1) Tensor-based fusion, with the representative models of Tensor Fusion Network (TFN) [21], Low-rank Multimodal Fusion (LMF) [2] and Locally confined modality fusion network [22]; **(2)** Translation-based fusion, with examples include Modality Translation Model (MCTN) [23] and Seq2Seq Modality Translation Model (SSMT) [24]; **(3)** Attention-based fusion, which exploits various attention mechanism components to fuse modalities. As an example, the Multi-attention Recurrent Network (MARN) (Zadeh et al., 2018) models interactions between modalities using a Multi-Attention Block and stores them in a hybrid memory. The Multimodal Transformer (MulT) [5] merges multimodal information via a feedforward fusion process from multiple directional modality transformers. The Recurrent Attended Variation Embedding Network (RAVEN) [25] builds human language by shifting word representations based on the nonverbal behavior's patterns. Detailed experimental results [7] show that attention-based fusion methods improve performance for sentiment analysis and emotion recognition tasks as compared with other model-based fusion methods. The reason is that attention-based fusion methods can implicitly model the inter-dynamic and intra-dynamic of different modalities.

However, most attention-based methods, which are typically modeled with Transformer [5,6,25], generally neglect the coherency of human emotions due to their parallel structure. Moreover, most existing attention-based mechanisms neglect the distinction of different modalities by simple concatenation [5].

Inspired by an emotional arousal model in psychology, a Deep Emotional Arousal Network (DEAN) is proposed in this paper. The DEAN model formulates an integral framework for multimodal sentiment analysis and emotion recognition. The components in DEAN incorporate the capability of understanding multimodalities communication, as humans do.

The advantages of DEAN are: (1) modelling the emotional coherence by introducing time-dependent interactions into the parallel structure of Transformer; (2) identifying the distinctions of different modalities by embedding a multimodal gating mechanism; and (3) providing an integral framework for human communication with multimodal information based on a human phycological model. Extensive performance evaluation experiments and ablation studies on three benchmark datasets are conducted. The experimental results show that DEAN outperforms the state-of-the-art models on these benchmark problems.

3. Deep Emotional Arousal Network

In this section, we explain the proposed DEAN model in detail. Specifically, DEAN is composed of three main components:

a Cross-modal Transformer: This module mimics the first subsystem of an established human emotional arousal model in physiology. More specifically, three pairs of cross-modal transformers (Fig 2b) are employed to model the inter- and intra-modality interaction among the modalities by leveraging attention mechanism. The modality with higher attention weight has greater importance.

b Multimodal BiLSTM System: This temporal structure is used to model the cognitive comparators in the psychological emotional arousal model. The process consists of three steps: (i) extracting the context-dependent features from each multimodality input by applying a Bidirectional LSTM network, which models the inter-dependencies among modalities with respect to time; (ii) comparing the current extracted features with those from the past memory; (iii) implementing the information transfer in time series to imitate the emotional coherence.

c Multimodal Gating Block: This gating block is adopted to imitate the functions of the activation mechanism in the psychological emotional arousal model. It is able to distinguish and fuse the information from each unimodality by controlling the output information of the target mode according to the importance of the target modality.

3.1. Cross-modal Transformer Module

Inspired by the success in natural language processing with the Transformer-based method, a Cross-modal Transformer [5] module is adopted and improved in DEAN to model the multi-channel perceptual analysis process in the human nervous system. This module implicitly fuses the multimodal inputs via a feed-forward fusion process. It is based on a pairwise cross-modal attention mechanism, which can explore the interactions between multimodal inputs and learn representations directly from aligned multimodal streams. For the unaligned multimodal inputs, a one-dimensional temporal convolutional layer is employed as a preprocessor to make the multimodal inputs aligned.

