Multi-Head Attention with Disagreement Regularization for Multimodal Sentiment Analysis

Hao Ai*
Center for Image and Information Processing, Xi'an University of Posts and Telecommunications
ah217846@163.com

Ying Liu
Center for Image and Information Processing, Xi'an University of Posts and Telecommunications
liuying_ciip@163.com

Jie Fang
Center for Image and Information Processing, Xi'an University of Posts and Telecommunications
jackfang713508@gmail.com

ABSTRACT

Multimodal sentiment analysis refers to the use of computers to analyze and recognize the emotions that people want to express through the extracted multimodal emotional features. It plays a major role in human-computer interaction and financial market forecasting. Most of the existing multimodal sentiment analysis method use contextual information for modeling, but the feature information is not abundant enough. there is no problem of mining useful sentiment feature information. This paper proposes a multimodal sentiment analysis method based on the multi-head attention mechanism. The method uses Bi-GRU to capture contextual relationships, and adopts the constrained multi-head attention mechanism to focus on learning different modalities the feature representation of the subspace. We evaluated our proposed method on two multimodal sentiment analysis benchmark datasets, CMU-MOSI and CMU-MOSEI, and achieved accuracy of 82.63% and 81.82% respectively, proving the effectiveness of our method.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**;

KEYWORDS

Multimodal sentiment analysis, Bidirectional Gated Recurrent Units, Emotional features, Multi-head attention mechanism

ACM Reference Format:
Hao Ai, Ying Liu, and Jie Fang. 2021. Multi-Head Attention with Disagreement Regularization for Multimodal Sentiment Analysis. In *2021 4th International Conference on Artificial Intelligence and Pattern Recognition (AIPR 2021), September 24–26, 2021, Xiamen, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3488933.3489004

1 INTRODUCTION

With the development of mobile we-media platforms such as YouTube, Douyin, and twitter, sentiment analysis has become a research hotspot. Unimodal sentiment analysis has achieved good results. For example, for text sentiment recognition, reference [1] draws on the idea of capsule network, and proposes RNN-based capsules for emotion recognition by constructing capsules for each emotion category. Reference [2] proposed using the capsule network to construct vector-based feature representations and cluster features through the EM routing algorithm. Reference [3] proposes a hierarchical transformer method, in which the low-level transformer is used to model word-level input, and the upper-level transformer is used to capture the context of discourse-level embedding. For acoustic sentiment analysis, reference [4] proposed using two cyclically connected capsule networks to extract features based on the capsule network to enhance spatial sensitivity and achieve better classification accuracy than the CNN-LSTM architecture. For visual sentiment recognition, GAN (Generative Adversarial Networks) is used in reference [5] and reference [6] for facial expression recognition due to its powerful generation ability.

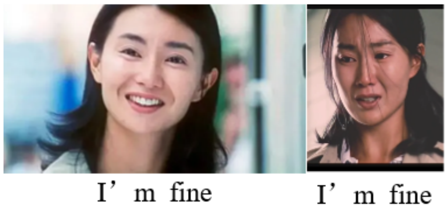

Figure 1: Limitations of unimodal sentiment analysis.

Although unimodal analysis recognition has achieved good results, there are still limitations [29], as shown in 1. As shown in the figure, we can clearly see that the sentiment of the characters on the left are positive, and the sentiment of the characters on the right are negative, but only through the text for sentiment polarity analysis (positive, negative and neutral) or emotional analysis (sad, happy, etc.), the same analysis result will be obtained, which is undoubtedly a very obvious error, so multimodal emotion recognition is particularly important.

Video provides a good source for extracting multimodal information. In addition to visual information, it also provides information on spoken acoustics and textual expressions. Reference [7] uses two independent CNNs to extract image and text features in tweets respectively, and input them into the third CNN to learn the emotional connection between the modalities. Similarly, reference [8] uses CNN to fuse image and text features to achieve the purpose of microblog emotion recognition. Reference [9] proposes a tensor fusion network (TFN) for unstable spoken language and accompanying gestures and sounds in online video, this method takes the

[*] liuying_ciip@163.com.

AIPR 2021, September 24–26, 2021, Xiamen, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8408-7/21/09…$15.00
https://doi.org/10.1145/3488933.3489004

multimodal emotion analysis problem as the dynamic inter-model and intra-model, and adopts the end-to-end learning method, which can solve the multimodal emotion analysis problem of inter-model and intra-model.

