INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Conversational Emotion Analysis via Attention Mechanisms

Zheng Lian[1,3], *Jianhua Tao*[1,2,3], *Bin Liu*[1], *Jian Huang*[1,3]

[1]National Laboratory of Pattern Recognition, CASIA, Beijing, China
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{zheng.lian, jhtao, liubin, jian.huang}@nlpr.ia.ac.cn

Abstract

Different from the emotion recognition in individual utterances, we propose a multimodal learning framework using relation and dependencies among the utterances for conversational emotion analysis. The attention mechanism is applied to the fusion of the acoustic and lexical features. Then these fusion representations are fed into the self-attention based bi-directional gated recurrent unit (GRU) layer to capture long-term contextual information. To imitate real interaction patterns of different speakers, speaker embeddings are also utilized as additional inputs to distinguish the speaker identities during conversational dialogs. To verify the effectiveness of the proposed method, we conduct experiments on the IEMOCAP database. Experimental results demonstrate that our method shows absolute 2.42% performance improvement over the state-of-the-art strategies.
Index Terms: conversational emotion analysis, multimodal fusion, contextual features, interaction strategy

1. Introduction

Recently, conversational emotion analysis has attracted attention due to its wide applications in human-computer interaction. Different from emotion recognition in individual utterances, conversational emotion analysis utilizes the relation among utterances to track the user's emotion states during conversations.

To improve the performance of conversational emotion analysis, prior works have been performed mainly in three directions: (1) After extraction of multimodal features, there is a question of how to fuse these features effectively. (2) After multi-modalities fusion, there is another question of how to utilize contextual features to help predict emotion states of the current utterance. (3) The last question is how to model the interaction of different speakers in conversational dialogs.

After extraction of multimodal features, there are mainly three strategies for multi-modalities fusion, namely feature-level fusion, decision-level fusion and model-level fusion [1]. In feature-level fusion, multimodal features are concatenated into a joint feature vector for emotion recognition. Although this method can significantly improve the performance [2, 3], it suffers from the curse of dimensionality. Decision-level fusion can eliminate the disadvantage of feature-level fusion. In decision-level fusion, multimodal features are modeled by corresponding classifiers first, and then recognition results from each classifier are fused by weighted sum or additional classifiers [4, 5]. However, it ignores interactions and correlations between different modalities. To deal with this problem, a vast majority of works are explored toward model-level fusion. It is a compromise for feature-level fusion and decision-level fusion methods, which is proposed to fuse intermediate representations of different features [6, 7, 8]. Recently, attention-based model-level fusion strategies have gained promising results. Chen et al. [9] proposed a multimodal fusion strategy, which can dynamically pay attention to different modalities at each timestep. Poria et al. [10] proposed an attention-based network for multimodal fusion, which fused multimodal features via the attention score for each modality. Inspired by the power of attention mechanisms for emotion recognition [9, 10], we also utilize the attention-base model-level fusion strategy in this paper.

According to the emotion generation theory [11], human perceive emotions not only through individual utterances but also by surroundings. To take into account the contextual effect in the dialogs, Vanzo et al. [12] used wider contexts (such as topics and hashtags) to help identify emotion states. Poria et al. [13] proposed a LSTM-based network to capture contextual information from their surroundings. Recently, the self-attention mechanism [14] has been verified to attend to longer sequences than many typical RNN-based models [14, 15]. This mechanism can provide an opportunity for injecting the global context of the whole sequence into each input utterance. However, this mechanism ignores information about the order of the sequential utterances in conversational dialogs [14], which is important for conversational emotion analysis. To deal with this problem, we use the bi-directional gated recurrent unit (GRU) layer [16], in combination with the self-attention mechanism for emotion analysis. With the help of the bi-directional GRU layer, the order of sequential utterances is injected into hidden vectors. Then the self-attention mechanism is utilized to capture the long-term contextual information.