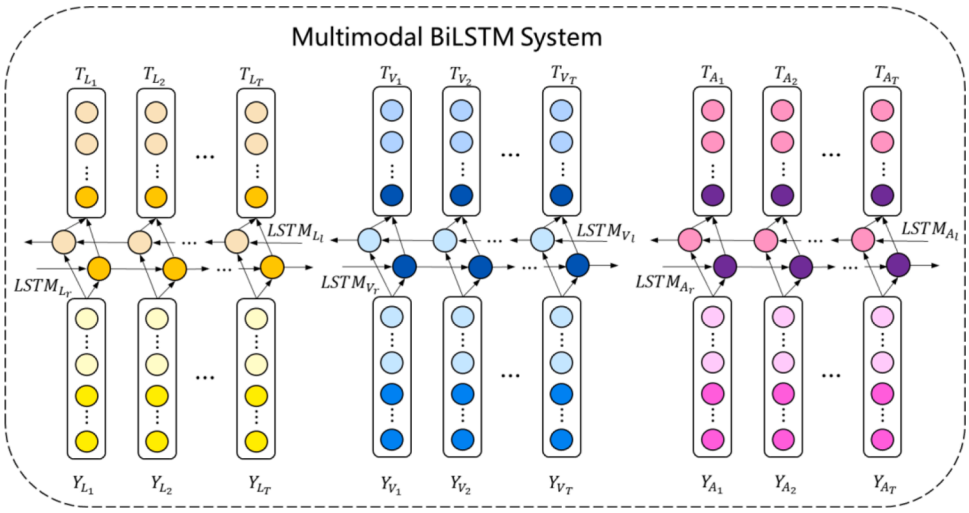

Fig. 4. Multimodal BiLSTM System

The Cross-modal Transformer of DEAN utilizes attention mechanism to enhance the target modality with other modalities at the low feature level. Due to high performance and similar computational cost to the basic attention [26], a multi-head attention is normally used in transformed-based models for multimodal sentiment analysis and emotion recognition. However, with the increasing of number of heads, the query-vectors and key-vectors become so low-dimensional that their dot product cannot constitute an informative matching function any longer. Therefore, a cross-talking attention mechanism is introduced by inserting a linear projection across the attention-heads to make each attention head depend on all the keys and queries. Since three modalities (i.e., Language, Visual and Audio) are considered in this paper, six Cross-modal Transformers Fig. 2b) are included in DEAN. Taking the Cross-modal Transformer of $V\rightarrow L$ as an example (Fig. 3), the Language modality (L) and Visual modality (V) are set as the target and auxiliary counterparts, respectively. The embedding feature of the Visual modality X_V is used to reinforce the Language modality X_L by learning the attention across the features of both L and V. As an example, a cross-talking attention for the pair of Language and Visual modality can be formulated by Eqs. (1)-(3).

$$Q_{L_i} = X_{L_i} W_{Q_i},\ K_{V_i} = X_{V_i} W_{K_i},\ V_{V_i} = X_{V_i} W_{V_i} \tag{1}$$

$$t\_head_i = softmax\left(\frac{Q_{L_i} K_{V_i}^T}{\sqrt{d_k}} W_{th_i}\right) V_{V_i} \tag{2}$$

$$CroTA = concat(t\_head_1, \ldots, t\_head_h)\ W_o \tag{3}$$

Where Q_{L_i}, K_{V_i} and V_{V_i} are the corresponding Query, Key and Value vectors for the i-th attention head, $i = 1, 2, \ldots, h$. $X_{L_i} \in *R*^{T\times d}$ and $X_{V_i} \in *R*^{T\times d}$ are the aligned input embedding features of Language and Visual modalities for attention-head i respectively; $W_{Q_i} \in *R*^{d\times d_k}$, $W_{K_i} \in *R*^{d\times d_k}$ and $W_{V_i} \in *R*^{d\times d_v}$ are the weight parameters to be learned; $\sqrt{d_k}$ is a scale factor; $W_{th} \in *R*^{T\times T\times h}$ is the parameter tensor across the attention-heads and $W_{th_i} \in *R*^{T\times T}$ is the parameter matrix of each attention-head. CroTA $\in *R*^{T\times d_v}$ is the output of the cross-talking attention, which is a linear projection of the concatenation of the outputs from all h attention heads.

In order to keep the initial information from the target modality together with the information reinforced by other modalities, a residual connection structure [27] is added after the cross-modal attention by using Eq. (4),

$$\widetilde{X}_{V\rightarrow L} = LaNorm(Q_L + CroTA(Q_L, K_V, V_V)) \tag{4}$$

where *LaNorm* denotes layer normalization [28].

The fusion feature of $\widetilde{X}_{V\rightarrow L}$, therefore, contains information from both the target modal and its enforced supplement provided by other modalities. To learn more meaningful interactions across modalities, $\widetilde{X}_{V\rightarrow L}$ is used as an input into a feedforward network *FFN*. It is augmented by its residual to yield the output of the Cross-modal Transformer of $V\rightarrow L$, i.e., $X_{V\rightarrow L}$, by using Eqs. (5) and (6),

$$X_{V\rightarrow L} = LaNorm(\widetilde{X}_{V\rightarrow L} + FFN(\widetilde{X}_{V\rightarrow L})) \tag{5}$$

$$FFN = W_2(ReLU(W_1(\widetilde{X}_{V\rightarrow L}) + b_1)) + b_2 \tag{6}$$

where $X_{V\rightarrow L} \in *R*^{T\times d_v}$.

In a similar way, the output of Cross-modal Transformer of $A\rightarrow L$ is obtained, which is denoted as $X_{A\rightarrow L}$. Then, the output of the Language modality with its overall interactions from other modalities is defined by Eq. (7).

$$Y_L = Concat(X_{V\rightarrow L}, X_{A\rightarrow L}) \tag{7}$$

Each pair of Cross-modal Transformers is used to model the interactions between different modalities, respectively. Therefore, the output of each pairwise fusion result pertaining to a different target modality is obtained by Eqs. (8) and (9).

$$Y_V = Concat(X_{L\rightarrow V}, X_{A\rightarrow V}) \tag{8}$$

$$Y_A = Concat(X_{L\rightarrow A}, X_{V\rightarrow A}) \tag{9}$$

where $Y_L \in *R*^{T\times 2d_v}$, $Y_V \in *R*^{T\times 2d_v}$ and $Y_A \in *R*^{T\times 2d_v}$ denote the output of different Cross-modal Transformer after being reinforced by other auxiliary modalities, respectively.

3.2. Multimodal BiLSTM System

Human emotion is coherent in time. The emotional change at the current moment is often affected by its past emotional memory. Due to the strength of the LSTM network in capturing long-distance semantic dependency, it is used to imitate the function of comparators in the psychological emotional arousal model. In DEAN, its Multimodal BiLSTM System (Fig. 4) can learn the sequential pattern of each modality by leveraging both forward and backward semantic dependency from training samples. As such, it is able to capture the coherence of emotion by amplifying the contribution of crucial factors in memory.

