Multi-modal Attention for Speech Emotion Recognition

Zexu Pan[1,2], *Zhaojie Luo*[4], *Jichen Yang*[3], *Haizhou Li*[1,3]

[1]Institute of Data Science, NUS, Singapore
[2]Graduate School for Integrative Sciences and Engineering, NUS, Singapore
[3]Department of Electrical and Computer Engineering,
National University of Singapore (NUS), Singapore
[4]Osaka University, Osaka, Japan

pan_zexu@u.nus.edu, luo@irl.sys.es.osaka-u.ac.jp, {eleyji, haizhou.li}@nus.edu.sg

Abstract

Emotion represents an essential aspect of human speech that is manifested in speech prosody. Speech, visual, and textual cues are complementary in human communication. In this paper, we study a hybrid fusion method, referred to as multi-modal attention network (MMAN) to makes use of visual and textual cues in speech emotion recognition. We propose a novel multi-modal attention mechanism, cLSTM-MMA, which facilitates the attention across three modalities and selectively fuse the information. cLSTM-MMA is fused with other uni-modal sub-networks in the late fusion. The experiments show that speech emotion recognition benefits significantly from visual and textual cues, and the proposed cLSTM-MMA alone is as competitive as other fusion methods in terms of accuracy, but with a much more compact network structure. The proposed hybrid network MMAN achieves state-of-the-art performance on IEMOCAP database for emotion recognition.

Index Terms: speech emotion recognition, multi-modal attention, early fusion, hybrid fusion

1. Introduction

Emotions play an important role in speech communication [1]. The recent advancement of artificial intelligence has equipped machines with intelligence quotient. It is equally important for machines to understand emotions, and to improve their emotional intelligence.

The fact that voice call is more informative than text messaging suggests that the affective prosody of speech delivers additional information that includes emotion. Similarly, speaking face-to-face is more effective than text messaging and voice call, which suggests that visual cues play an important role. Humans express emotion through prosody, gesture, and lexical choice. Emotion is quantized by physiological arousal and hedonic valence level [2], which are only partially expressed through speech. The use of specific phrases further indicates our valence level and our body language carries the remaining arousal and valence. It is found that humans rely more on multi-modalities than uni-modal [3] to understand emotions.

Multi-modal speech emotion recognition has been an area of research for decades. *Cho et al.* [4] used text to aid speech in the MCNN network. Similarly *Hossain et al.* [5] and *Xue et al.* [6] used visual cues to augment speech using SVM and SymcHDP networks. It is evident that emotion recognition benefits from the fusion of speech, vision and text information [7–11]. However, it has not been an easy task to fuse the information from different modalities. As the information coming from different modality is neither completely independent nor correlated, the fusion mechanism is expected to pick up the right information from the right modality.

Early or late fusions are the typical options in multi-modal classifier design in emotion recognition. The state-of-the-art method introduced the contextual long short-term memory block (cLSTM) and built a late fusion network (cLSTM-LF) [12,13]. The predictions of uni-modal models are fused to make a final prediction. It is effective at modelling modality-specific interactions but not cross-modal interactions [14].

There are also studies to explore the interaction between modalities with early fusion [15–17]. *Sebastian et al.* [15] concatenated the low-level features and passed them through a convolutional neural network. *Georgiou et al.* [17] concatenated features from different modality at various levels and used multi-layer perceptron for emotion prediction. With early fusion, we are able to explore the interaction between raw features across modalities, that is good. However, the raw features represent different physical properties of the signals in the respective modalities. Therefore, the classifier network will have to learn both the feature abstraction of respective modalities, and the interaction of them at the same time, that is not easy. Furthermore, simple concatenation utilizes whatever information from the input streams that may or may not be relevant to the classification tasks. Early fusion also potentially suppresses modality-specific interactions [18]. In general, concatenation based early fusion methods do not outperform the late fusion methods in emotion recognition [14,19].