As a person's emotion state can be influenced by the interlocutor's behaviors in conversational dialogs [17], speaker information can be utilized to improve the performance of emotion recognition. A vast majority of works have been explored to model the interaction of different speakers in recent years [17, 18, 19, 20]. Lee et al. [17] proposed a dynamic bayesian network to model the conditional dependency between two interacting partners' emotion states in a dialog. Zhang et al. [18] proposed a multi-speaker emotion recognition model. The emotion probabilities of previous utterances of each speaker were utilized to estimate emotion of the current utterance. Chen et al. [19, 20] combined feature sequences of different speakers to imitate the interaction patterns. These methods [17, 18, 19, 20] need explicit speaker identity of each utterance. But speaker identities are not always available in sensitive or anonymous conversations. To handle this situation, we propose a new interactive strategy in this paper. We first extract speaker embeddings for each utterance using the pre-trained speaker verification system [21]. Then these embeddings are utilized as additional inputs to distinguish speaker identities in conversations.

This paper proposes a multimodal learning framework for conversational emotion analysis. The main contributions of this paper lie in four aspects: 1) to prioritize important modalities, we use the attention mechanism for multi-modalities fu-

http://dx.doi.org/10.21437/Interspeech.2019-1577

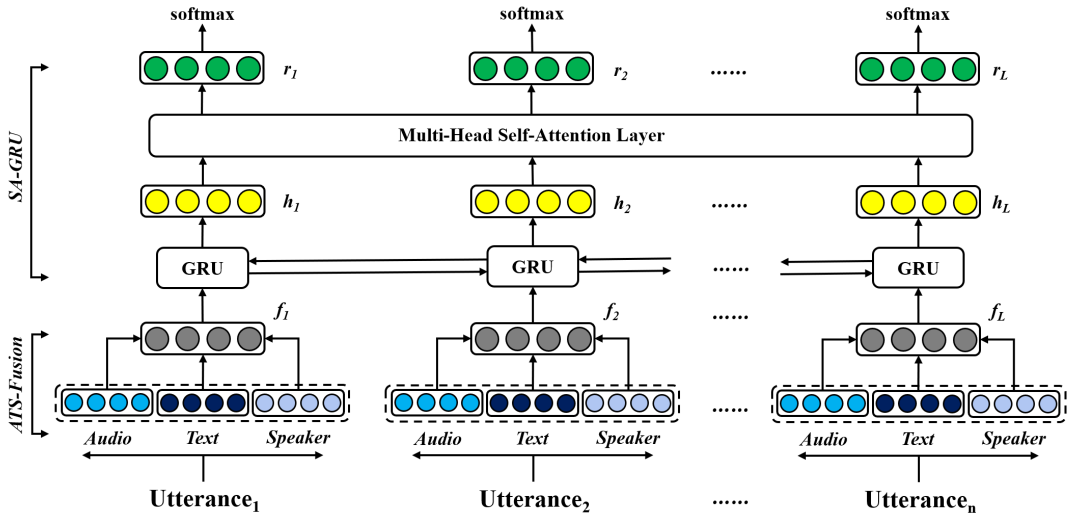

Figure 1: *Overall structure of the proposed framework: multimodal input features are fused by ATS-Fusion, followed by SA-GRU for emotion classification.*

sion; 2) to capture long-term contextual information, we use the bi-directional GRU layer, in combination with the self-attention mechanism; 3) to model the interaction of different speakers when explicit speaker identities are unavailable, we utilize speaker embeddings as additional inputs; 4) our proposed method is superior to state-of-the-art approaches for conversational emotion analysis. To the best of our knowledge, it is the first time that the self-attention mechanism is utilized for conversational emotion analysis. In the meantime, we provide an approach to modeling the interaction when explicit speaker identities are unavailable.

2. Proposed Method

To infer the emotion state in realistic conversational dialogs, we propose a multimodal learning framework. As illustrated in Fig. 1, the proposed framework contains two components: the Audio-Text-Speaker Fusion component (ATS-Fusion) for multi-modalities fusion and the Self-Attention based GRU component (SA-GRU) for emotion classification. ATS-Fusion takes the audio, text and speaker information as the inputs and outputs robust representations using the attention mechanism. SA-GRU uses the bi-directional GRU layer, in combination with the multi-head self-attention layer to amplify the important contextual evidents for emotion classification. Meanwhile, to model the interaction of different speakers when explicit speaker identities are unavailable, we utilize speaker embeddings as additional inputs.