In previous studies, the discourse was only regarded as an independent part without considering the relationship between the video pictures before and after the discourse, but in fact, for a sentence, the picture changes before and after also have a certain internal relationship [10]. Therefore, the reference [11] uses 3 GRUs to modal the speaker information, the context information of the previous dialogue and the emotional information. This structure is divided into three parts: Global GRU, Party GRU and Emotion GRU. Global GRU is used to calculate and update the Global State at each moment. Party GRU is used to calculate and update the status of the speaker at the current moment. Emotion GRU is used to calculate the emotional representation of the current dialogue content. Reference [12] proposes a session level multimodal emotion recognition method based on transformer. The structure includes two transformers. One transformer is used to capture the time-dependent tolerance between unimodal features, and the other transformer is used to learn the multimodal interaction of misalignment multimodal features. Reference [13] proposed a multimodal sentiment analysis method that uses Bi-GRU and self-attention mechanism to capture contextual relationships. Since self-attention cannot learn features from different subspaces, there is a problem of insufficient feature expression ability [14]. Multi-head Attention [14] It is an extension of self-attention, which can calculate attention from multiple different dimensions, so that the model can learn features in different subspaces. Reference [14] pointed out that the use of single-head attention mechanism can learn some long distance dependency of words within a sentence, and multi-head attention can not only strengthen this learning ability, but even understand the syntax and semantic structure of the sentence. To solve the above problems, this paper proposes a multimodal sentiment analysis method based on multi-head attention mechanism. And in this method, three multimodal information, Vt (Vision), At (Acoustics), Tt (Text), are used to predict sentiment better.

The overall structure of this paper is as follows: Section 2 discusses the methods proposed in this paper in detail; Section 3 conducts experimental analysis; Section 4 summarizes the full paper.

2 ALGORITHM DESCRIPTION IN THIS PAPER

This paper proposes a multimodal sentiment analysis method based on the multi-head attention mechanism. First, extracting the original features of the three modalities, and then inputting the original features of the three modalities into the Bi-GRU network to obtain the emotional representation of the corresponding modal. Then it is input into multi-head attention mechanism through the dense layer. The purpose of the dense layer is to transform the previously extracted features in the dense layer through nonlinear changes, extracting the associations between these features, and then mapping them to the output space. The output representation of the dense layer is connected with the output after multi-head attention mechanism. The purpose is to raise the gradient flow to the lower layer, and finally send it to softmax for sentiment analysis. The overall architecture is shown in Figure 2.

2.1 Feature Extraction

In the feature extraction stage, we extract trimodal information of text, acoustics and vision in the two datasets of MOSI and MOSEI.

For the MOSEI and MOSI datasets, we use the pre-trained Glove [15] word embedding model to encode each word into a word vector with a dimension of 300 for the text information. Visual features are extracted using 3D-CNN [28]. We believe that 3D-CNN can not only learn relevant features from each frame, but also learn the correlation between consecutive frames. Acoustic features We use COVAREP [16] to extract shallow features, including Mel-scale Frequency Cepstral Coefficients (MFCC), pitch tracking, voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters, maximum dispersion quotients.

2.2 Context Modeling - Bi-GRU

GRU (Gate Recurrent Unit) [19] is similar to LSTM (Long-Short Term Memory) [18]. It can effectively capture the context information and is a variant of Recurrent Neural Network (RNN) [17], Which is designed to solve the problems of gradient explosion and gradient disappearance in long-term memory and backpropagation. GRU and LSTM behave very similarly in most cases, but the calculation is easier. The structure is as shown in the 3.

As shown in 3, the GRU first obtains the reset gating r_t and the update gating z_t through the previous transmiss in state h_{t-1} and the input x_t of the current node. σ is the sigmoid function, through this function, the data can be transformed to the range of 0~1 to act as a gating signal, reset the gating and update the gating calculation as shown in equations (1) and (2).

$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right) \tag{1}$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right) \tag{2}$$

After getting the gating signal, first use the reset gating to get the data after "reset" $h'_{t-1} = h_{t-1} \odot r_t$, Then h'_{t-1} is spliced with the input x_t, and then a tanh activation function is used to scale the data to the range of -1~1 to obtain $\widetilde{h_t}$. $\widetilde{h_t}$ mainly contains the current input x_t data, and $\widetilde{h_t}$ is added to the current hidden state in a targeted manner, which is equivalent to "memory the state at the previous moment", and the calculation is as shown in equation (3).

$$\widetilde{h_t} = tanh\left(W \cdot \left[r_t \odot h_{t-1}, x^t\right]\right) \tag{3}$$

After obtaining $\widetilde{h_t}$, the "memory update" stage is finally carried out. At this stage, the two steps of forgetting and remembering are carried out at the same time, and the calculation is as shown in equation (4).

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \tag{4}$$

The gating signal z_t here ranges from 0 to 1. The closer to 1, the more data "remembered", and the closer to 0, the more "forgotten". $\odot$ in all the above formulas represents the multiplication of the corresponding elements in the operation matrix.