Let Y_t^m be the output of the m-th Cross-modal Transformer at time t, where $m \in$ {Language, Visual, Audio}. Y_t^m goes through the input gate,

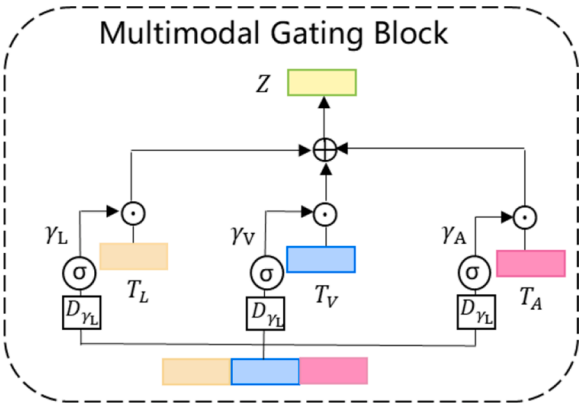

Fig. 5. Multimodal Gating Block

forgetting gate and output gate of LSTM in a bidirectional way. It is compared with T^m_{t-1} in memory to yield the output, T^m_t. The outputs of LSTM in both forward and backward manners at time t are concatenated to form the output of the BiLSTM module. The output of LSTM is updated by Eqs. (10)-(15).

$$i^m_t = \sigma\left(W^m_i\left[T^m_{t-1}, Y^m_t\right] + b^m_i\right) \tag{10}$$

$$\mathrm{f}^{\mathrm{m}}_{\mathrm{t}} = \sigma\left(\mathrm{W}^{\mathrm{m}}_{\mathrm{f}}\left[\mathrm{T}^{\mathrm{m}}_{\mathrm{t-1}}, \mathrm{Y}^{\mathrm{m}}_{\mathrm{t}}\right] + \mathrm{b}^{\mathrm{m}}_{\mathrm{f}}\right) \tag{11}$$

$$\mathrm{o}^{\mathrm{m}}_{\mathrm{t}} = \sigma\left(\mathrm{W}^{\mathrm{m}}_{\mathrm{o}}\left[\mathrm{T}^{\mathrm{m}}_{\mathrm{t-1}}, \mathrm{Y}^{\mathrm{m}}_{\mathrm{t}}\right] + \mathrm{b}^{\mathrm{m}}_{\mathrm{o}}\right) \tag{12}$$

$$\widetilde{\mathrm{c}}^{\mathrm{m}}_{\mathrm{t}} = \tanh\left(\mathrm{W}^{\mathrm{m}}_{\widetilde{\mathrm{c}}}\left[\mathrm{T}^{\mathrm{m}}_{\mathrm{t-1}}, \mathrm{Y}^{\mathrm{m}}_{\mathrm{t}}\right] + \mathrm{b}^{\mathrm{m}}_{\widetilde{\mathrm{c}}}\right) \tag{13}$$

$$\mathrm{c}^{\mathrm{m}}_{\mathrm{t}} = \mathrm{f}^{\mathrm{m}}_{\mathrm{t}} \odot \mathrm{c}^{\mathrm{m}}_{\mathrm{t-1}} + \mathrm{i}^{\mathrm{m}}_{\mathrm{t}} \odot \widetilde{\mathrm{c}}^{\mathrm{m}}_{\mathrm{t}} \tag{14}$$

$$\mathrm{h}^{\mathrm{m}}_{\mathrm{t}} = \mathrm{o}^{\mathrm{m}}_{\mathrm{t}} \odot \tanh\left(\mathrm{c}^{\mathrm{m}}_{\mathrm{t}}\right) \tag{15}$$

where i^m_t, f^m_t and o^m_t denote the input gate, forgetting gate and output gate of LSTM for the *m-th* modality at time t, respectively, W^m_i, W^m_f, W^m_o, $W^m_{\widetilde{c}}$ are their corresponding weight matrices, $\odot$ denotes the Hadamard product (element-wise product), and σ is the sigmoid activation function. The output of the multimodal BiLSTM module is denoted by $T_L \in *R*^{T\times d_L}$, $T_V \in *R*^{T\times d_V}$, $T_A \in *R*^{T\times d_A}$ respectively. It concatenates the outputs of the forward and backward LSTM models, which guarantees the temporal interaction of multimodalities. The input of BiLSTM for each modality comes from the output of its corresponding Cross-modal Transformer at timestep t, which preserves the spatial interaction of different modalities. As a result, the Multimodal BiLSTM System captures both the dynamics of intra-modality and inter-modality from the spatial and temporal viewpoints.

3.3. Multimodal Gating Block

A Multimodal Gating Block (Fig. 5) is constructed to model the activation mechanism of the human emotional arousal model. This block reinforces the target modality and controls the output of each target modality by implicitly taking its importance into account.