2.1. Speaker encoder

Speaker embeddings should capture the speaker characteristics. We extract speaker embeddings from each utterance using the pre-trained speaker verification system in [22]. This system maps variable-length utterances to fixed-dimensional embeddings, known as x-vector [22]. The architecture is based on the end-to-end system described in [21]. It consists of five layers that operate on speech frames, a statistics pooling layer that aggregates over the frame-level representations, two additional layers that operate at the segment-level, and finally a softmax output layer. In the training process, this system also utilizes data augmentation to multiply the amount of training samples and improve robustness.

2.2. Attention-based multi-modalities fusion: ATS-Fusion

Not all modalities are equally relevant in emotion classification. To prioritize important modalities, we utilize the attention mechanism [10, 23, 24] in ATS-Fusion. This method takes audio, text and speaker information as inputs and outputs an attention score for each modality.

Let us assume a video to be considered as $V = [u_1, u_2, ..., u_i, ..., u_L]$, where u_i is the i^{th} utterance in the video and L is the number of utterances in the video. We first extract acoustic features a_i, lexical features t_i and speaker features s_i from the utterance u_i. Then we equalize the feature dimensions of all three inputs to size d and process utterance-level concatenation:

$$u_i^{cat} = Concat(W_a a_i, W_t t_i, W_s s_i) \tag{1}$$

where W_a, W_t and W_s are feature embedding matrices for acoustic features, lexical features and speaker features, respectively. Here, $W_a a_i \in \mathbb{R}^{d\times 1}$, $W_t t_i \in \mathbb{R}^{d\times 1}$, $W_s s_i \in \mathbb{R}^{d\times 1}$ and $u_i^{cat} \in \mathbb{R}^{d\times 3}$.

The attention weight vector α_{fuse} and the fusion representation f_i of the input utterance u_i are computed as follows:

$$P_F = tanh(W_F u_i^{cat}) \tag{2}$$

$$\alpha_{fuse} = softmax(w_F^T P_F) \tag{3}$$

$$f_i = u_i^{cat} \alpha_{fuse}^T \tag{4}$$

where $W_F \in \mathbb{R}^{d\times d}$ and $w_F \in \mathbb{R}^{d\times 1}$ are trainable parameters. Here, $P_F \in \mathbb{R}^{d\times 3}$, $\alpha_{fuse} \in \mathbb{R}^{1\times 3}$ and $f_i \in \mathbb{R}^{d\times 1}$.

2.3. Self-attention based emotion classifier: SA-GRU

In the conversational emotion analysis, the emotion state of the target utterance is temporally and contextually dependent to surroundings. To take into account such dependencies, we use the bi-directional GRU layer, in combination with the multi-head self-attention layer for emotion classification.

2.3.1. Gated recurrent unit

GRU is usually adopted as the basic unit in RNN as it is able to solve the vanishing gradient problem in the conventional RNN training. Since each GRU cell consists of an update gate and a reset gate to control the flow of information, it is capable of modeling long-term dynamic dependencies.

Let $F = [f_1, f_2, ..., f_t, ..., f_L]$ be the input to the GRU network, where f_t is the fusion representation of the utterance u_t (in Sec 2.2) and L is the number of utterances in the video. To obtain contextually dependent utterance representations $H = [h_1, h_2, ..., h_t, ..., h_L]$, the GRU network computes the hidden vector sequence H from F with the following equations:

$$z_t = \sigma_g(W_z f_t + U_z h_{t-1} + b_z) \quad (5)$$

$$r_t = \sigma_g(W_r f_t + U_r h_{t-1} + b_r) \quad (6)$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \sigma_h(W_h f_t + U_h(r_t \circ h_{t-1}) + b_h) \quad (7)$$

where σ_g is the sigmoid activation function, σ_h is the hyperbolic tangent and $\circ$ is element-wise multiplication. z and r are the update gate vectors and reset gate vectors, respectively. W, U and b are weight matrices and bias vectors for each gate.