Transformer has been effective in natural language processing that features a self-attention mechanism where each input feature embedding is first projected into query, key and value embeddings [20]. In multi-modal situation, the query is from one modality while key and value are from another modality. The attentions between two modalities are computed by cosine similarities between the query and the key. The values are then fused based on the attention scores. The attention mechanism in Transformer is one of the effective solutions to learn cross-modality correlation [21,22]. *Tsai et al.* [21] used directional pairwise cross-modal attention for sentiment analysis. They show positive results with two-modalities attention. In this paper, we would like to explore a mechanism for three-modalities attention for the first time. We believe that speech, visual and text modalities provide complementary evidence for emotion. Three modalities cross-modal attention allows us to take advantage of such evidence.

We propose a multi-modal attention mechanism in place of concatenation to model the correlation between three modalities in cLSTM-MMA. As cLSTM-MMA takes the multi-modal features as input, we consider it as an early fusion sub-network. It

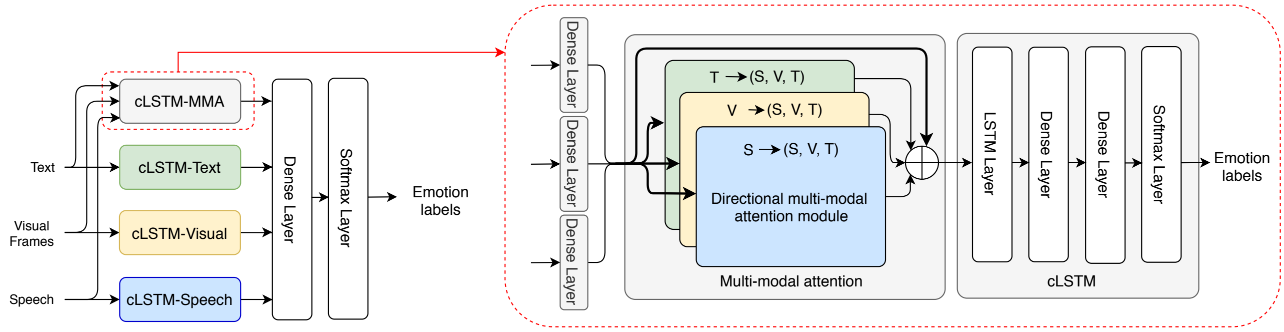


Figure 1: *On the left panel is the proposed multi-modal attention network (MMAN). It consists of a multi-modal attention sub-network (cLSTM-MMA) for early fusion and three uni-modal sub-networks cLSTM-Text, cLSTM-Visual and cLSTM-Speech. The predictions of the four sub-networks are fused with a dense and a softmax layer in late fusion. The architecture of the cLSTM-MMA sub-network is shown in the red dotted box on the right panel. The symbol ⊕ represents concatenation and S, V, T represents speech, visual and text respectively. The cLSTM-MMA consists of three independent dense layers for uni-modal feature embeddings standardisation, multi-modal attention with three parallel directional multi-modal attention modules and finally a cLSTM with one LSTM layer inside.*

consists of three parallel directional multi-modal attention modules for multi-modal fusion. In each module, a query is first computed from a modality. It is then used to compute the cross-modal attention and the self-attention scores to find the relevant information answering to this query. The three parallel modules have distinct queries from three different modalities specifically. Thus, allowing the network to attend for different interactions based on the different queries jointly. The multi-modal attention can be easily scaled up if more than three modalities are present. To take advantage of both the late fusion and early fusion to account for modality-specific and cross-modal interactions, we propose a hybrid multi-modal attention network (MMAN) which fuses the predictions of the cLSTM-MMA and uni-modal cLSTM sub-networks for the final prediction.

The rest of the paper is organized as follows. Section 2 presents the details of the proposed multi-modal attention network. Section 3 describes the experimental setup. Section 4 reports the results and evaluations. Finally, conclusions are drawn in Section 5.