The concatenation of T_L, T_V, T_A is firstly filtered by three gates constructed with three feedforward neural networks, respectively, in order to obtain the weight-like vector for each modality. The output of the Multimodal Gating Block is consequently obtained by considering the distinction of different modalities. The process is implemented using Eqs. (16)-(20).

$$T^{[L,V,A]} = Concat(T_L, T_V, T_A) \tag{16}$$

$$\gamma_L = D_{\gamma_L}\left(T^{[L,V,A]}\right) \tag{17}$$

$$\gamma_V = D_{\gamma_V}\left(T^{[L,V,A]}\right) \tag{18}$$

$$\gamma_A = D_{\gamma_A}\left(T^{[L,V,A]}\right) \tag{19}$$

$$Z = Concat(\gamma_L \odot T_L, \gamma_V \odot T_V, \gamma_A \odot T_A) \tag{20}$$

where $\gamma_L \in *R*^{T\times d_L}$, $\gamma_V \in *R*^{T\times d_V}$, $\gamma_A \in *R*^{T\times d_A}$ represent the output of each target modality based on the gating mechanism, respectively.

The information flow of DEAN is shown in the following Algorithm.

Algorithm: Deep Emotional Arousal Network (DEAN), Cross-modal Transformer (CT), Multimodal BiLSTM System (MLS) and Multimodal Gating Block (MGB), where $m \in M = \{L, V, A\}$.

DEAN (X^m)
  $Y^m \leftarrow$ CT (X^m)
  $T^m \leftarrow$ MLS (Y^m)
  Z $\leftarrow$ MGB ($T^m, \cup_{m\in M}\{T^m\}$)
  return Z
  CT (X^m)
  for $m_\alpha \in M$ **do**: ◁ for all the M modalities
  for $m_\beta \in M - m_\alpha$ **do**:
    $Q_{m_{\alpha i}} \leftarrow W_{Q_i} X_{m_{\alpha i}}$, $K_{m_{\beta i}} = W_{K_i} X_{m_{\beta i}}$, $V_{m_{\beta i}} = W_{V_i} X_{m_{\beta i}}$
    $t\_head_i = softmax\left(\frac{Q_{m_{\alpha i}} K^T_{m_{\beta i}}}{\sqrt{d_k}} W_{th_i}\right) V_{m_{\beta i}}$
  $CroTA = concat(t\_head_1, \ldots, t\_head_h)\, W_o$
  $\widetilde{X}_{m_{\beta\to m_\alpha}} \leftarrow$ LaNorm$(Q_{m_\alpha} + CroTA)$
  $FFN \leftarrow W_{k_2}(ReLU(W_{k_1}\widetilde{X} + b_{k_1})) + b_{k_2}$
  $X_{m_{\beta\to m_\alpha}} \leftarrow$LaNorm$(\widetilde{X}_{m_{\beta\to m_\alpha}} + FFN)$
  return $X_{m_{\beta\to m_\alpha}}$
  $Y^m \leftarrow Concat_{m_\beta \in M - m_\alpha} X_{m_{\beta\to m_\alpha}}$
  return Y^m
BiLSTM (Y^m)
  for $t = 1, \ldots, T$ **do**:
  LSTM_STEP (Y^m_t, h^m_{t-1})
  for $m \in M$ **do**: ◁ for M modalities
  update each LSTM module by using
  Eqs. (9)-(14)
  return h^m_t
  $T^m \leftarrow \cup_{t\in T}\{h^m_t\}$
  return T^m
  MGB ($T^m, \cup_{m\in M}\{T^m\}$)
  $\gamma_m \leftarrow D(\cup_{m\in M}\{T^m\}; \theta_{\gamma_m})$
  Z $\leftarrow Concat_{m\in M}\, \gamma_m \odot T^m$
  return Z

4. Experiments

4.1. Datasets

To evaluate the efficiency of the proposed DEAN model, extensive experiments are performed using three benchmark problems pertaining to multimodal sentiment analysis and multimodal emotion recognition. For the sentiment analysis task, CMU-MOSI and CMU-MOSEI are selected as the benchmark datasets, while for the emotion recognition task, IEMOCAP dataset is used for performance evaluation and comparison.

CMU-MOSI consists of 2,199 short monologue video clips of 89 speakers from different national backgrounds on YouTube. Acoustic and visual features of CMU-MOSI are extracted at a sampling rate of 15 and 12.5 Hz respectively, and textual data are segmented per word. CMU-MOSEI has 3,228 monologue video clips containing a small number of characters for a total of 65 hours. Both datasets (MOSI and MOSEI) have been labeled on a continuous scale of [-3,3]. The IEMOCAP dataset covers 302 video sessions of 10 actors lasting up to 11 hours. Each segment has a corresponding emotional label, i.e., anger, excitement, fear, sadness, surprise, frustration, happiness, disappointment and neutrality.