2.3.2. Multi-head self-attention layer

To focus on only relevant utterances in emotion classification of the target utterance, a multi-head self-attention layer is used.

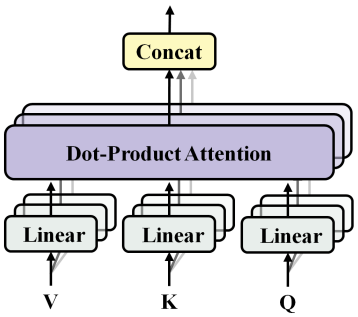

Figure 2: *Multi-head self-attention layer: it consists of several dot-product attention layers running in parallel.*

The outputs of the GRU network $H = [h_1, h_2, ..., h_L]$ are passed into the multi-head self-attention layer (in Fig. 2). To extract queries Q, keys K and values V, we linearly project the output H for h times with different linear projections, which are computed as follows:

$$V = Concat(HW_1^V, ..., HW_h^V) \quad (8)$$

$$K = Concat(HW_1^K, ..., HW_h^K) \quad (9)$$

$$Q = Concat(HW_1^Q, ..., HW_h^Q) \quad (10)$$

where $H \in \mathbb{R}^{L \times d}$, $W_i^Q \in \mathbb{R}^{d \times (d/h)}$, $W_i^K \in \mathbb{R}^{d \times (d/h)}$, $W_i^V \in \mathbb{R}^{d \times (d/h)}$ and h is the number of heads.

On each of these projected versions of queries HW_i^Q, keys HW_i^K and values HW_i^V, we perform the dot-product attention with following equations:

$$head_i = softmax((HW_i^Q)(HW_i^K)^T)((HW_i^V) \quad (11)$$

Then outputs of attention functions $head_i \in \mathbb{R}^{L \times (d/h)}, i \in [1, h]$ are concatenated together as final values R. As $R \in \mathbb{R}^{L \times d}$, where L is the number of utterances in a dialog. The matrix R can be represented as $R = [r_1, r_2, ..., r_t, ..., r_L]$, where $r_t \in \mathbb{R}^d$ for $t = 1$ to L. Finally, R is fed into a linear projection layer and a softmax layer for emotion recognition.

3. Experiments and Discussion

3.1. Corpus description

The IEMOCAP [25] database contains about 12.46 hours of audio-visual conversations in English. There are five sessions with two actors each (one female and one male) and each session has different actors. Each session has been segmented into utterances, which are labeled into ten discrete labels (e.g., happy, sad, angry). To compared with state-of-the-art approaches [13, 26], we form a four-class emotion classification dataset containing $happy(1636), angry(1103), sad(1084), neutral(1708)$, where happy and excited categories are merged into a single happy category. Thus 5531 utterances are involved. To ensure that models are trained and tested on speaker independent sets, utterances from the first 8 speakers are used as the training set and utterances from other speakers are used as the testing set.

3.2. Experimental setup

Features: Acoustic features are extracted from waveforms using the openSMILE [27] toolkit. Specifically, we use the IS13-ComParE configuration file in openSMILE. Totally, 6373-dimensional utterance-level acoustic features are extracted, containing Mel-frequency cepstral coefficients (MFCCs), voice intensity, pitch, and their statistics (e.g., mean, root quadratic mean); Lexical features are extracted from transcriptions through two steps. In particular, we first get 1024-dimensional vector representations of words using the deep contextualized word representation — ELMo [28]. These word vectors are learned from the deep bidirectional language model trained on the 1 Billion Word Benchmark [29]. To obtain utterance-level lexical features, we then calculate mean values of these word vectors; Speaker embeddings are extracted from raw input utterances using the Kaldi Speech Recognition Toolkit [30]. Specifically, we use the x-vector system [22] trained on the NIST SREs [31, 32]. Finally, 512-dimensional speaker embeddings are extracted from the x-vector system.