2. Multi-modal attention network

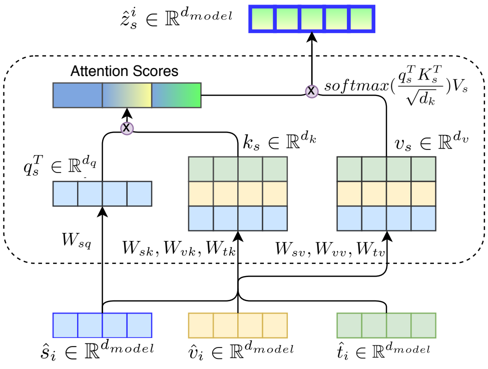
The proposed hybrid fusion network MMAN is shown on the left panel of Figure 1. We have the speech, visual and text feature embeddings of the same utterance as the input. The MMAN consists of a cLSTM multi-modal attention sub-network (cLSTM-MMA) for early fusion, and three uni-modal sub-networks cLSTM-Speech, cLSTM-Visual and cLSTM-Text for late fusion. The outputs of the four sub-networks are fused with a dense and a softmax layer.

2.1. Multi-modal attention

The architecture of cLSTM-MMA sub-network is shown in the red dotted box on the right panel of Figure 1. The cLSTM-MMA consists of three independent dense layers for uni-modal feature embeddings standardisation, multi-modal attention with three parallel directional multi-modal attention modules and finally a cLSTM with one LSTM layer inside.

2.1.1. Modality dimension standardization

The three inputs that represent one utterance are first encoded as the feature embeddings of different dimensions. We first standardize all feature embeddings into the same dimension d_{model} to facilitate the subsequent processing.



$\hat{z}_s^i \in \mathbb{R}^{d_{model}}$
Attention Scores
$softmax(\frac{q_s^T K_s^T}{\sqrt{d_k}})V_s$
$q_s^T \in \mathbb{R}^{d_q}$
$k_s \in \mathbb{R}^{d_k}$
$v_s \in \mathbb{R}^{d_v}$
W_{sq}
W_{sk}, W_{vk}, W_{tk}
W_{sv}, W_{vv}, W_{tv}
$\hat{s}_i \in \mathbb{R}^{d_{model}}$
$\hat{v}_i \in \mathbb{R}^{d_{model}}$
$\hat{t}_i \in \mathbb{R}^{d_{model}}$

Figure 2: *The details of the directional multi-modal attention module $S \rightarrow (S, V, T)$ with query from speech. The inputs to this module are the uni-modal feature embeddings $(\hat{s}_i, \hat{v}_i, \hat{t}_i)$ after the standardization dense layers*

Let's denote the dataset as $\mathcal{D} = \{s_i, v_i, t_i, y_i\}_{i=1:M}$ where s_i, v_i, t_i and y_i represent the speech, visual, text feature embeddings and the emotion labels of utterance i. M is the number of utterances in a conversation. With $s_i \in \mathbb{R}^{d_s}$, $v_i \in \mathbb{R}^{d_v}$ and $t_i \in \mathbb{R}^{d_t}$ where d_s, d_v, d_t are dimensions of corresponding speech, visual and text features. By passing the original feature embeddings through the individual dense feed forward layers as shown in Figure 1, we standardize the outputs into the same dimension $\hat{s}_i \in \mathbb{R}^{d_{model}}$, $\hat{v}_i \in \mathbb{R}^{d_{model}}$ and $\hat{t}_i \in \mathbb{R}^{d_{model}}$.

2.1.2. Directional multi-modal attention module

Taking the directional multi-modal attention module with speech query for illustration. It is represented by $S \rightarrow (S, V, T)$ as shown in the blue module in Figure 1. This module computes the directional attention from speech to visual and text as well as the self-attention of speech. The detail of this speech query module is illustrated in Figure 2.

We use the query, key and value representation to compute the attention. We compute the query of speech q_s through a learnable weights $W_{sq} \in \mathbb{R}^{d_{model} \times d_q}$ as shown in Equation 1.

$$q_s = {W_{sq}}^T \hat{s}_i \tag{1}$$

where d_q is the dimension of the query vector.