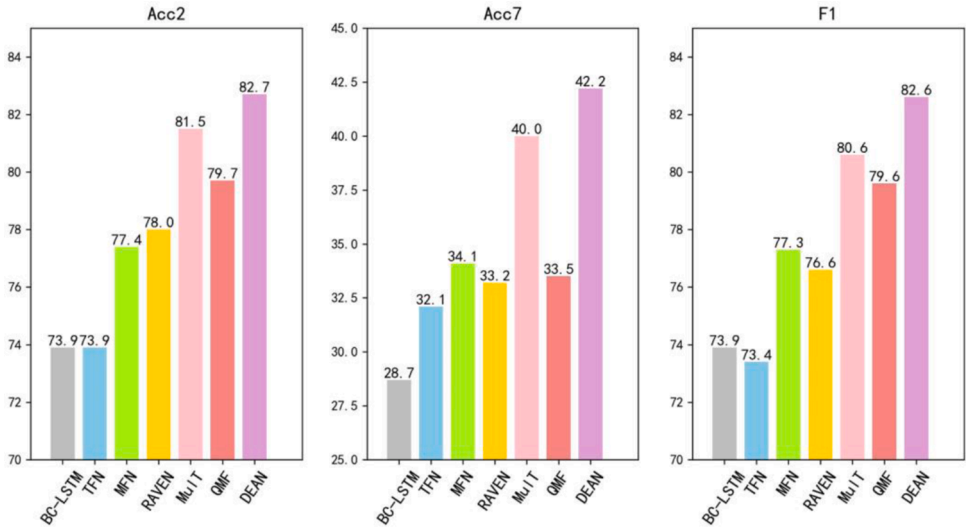

Fig. 6. Comparison results in term of three main metrics on CMU-MOSI

4.2. Baseline Models

The performance of DEAN is compared with those from seven state-of-the-art models in multimodal sentiment analysis and emotion recognition. These baseline models cover the main categories of models for multimodal sentiment analysis and emotion recognition proposed in recent years, which include the LSTM-based fusion model, Tensor-based model, Memory-based model, Attention-based model and other novel models.

BC-LSTM (Bidirectional Contextual LSTM) [2] is a multimodal sentiment analysis model that captures context information in a video, where the regular LSTM is replaced with a Bi-directional LSTM.

TFN (Tensor Fusion Network) [21] is a tensor-based fusion model, which explicitly aggregates unimodal, bimodal and trimodal interactions and captures view-specific and cross-view dynamics by creating a multidimensional tensor.

MFN (Memory Fusion Network) [29] is a memory-based fusion network that constructs a multimodal gated memory. The network is composed of a Delta-memory Attention Network, a Multi-view Gated Memory and the System of LSTMs, where the memory cell is updated together with the evolution of hidden states in three unimodal LSTM modules.

Graph-MFN (Graph Memory Fusion Network) [30] uses a dynamic fusion graph to model cross-modal interactions on the basis of the cyclic architecture of MFN. Graph-MFN replaces the Delta memory Attention network in MFN with a Dynamic Fusion Graph to make the network more interpretable.

RAVEN (Recurrent Attended Variation Embedding Network) [25] uses multimodal shifted word representations based on the visual and acoustic modalities. It effectively models the dynamic change of word representation space in a non-linguistic context.

MulT (Multimodal Transformer) [5] uses the Transformer structure to model unaligned multimodal sequence interactions. It achieves better performance on CMU-MOSI, CMU-MOSEI and IEMOCAP datasets.

QMN (Quantum-like Multimodal Network) [31] utilizes the mathematical formalism of quantum theory (QT) and a LSTM network to

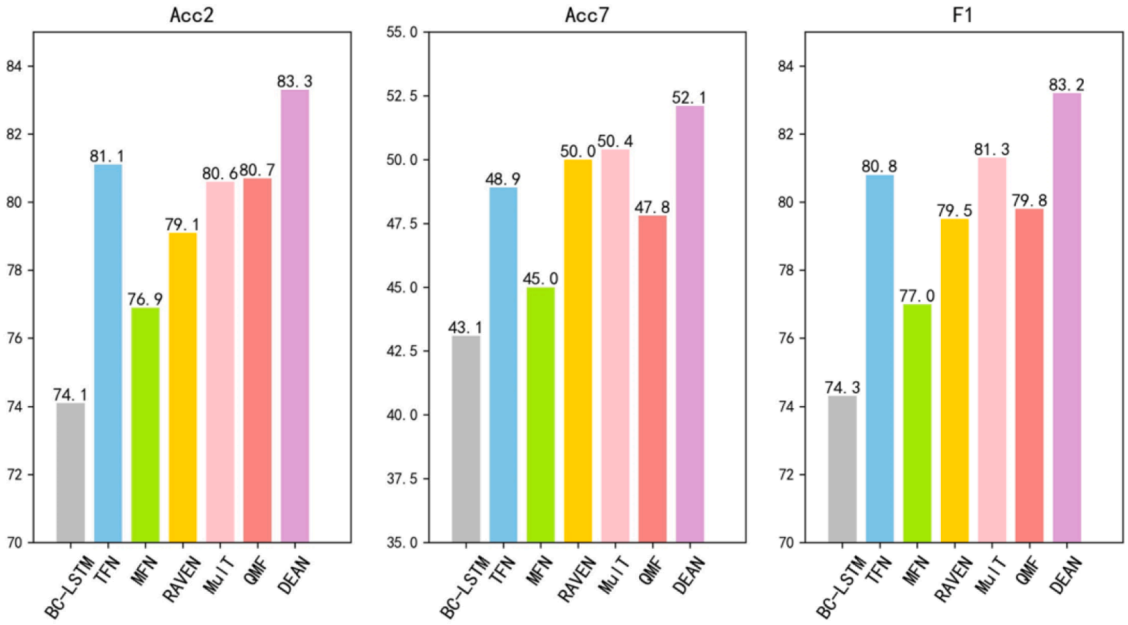

Fig. 7. Comparison results in term of three main metrics on CMU-MOSEI

Table 1
Results for emotion recognition on IEMOCAP

Metric	Happy		Sad		Angry		Neutral	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BC-LSTM	83.1	81.7	82.1	81.7	85.0	84.2	66.1	64.1
MFN	90.2	85.8	**88.4**	86.1	87.5	86.7	72.1	68.1
RAVEN	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
MulT	**90.7**	88.6	86.7	86.0	87.4	87.0	72.4	70.7
DEAN (ours)	90.6	**89.6**	86.9	**86.3**	**88.7**	**88.6**	**72.4**	**71.7**