Settings: ATS-Fusion contains three fully-connected layers of size $d = 100$. These layers map acoustic, lexical and speaker features into fixed dimensionality, respectively. SA-GRU contains the bi-directional GRU layer with 100 units, followed with a self-attention layer (100 dimensional states and 4 attention heads). We use the Adam optimization scheme with a learning rate of 0.0001 and a batch size of 20. Cross-entropy loss is used as the loss function of the emotion recognition task. To prevent over-fitting, we use dropout [33] with $p = 0.2$ and $L2$ regularization. To alleviate the impact of the weight initialization, each configuration is tested 20 times. The unweighted accuracy (UA) is chosen as our evaluation criterion.

3.3. Contribution of individual components

In this section, we evaluate the contribution of each component in the framework. Four comparison systems with different combination of individual components are implemented to compare with the proposed framework.

(1) System 1 (S1): Ignoring speaker embeddings, we only use acoustic and lexical features for emotion recognition. The

attention mechanism in Sec 2.2 is used for multi-modalities fusion, marked as the Audio-Text Fusion component (AT-Fusion).

(2) System 2 (S2): Instead of using ATS-Fusion, we first equalize dimensions of acoustic, lexical and speaker features into $d = 100$ via feed forward layers. Then we simply add these multimodal representations together. This fusion approach is marked as ADD.

(3) System 3 (S3): Instead of feeding the outputs of ATS-Fusion to SA-GRU, we only use the self-attention layer.

(4) System 4 (S4): Instead of feeding the outputs of ATS-Fusion to SA-GRU, we only use the bi-directional GRU layer.

(5) System 5 (S5): It is our proposed framework for conversational emotion recognition. Acoustic, lexical and speaker features are fused by ATS-Fusion. Then the outputs of ATS-Fusion are fed into SA-GRU for emotion classification.

Table 1: *Experimental results with different combination of components.*

	Modalities	Fusion	Classifier	UA(%)
S1	A+T	AT-Fusion	SA-GRU	76.40
S2	A+T+S	ADD	SA-GRU	76.65
S3	A+T+S	ATS-Fusion	Attention	66.00
S4	A+T+S	ATS-Fusion	Bi-GRU	77.29
S5	A+T+S	ATS-Fusion	SA-GRU	78.02

To verify the importance of speaker embeddings, we compare the performance of S1 and S5. Experimental results in Table 1 demonstrate that S5 gains better performance, 78.02%, than S1, 76.40%. The speaker embeddings indicate the speaker's identity of each utterance. They reflect personal information and role switches in the conversation, which are important for conversational emotion recognition [18, 19]. Therefore, cooperating speaker embeddings with other multimodal features can improve the performance of emotion recognition.

To verify the effectiveness of ATS-Fusion, we compare the performance of S2 and S5. Experimental results show that S2 achieves a limited performance in emotion recognition. ATS-Fusion takes audio, text and speaker information as inputs and prioritizes important modalities via attention weights. This approach is similar to human perceptions since humans can dynamically focus on more trustful modalities to understand emotions [9, 10]. While ADD is a special case of ATS-Fusion. It treats each modality equally and cannot select relevant modalities for emotion recognition. Therefore, ATS-Fusion is more suitable for multimodal fusion than ADD.

To verify the effectiveness of SA-GRU, we compare the performance of S3, S4 and S5. We find S5 is superior to S3 and S4. The bi-directional GRU layer (in S4) models the temporal and contextual dependence in conversational dialogs. The self-attention layer (in S3) amplifies the important contextual evidences for emotion analysis of the target utterance. To employ the power of these components, our proposed SA-GRU (in S5) combines the bi-directional GRU layer with the self-attention layer for emotion recognition. Experimental results indicate the effectiveness of SA-GRU, which can gain better performance than the single bi-directional GRU layer (or the single self-attention layer).

Meanwhile, experimental results in Table 1 demonstrate that S3 gains much lower performance than S4. The self-attention layer contains no recurrence and no convolution networks. If we shuffle the order of utterances in the dialogs, we will get the same output. However, the order of sequential utterances is important for conversational emotion analysis [14]. With the help of the GRU layer, the order of sequential utterances is injected into hidden vectors, which can further improve the performance of emotion recognition.

3.4. Comparison to state-of-the-art approaches

To show the effectiveness of the proposed method, we compare our method with currently advanced approaches. Experimental results of different methods are shown in Table 2.