The keys K_s and values V_s are computed using learnable weights $W_{sk}, W_{vk}, W_{tk} \in \mathbb{R}^{d_{model} \times d_k}, W_{sv}, W_{vv}, W_{tv} \in \mathbb{R}^{d_{model} \times d_v}$, where d_k, d_v are dimensions of key and value vector. The computation is shown in Equation 2 and 3.

$$K_s = concat\{\hat{s}_i^T W_{sk}, \hat{v}_i^T W_{vk}, \hat{t}_i^T W_{tk}\} \tag{2}$$

$$V_s = concat\{\hat{s}_i^T W_{sv}, \hat{v}_i^T W_{vv}, \hat{t}_i^T W_{tv}\} \tag{3}$$

The cross-modal and self attention scores are computed by the dot product of the query q_s and keys K_s. It is then used to compute the weighted sum of the values $\hat{z}_s^i$, which represents the interaction of different modalities answering to speech query. The directional multi-modal attention from speech query $D_{S \to (S,V,T)}$ is given in Equation 4 and illustrated in Figure 2.

$$\begin{aligned} \hat{z}_s^i &= D_{S \to (S,V,T)}(\hat{s}_i, \hat{v}_i, \hat{t}_i) \\ &= softmax(\frac{q_s^T K_s^T}{\sqrt{d_k}})V_s \end{aligned} \tag{4}$$

The same computing procedure is applied to text and visual directional multi-modal attention modules except that each module has its own learnable weights computing the query to facilitate the learning of different interactions based on different directional queries. The outputs from three parallel attention modules are concatenated with a skip connection.

2.1.3. Contextual long short-term memory block

The output from multi-modal attention is passed through a cLSTM block with one LSTM layer as shown in Figure 1 to capture the contextual cues between consecutive utterances in a conversation [13].

2.2. Uni-modal sub-networks

The cLSTM-Speech, cLSTM-Visual and cLSTM-Text sub-networks are all built using cLSTM block with two LSTM layers except that their inputs are different. Their network hyper-parameters are customized to suit different modalities. The cLSTM-MMA and three uni-modal sub-networks are separately trained. Their weights are fixed during the training of the late fusion dense layer in the MMAN.

3. Experimental setup

3.1. Dataset

The IEMOCAP dataset [23] is used to evaluate the proposed network. The dataset contains 10K videos split into 5 minutes of dyadic conversations for human emotion analysis. Each conversation is split into spoken utterance. Each utterance consists of corresponding transcription, speech waveform and visual frames. To align with previous works, we consider the emotion classes of angry, happy (excited), sad (frustrated) and neutral for multi-class classification but without excited and frustrated for binary sentiment classification system. The train and the test sets are disjoint for speakers. The speakers in the training set are not contained in the test set as we assume the speakers are unknown at the inference time. The details of the dataset are provided in Table 1.

Table 1: *The number of utterances labelled happy (HPY), sad, neutral (NEU), angry (ANG), excited (EXC) and frustrated (FRU) in the training and testing set of IEMOCAP*

	HPY	SAD	NEU	ANG	EXC	FRU
Train	504	839	1324	933	742	1468
Test	144	245	384	170	299	381

3.2. Uni-modal feature extraction

We follow *Poria et al.* for low level feature extraction [13]. The input video of an utterance is first separated into corresponding text, video frames and speech modalities and extraction is done by using individual pre-trained networks transferred from other tasks. The feature of each utterance is extracted as a fixed-length vector for each modality.

Speech: OpenSMILE toolkit [24] with IS13-ComParE [25] is used to for feature extraction. It is performed with 30 Hz frame-rate and 100 ms sliding window. The features include Mel Frequency Cepstral Coefficient (MFCC), spectral centroid, spectral flux, beat histogram, beat sum, voice intensity, pitch, mean and root quadratic mean, etc [12].

Visual: We use a 3D-CNN [26] pre-trained from human action recognition to extract their body language. The 3D-CNN is applied to the consecutive visual frames of the speaker's upper body. It learns the relevant features of each frame and the changes among the given number of consecutive frames, which are the motion cues.