capture the interactions between different modalities from different speakers. QMN consists of a multi-modal decision fusion method and a strong/weak influence model to represent the interactions within/between utterances.

5. Results and Discussion

Various experiments are conducted with CMU-MOSI and CMU-MOSEI for multimodal sentiment analysis and IEMOCAP for emotion recognition. Each dataset is divided into 70% for training, 10% for validation, and 20% for test, respectively.

The comparison results of multimodal sentiment analysis are shown in Figs. 6 and 7, respectively, with the best scores highlighted in bold. The results of DEAN are calculated by averaging 40 runs. The best hyperparameters of DEAN are $l_{dim} = v_{dim} = a_{dim} = 30$, batch size $bs = 24$, number of heads $h = 5$ and learning rate $lr = 0.001$.

Fig. 6 shows that DEAN outperforms other methods in discriminating the finer-grained human emotions. This is due to the capability of DEAN in capturing the characteristics of the human emotion arousal model. This capability is in line with the statement published in an article in Science, i.e., "it is currently unclear which aspects of the biological circuitry are computationally essential and could be useful for network-based Artificial Intelligence system, but the differences in structure are prominent." (Ullman et al., [32]). Combining deep learning with brain-like innate structures endows DEAN with the power of handling complicated interactions among multimodality inputs. From Fig. 6 we notice that DEAN increases ACC7 by 2.2% and decreases MAE by 1.8, as compared with those from MulT on the CMU-MOSI dataset. Fig. 7 shows that DEAN achieves great performance on the CMU-MOSEI dataset.

In order to obtain fine-grained emotion understanding, we conduct extra experiments for emotion recognition with the IEMOCAP dataset. Table 1 shows the experimental results. DEAN achieves better scores in the categories of happy, angry and neutral emotions as compared with those from state-of-the-art methods, except for the category of sad. For this sad emotion, memory-based models produce the highest score among all compared models. Recognizing the neural emotion is the most challenging task for all models, as listed in Table 1.

5.1. Ablation Studies

In this section, various ablation studies have been conducted on both CMU-MOSEI and CMU-MOSI datasets. The aims are: (1) to disclose the influence of each individual module on the proposed model, (2) to investigate the importance of each modality, (3) to explore the interactions between modalities.

5.1.1. The influence of each individual module on the proposed model

To investigate the effect of each individual module of DEAN, we gradually remove each component from DEAN, as follows.

DEAN: The original proposed model, which consists of three modules, i.e., (1) Cross-modal Transformer, (2) Multimodal BiLSTM System and (3) Multimodal Gating Block, is used as the baseline for comparison.

DEAN w/o GATE: The Multimodal Gating Block is removed from DEAN. In this case, the model is similar to an attention-based LSTM model [31], which lacks the capability of controlling the output of target modality. From the standpoint of psychology, DEAN w/o GATE is analogous to the perceptual reorganization of patients after synesthesia imbalance.

DEAN w/o BiLSTM: The multimodal BiLSTM system is removed from DEAN. The output only implements the spatial fusion of multimodal information while ignoring the temporal fusion information. It is designed to evaluate the importance of spatio-temporal fusion between modalities.

DEAN w/o BiLSTM & GATE: Removing both Multimodal Gating Block and Multimodal BiLSTM system from DEAN results in an attention-based fusion model. In this case, we can explore that whether a Transformer (i.e., the Cross-modal Transformer module) can fully replace RNN models in sequential modeling.

Table 2
Experimental results of module ablation on CMU-MOSEI

With cross-talking attention	Acc7	Acc2	F1	MAE
DEAN	**52.3**	**83.3**	**83.2**	**0.571**
DEAN w/o GATE	52.0	83.0	82.8	0.573
DEAN w/o BiLSTMs	51.6	82.5	82.3	0.579
DEAN w/o BiLSTMs & GATE	51.2	81.9	81.8	0.583
Without cross-talking attention	Acc7	Acc2	F1	MAE
DEAN	52.1	82.8	82.9	0.573
DEAN w/o GATE	51.4	82.0	82.3	0.591
DEAN w/o BiLSTMs	50.8	82.2	81.6	0.596
DEAN w/o BiLSTMs & GATE	50.3	80.9	80.9	0.599

Table 3
Experimental results of module ablation on CMU-MOSI

With cross-talking attention	Acc7	Acc2	F1	MAE
DEAN	**42.2**	**82.7**	**82.6**	**0.843**
DEAN w/o GATE	41.9	82.4	82.3	0.849
DEAN w/o BiLSTMs	41.5	82.0	82.1	0.852
DEAN w/o BiLSTMs & GATE	40.0	81.2	81.2	0.854
Without cross-talking attention	Acc7	Acc2	F1	MAE
DEAN	41.3	82.1	82.0	0.857
DEAN w/o GATE	40.9	81.6	81.5	0.867
DEAN w/o BiLSTMs	40.5	81.4	81.2	0.862
DEAN w/o BiLSTMs & GATE	40.1	80.6	80.4	0.869