Table 2: *The performance of state-of-the-art approaches and the proposed approach on the IEMOCAP database.*

Approaches	UA (%)
Rozgić et al. (2012) [26]	67.40
Jin et al. (2015) [34]	69.20
Poria et al. (2017) [13]	75.60
Li et al. (2018) [35]	74.80
Proposed method	**78.02**

Compared with our proposed method, these approaches [13, 26, 34, 35] also utilized acoustic and lexical features for emotion recognition. Rozgić et al. [26] combined decision trees with support vector machines (SVM) for the sentence-level multimodal emotion recognition. Jin et al. [34] investigated different ways to combine acoustic and lexical features, including decision-level fusion and feature-level fusion. Poria et al. [13] proposed a LSTM based network to capture contextual information from their surroundings in the same video. Li et al. [35] proposed a multi-modal, multi-task deep learning framework to infer the user's emotive states.

Experimental results in Table 2 demonstrate the effectiveness of the proposed method. Our method shows absolute 2.42% performance improvement over state-of-the-art approaches. It proves that modeling long-term dynamic dependencies and considering the speaker information can improve the performance of emotion recognition.

4. Conclusions

This paper proposes a multimodal learning framework to infer the emotion state in realistic conversational dialogs. To evaluate the effectiveness of the proposed method, we conduct experiments on the IEMOCAP database. Experimental results reveal that ATS-Fusion, which is employed to fuse multimodal features, can provide robust representations of each utterance for following SA-GRU. SA-GRU, which amplifies the important contextual evidents for emotion classification, can gain better performance than the single bi-directional GRU layer (or the single self-attention layer). With the help of speaker embeddings, the proposed framework can utilize the interactive information in conversational dialogs, which can further improve the performance of emotion recognition. Experimental results on the IEMOCAP database demonstrate that the proposed framework is superior to state-of-the-art strategies.

5. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2018YFB1005003), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61771472), and the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100).

6. References

[1] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA transactions on signal and information processing*, vol. 3, 2014.

[2] F. Eyben, S. Petridis, B. Schuller, and M. Pantic, "Audiovisual vocal outburst classification in noisy acoustic conditions," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5097–5100.

[3] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[4] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 396–406.

[5] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2462–2465.

[6] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Two-level hierarchical alignment for semi-coupled hmm-based audiovisual emotion recognition with temporal course," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1880–1895, 2013.

[7] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic bayesian network models," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 609–618.

[8] M. Paleari, R. Benmokhtar, and B. Huet, "Evidence theory-based multimodal emotion recognition," in *International Conference on Multimedia Modeling*. Springer, 2009, pp. 435–446.

[9] S. Chen and Q. Jin, "Multi-modal conditional attention fusion for dimensional emotion prediction," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 2016, pp. 571–575.

[10] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 1033–1038.

[11] J. J. Gross and L. Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

[12] A. Vanzo, D. Croce, and R. Basili, "A context-based model for sentiment analysis in twitter," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 2345–2354.

[13] 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 (Volume 1: Long Papers)*, vol. 1, 2017, pp. 873–883.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[15] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," *arXiv preprint arXiv:1801.10198*, 2018.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[18] R. Zhang, A. Ando, S. Kobashikawa, and Y. Aono, "Interaction and transition model for speech emotion recognition in dialogue." in *INTERSPEECH*, 2017, pp. 1094–1097.

[19] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

[20] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 65–72.

[21] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[23] 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.

[24] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[25] 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*, vol. 42, no. 4, p. 335, 2008.

[26] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[27] 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.

[28] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[29] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, "One billion word benchmark for measuring progress in statistical language modeling," *arXiv preprint arXiv:1312.3005*, 2013.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[31] A. F. Martin and C. S. Greenberg, "The nist 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[32] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation." in *Interspeech*, 2017, pp. 1353–1357.

[33] 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*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4749–4753.

[35] R. Li, Z. Wu, J. Jia, J. Li, W. Chen, and H. Meng, "Inferring user emotive state changes in realistic human-computer conversational dialogs," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 136–144.