Text: Word2vec [27] is used to embed each word of an utterance's transcript into word2Vec vectors. The embedded words are concatenated, padded and standardized to a 1-dimensional vector by passing through a CNN [28].

3.3. Reference baselines

Three baselines are constructed from the state-of-the-art model [13]. They are all built based on cLSTM block. The first baseline uses speech data only while the other two uses speech, visual and text data.

Speech-only cLSTM (cLSTM-Speech): The speech-only baseline receives speech features only, the speech features are passed through a cLSTM block with two LSTM layers for prediction.

Multi-modal cLSTM with early fusion (cLSTM-EF): The cLSTM-EF baseline receives concatenated speech, visual and text feature embeddings as input. The concatenated features are passed through a cLSTM block with two LSTM layers for prediction.

Multi-modal cLSTM with late fusion (cLSTM-LF): The cLSTM-LF baseline has a hierarchical structure. The lower level consists of three uni-modal networks cLSTM-Speech, cLSTM-Text and cLSTM-visual. At the higher level, the predictions of the three uni-modal networks are concatenated and passed through another cLSTM block for final prediction.

4. Evaluation results

4.1. Comparison with baselines

We reported the results of our proposed models and baselines of multi-class recognition systems in terms of accuracy and recall rates in Table 2 and Figure 3.

Table 2: *Accuracy (%) and number of network parameters of the baselines and our models in multi-class classification*

Model	Accuracy	No. of Parameters (million)
cLSTM-Speech	57.12	**1.27**
cLSTM-EF	69.75	2.00
cLSTM-LF	71.78	4.73
cLSTM-MMA	71.66	**1.27**
MMAN	**73.94**	4.39

Table 3: *A comparative study of binary sentiment analysis with different multi-modal implementations in terms of classification accuracy (%)*

Model	Happy	Sad	Neutral	Angry
MFM [29]	90.2	**88.4**	72.1	87.5
RAVEN [14]	87.3	86.2	69.7	87.3
FMT [30]	88.8	88.0	74.0	89.7
MulT [22]	90.7	86.7	72.4	87.4
cLSTM-MMA	92.15	82.93	**76.99**	**93.32**
MMAN	**92.68**	83.14	75.61	92.90

4.1.1. Benefits of textual and visual cues

We first presented the recognition accuracy of the Speech-only baseline cLSTM-Speech, which obtains a recognition accuracy of 57% which is more than 10 absolute percentage points lower than any other the multi-modal methods in Table 2.

We compared its confusion matrix with our cLSTM-MMA model in Figure 3, since they have similar network size. cLSTM-Speech's recall rates for neutral is very low but very high for sad as see in Figure 3. This unbalance phenomenon is alleviated by cLSTM-MMA, which uses multi-modal information as seen from the cLSTM-MMA confusion matrix. This shows that the visual and textual cues do complement speech's ambiguity in emotion recognition.

4.1.2. Multi-modal attention vs concatenation

The cLSTM-MMA is 2% higher than cLSTM-EF in terms of accuracy as shown in Table 2. This means that the proposed multi-modal attention is more prevalent in computing the interaction between modalities compared to concatenation method with early fusion. Besides, the cLSTM-MMA has 40% fewer parameters compared to the cLSTM-EF baseline.

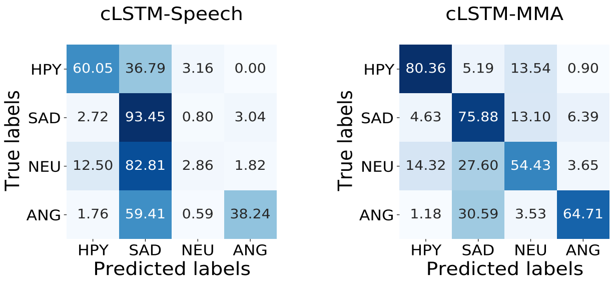


Figure 3: *Normalised confusion matrix of the Speech-only baseline cLSTM-Speech and proposed cLSTM-MMA network. Diagonal entries represent the recall rates of each emotion.*