Table 4
Ablation study results on importance of individual modality

Unimodality	Acc7	Acc2	F1	MAE
Language	47.5	78.1	78.7	0.647
Audio	45.1	66.2	70.3	0.741
Visual	43.4	65.6	69.9	0.762
Multi-modality with DEAN	**52.1**	**83.3**	**83.2**	**0.571**

Table 2 and Table 3 show the experimental results of ablation studies for sentiment analysis on CMU-MOSI. The comparison results indicate that all the DEAN models with cross-talking attention on the Cross-modal transformer outperforms the counterparts without the cross-talking attention. The DEAN with all modules achieves the best result on all metrics. Taking Acc7 with cross-talking attention as an example, it is found from Table 2 that DEAN w/o BiLSTM and DEAN w/o GATE produce decreased accuracy rates of 51.6% and 52.0%, as compared with 52.3% of DEAN, respectively. Similar results are shown in Table 3. This empirical outcome indicates that the Multimodal BiLSTM system is important, which supports our claim that the coherence of human emotion is critical for sentiment analysis. The experimental results also bring an insight that RNN-like structure is still helpful to improve the temporal aspects of transformer-based models.

5.1.2. The importance of individual modality

To investigate the importance of each modality, we conduct several experiments with and without language, audio and visual modalities for

Table 5
The experimental results of interaction for bi-modality (the modality that arrow points to is the target one)

Modality	Acc7	Acc2	F1	MAE
V→L	50.7	78.9	79.7	0.590
A→L	50.3	78.4	79.6	0.595
L→V	49.5	77.8	79.4	0.626
A→V	44.7	69.7	71.7	0.755
L→A	48.6	77.6	79.4	0.628
V→A	44.9	69.5	72.4	0.751

Table 6
The experimental results of interaction among tri-modality (the modality that arrow points to is the target one)

	Acc7	Acc2	F1	MAE
V, A→L	51.7	82.1	82.3	0.577
L, A→V	51.1	80.8	80.7	0.596
L, V→A	50.9	79.6	80.4	0.583
DEAN	**52.1**	**83.3**	**83.2**	**0.571**

sentiment analysis on CMU-MOSEI. The embedding vector of each individual modality is used as the input for its corresponding Transformer separately. The ablation study results are illustrated in Table 4.

We can draw a conclusion from Table 4 that the modality of Language plays an important role in sentiment analysis, as compared with Audio and Visual modalities. This is because Language is considered as a pivot modality for sentiment analysis when Transformer-based methods are used. The experimental results also indicate that the Multimodal Gating Block is indispensable for distinguishing the contribution of each modality, since the integrated model with all three modules achieves the best performance, as compared with those from individual modality input.

5.1.3. The interactions between modalities

To study the interactions between modalities, we conduct bi-modality and tri-modality experiments. The aim is to observe the interaction between auxiliary modality and target modality, where language, visual and audio are set as the target modality and auxiliary modality respectively. The experimental results are shown in Table 5 and Table 6.

Based on Table 5 and Table 6, the performance of the target modal can be enhanced by other auxiliary modalities, in the way of bi-modality or tri-modality. Specifically, the tri-modality combination achieves the highest performance, as compared with unimodal and bimodal sentiment analysis. The performance of taking language as the target modality is better than those of visual or audio as the target modality, regardless of whether one or two auxiliary modalities are used. Table 5 shows that taking language as the target while visual and audio as auxiliary modalities yields better results, i.e., 50.7% and 50.3% for language, 49.5% and 44.7% for visual, and 48.6% and 44.9% for audio. Table 6 shows similar observations.

In addition, language plays a crucial role as an auxiliary modality. Compared with L→V, Acc7 of A→V decreases by 4.8% in the absence of feature inputs from the language modality. This observation further indicates the importance of language in multimodal sentiment analysis.

6. Conclusions

Inspired by the human emotional arousal model in psychology, we have proposed a Deep Emotional Arousal Network (DEAN) for multimodal sentiment analysis and emotion recognition tasks in this paper. DEAN provides an integrated framework for modelling human communication with multimodality information. It is capable of representing the emotional coherence by incorporating time-dependent interactions into the parallel structure of a Transformer model and identifying the distinction of different modalities by embedding a Multimodal Gating Block. A series of comprehensive evaluation and analysis studies on three benchmark datasets have been conducted. The empirical results indicate the effectiveness of DEAN in multimodal sentiment analysis and emotion recognition tasks, outperforming several state-of-the-art models on all the three benchmark problems.

The research of multimodal sentiment analysis and emotion recognition is highly correlated to the studies on multisensory integration in brain neuroscience [33]. We aim to integrate the recent findings from the multisensory literature into DEAN for multimodal sentiment analysis and emotion recognition in future.