Table 4: *A comparative study of multi-class emotion recognition with different multi-modal implementations*

Model	Accuracy (%)
Rozgic et al. [31]	69.4
Poria et al. [19]	71.59
Tripathi et al. [12]	71.04
MMAN	**73.94**

4.1.3. Comparison with late fusion

The cLSTM-MMA achieves comparable accuracy with the state-of-the-art late-fusion model cLSTM-LF with only a quarter of its' parameters as shown in Table 2. The proposed hybrid MMAN network outperforms all the multi-modal networks and achieves the state-of-the-art accuracy of 73.98% using the same amount of parameters as cLSTM-LF, suggesting that both modality-specific and cross-modal interactions are important in emotion recognition.

4.2. Comparison with previous works

We compared the accuracies of our model with other binary sentiment classification systems using speech, visual and text in Table 3. The cLSTM-MMA has superior performance over the pairwise correlation network MulT and others (with the exception of the sad emotion). Showing that correlation between three modalities is superior then pairwise correlation. Interestingly, MMAN has similar performance with cLSTM-MMA, suggesting that modality-specific interaction may not contribute much in binary sentiment classification case.

Table 4 summaries the performance of previous multi-class emotion recognition network using speech visual and text. Our proposed MMAN achieves a state-of-the-art result of 73.94% on dataset IEMOCAP. Also, most of the methods of the previous work that achieved comparable results are based on BLSTM which have access to future utterance information when deciding for the current utterance, thus the comparison reported in this paper is in their favour. Nevertheless, MANN outperforms all reference baselines.

5. Conclusion

In this work, we presented a hybrid fusion model MMAN using visual and textual cues to aid speech in emotion recognition. We proposed the multi-modal attention in early fusion which features parallel directional attention between modalities in place of concatenation. The attention mechanism enables better data association between modalities and has a significantly less amount of parameters needed. Through experiments, we showed that the multi-modal attention alone is as competitive as other fusion methods with a much more compact network. Our hybrid model achieved the state-of-the-art result on IMEOCAP dataset for emotion recognition.

6. Acknowledgement

This research work is partially supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), and in part by Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) from the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme.

7. References

[1] M. Sreeshakthy and J. Preethi, "Classification of human emotion from deap eeg signal using hybrid improved neural networks with cuckoo search," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol. 6, no. 3-4, pp. 60–73, 2016.

[2] L. F. Barrett, "Solving the emotion paradox: Categorization and the experience of emotion," *Personality and social psychology review*, vol. 10, no. 1, pp. 20–46, 2006.

[3] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Current opinion in neurobiology*, vol. 11, no. 4, pp. 505–509, 2001.

[4] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *arXiv preprint arXiv:1911.00432*, 2019.

[5] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[6] J. Xue, Z. Luo, K. Eguchi, T. Takiguchi, and T. Omoto, "A bayesian nonparametric multimodal data modeling framework for video emotion recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 601–606.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[8] 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.

[9] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[10] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

[11] 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.

[12] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2019.

[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)*, 2017, pp. 873–883.

[14] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7216–7223.

[15] J. Sebastian and P. Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," in *Proc. Interspeech*, 2019, pp. 51–55.

[16] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[17] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," *Proc. Interspeech 2019*, pp. 1646–1650, 2019.

[18] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and 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 (Volume 1: Long Papers)*, 2018, pp. 2247–2256.

[19] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[20] 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.

[21] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[22] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5612–5623.

[23] 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.

[24] 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.

[25] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[26] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[27] G. C. J. D. Tomas Mikolov, Kai Chen, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[29] Y. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *ICLR*, 2019.

[30] A. Zadeh, C. Mao, K. Shi, Y. Zhang, P. P. Liang, S. Poria, and L.-P. Morency, "Factorized multimodal transformer for multimodal sequential learning," *arXiv preprint arXiv:1911.09826*, 2019.

[31] 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.