CRediT authorship contribution statement

Feng Zhang: Conceptualization, Methodology, Writing - Original Draft. **Xi-Cheng Li:** Experiments Validation, Writing Assistant - Original Draft;

Chee-Peng Lim: Review & Editing, Investigation, Editing. **Qiang Hua:** Resources, Supervision; Funding Acquisition, Editing. **Chun-Ru Dong:** Conceptualization, Editing. **Jun-Hai Zhai:** Experiments Validation, Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the editors and anonymous reviewers for their valuable time and insightful comments. This research is partly supported by the National Nature Science Fund Project NSFCs (No. 61773150), the Natural Science Foundation of Hebei Province (No. F2018201115), the Key Scientific Research Foundation of Education Department of Hebei Province (ZD2019021) and the key R&D program of science and technology foundation of Hebei Province (19210310D).

References

[1] K.S. LaBar, R. Cabeza, Cognitive neuroscience of emotional memory, Nature Reviews Neuroscience 7 (1) (2006) 54–64.
[2] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.P. Morency, Context-dependent sentiment analysis in user-generated videos, *Proceedings of the 55th annual meeting of the association for computational linguistics* volume 1 (2017) 873–883.
[3] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2247–2256. ACL 2018.
[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q.V Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: *Advances in Neural Information Processing Systems 2019*, NIPS 2019, 2019, pp. 5753–5763.
[5] Y.H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the conference. Association for Computational Linguistics, 2019, July, p. 6558. Meeting, Vol. 2019.
[6] J.B. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis, Second Grand-Challenge and Workshop on Multimodal Language (2020) (Challenge-HML), 2020.
[7] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, D Song, What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis, Information Fusion 66 (2021) 184–197.
[8] P.H. Lindsay, D.A. Norman, Human information processing: An introduction to psychology, Academic press, 2013.
[9] E. Acar, F. Hopfgartner, S Albayrak, A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material, Multimedia Tools and Applications 76 (9) (2017) 11809–11837.
[10] Sheng-hua Zhong, Jiaxin Wu, Jianmin Jiang, Video summarization via spatio-temporal deep architecture, Neuro-computing 332 (2019) 224–235.
[11] Y. Zhu, M. Tong, Z. Jiang, S. Zhong, Q Tian, Hybrid feature-based analysis of video's affective content using protagonist detection, Expert Systems with Applications 128 (2019) 316–326.

[12] S. Sivaprasad, T. Joshi, R. Agrawal, N. Pedanekar, Multimodal continuous prediction of emotions in movies using long short-term memory networks, in: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 413–419.
[13] D. Gui, S.H. Zhong, Z Ming, Implicit affective video tagging using pupillary response, in: International Conference on Multimedia Modeling, Springer, Cham, 2018, pp. 165–176.
[14] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing 45 (11) (1997) 2673–2681.
[15] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, D. Xu, Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition, in: proceedings of the 26th international joint conference on artificial intelligence, 2017, pp. 3595–3601.
[16] E. Morvant, A. Habrard, S Ayache, Majority vote of diverse classifiers for late fusion, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2014. Joensuu, Finland, August 20-22.
[17] P.P. Liang, Z. Liu, A. Zadeh, L. Morency, Multimodal language analysis with recurrent multistage fusion, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 150–161. ACM 2018.
[18] V. Vielzeuf, S. Pateux, F Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 569–576.
[19] T.Y. Lin, A. Roy Chowdhury, S Maji, Bilinear CNN models for fine-grained visual recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449–1457.
[20] T. Baltrušaitis, N. Banda, P Robinson, Dimensional affect recognition using continuous conditional random fields, in: *2013 the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (FG), 2013, pp. 1–8.
[21] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114.
[22] S. Mai, S. Xing, H. Hu, Locally confined modality fusion network with a global perspective for multimodal human affective computing, IEEE Transactions on Multimedia 22 (1) (2019) 122–137.
[23] H. Pham, P.P. Liang, T. Manzini, L.P. Morency, B Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, Proceedings of the AAAI Conference on Artificial Intelligence 33 (1) (2019) 6892–6899.
[24] Pham H., Manzini T., Liang P. P., and Poczos B. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. *arXiv preprint arXiv*:1807.03915.
[25] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.P. Morency, Words can shift: Dynamically adjusting word representations using nonverbal behaviors, Proceedings of the AAAI Conference on Artificial Intelligence 33 (1) (2019) 7216–7223.
[26] Shazeer N., Lan Z., Cheng Y., Ding N., & Hou, L. 2020. Talking-heads attention. arXiv preprint arXiv:2003.02436.
[27] Chen M. X., Firat O., Bapna A., Johnson M., Macherey W., Foster G., and Hughes M. 2018. The best of both worlds: Combining recent advances in neural machine translation. *arXiv preprint arXiv*:1804.09849.
[28] Ba J. L., Kiros J. R., & Hinton G. E. 2016. Layer normalization. *arXiv preprint arXiv: 1607.06450*.
[29] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence 32, 2018.
[30] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 1 (2018) 2236–2246.
[31] C. Li, Z. Bao, L. Li, Z. Zhao, Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition, Information Processing & Management 57 (3) (2020), 102185.
[32] S. Ullman, Using neuroscience to develop artificial intelligence, Science 363 (6428) (2019) 692–693.
[33] B.E. Stein, T.R. Stanford, Multisensory integration: current issues from the perspective of the single neuron, Nature reviews neuroscience 9 (4) (2008) 255–266.