In the unidirectional GRU network structure, data is always transmitted sequentially from front to back in one direction. If the state can be transferred from the previous moment and the next

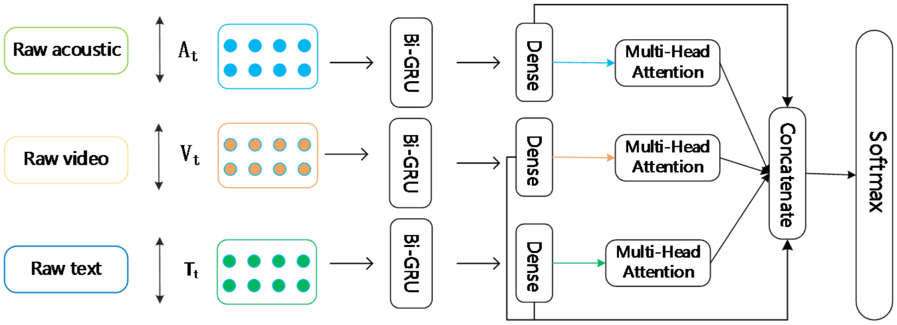

Figure 2: Overall architecture of the proposed method.

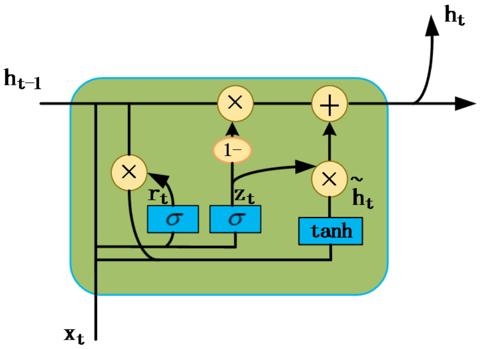

Figure 3: Structure of GRU.

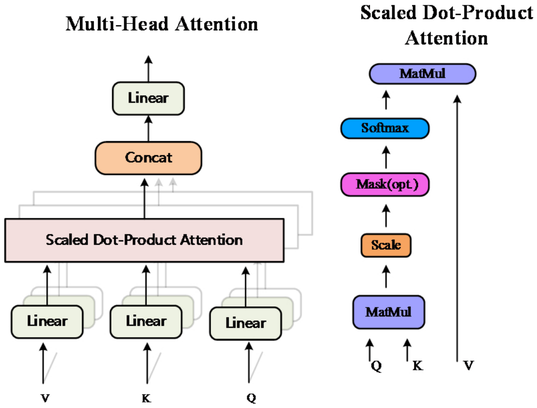

Figure 4: Multi-Head Attention.

moment together, it will be more helpful to extract deep-level features, from this, Bi-GRU is composed of two GRU units in opposite directions, and the output state is jointly determined by the two GRUs.

2.3 Multi-Head Attention

The Multi-head attention model is composed of multiple scaled dot-product attention basic units stacked, and its structure is shown in 4.

The input matrix has three values of Q, K, and V. There are h layers in the Scaled Dot-Product Attention section, and the attention calculation expression for each layer is shown in equation (5).

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

Where d is the number of hidden units in the neural network. Because of the self-attention mechanism adopted by multi-head attention, the input Q = K = V. In the multi-head attention model, firstly, the Q, K, and V vectors are linearly transformed. secondly, Q and each K use the dot product similarity function to calculate their weights, scale by dividing by a K dimension to avoid the inner product value being too large, and use the softmax function to normalize these weights. Finally, the weight and the corresponding key value are weighted and summed to obtain the Attention. After h times of zooming attention calculations, multiple heads are obtained, and the heads of each time are spliced, and the final result MultiHead is obtained through linear transformation, the calculation formulas are shown in equation (6) and Equation (7).

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6}$$

$$MultiHead\,(Q, K, V) = Concat\,(head_1, \ldots, head_h)\,W^o \tag{7}$$

Among them, W^o represents the weight of the linear transformation, and $MultiHead(Q, K, V)$ represents the result of the final calculation. Through multiple Attention calculations, the model can learn more feature information from different subspaces.

If multiple heads are set in Multi-Head Attention to allow the model to focus on different information angles, but in specific practice, if the head of each attention grows natively, it may cause information redundancy, that is, multiple head extraction the features are very consistent.

In order to allow each attention to focus on different information angles, we use the method proposed in [16] to constrain multiple heads, that is, increase the difference in attention weights among different heads.

Table 1: Experimental results of different modalities numbers on MOSI and MOSEI datasets

Number of modalities	Tt	Vt	At	Acc%(MOSI)	Acc%(MOSEI)
Unimodal	√	-	-	80.24	79.21
	-	√	-	65.22	75.63
	-	-	√	64.20	76.44
Bimodal	√	√	-	80.50	79.88
	√	-	√	80.12	80.21
	-	√	√	65.46	78.92
Trimodal	√	√	√	82.63	81.82

3 EXPERIMENT AND RESULT ANALYSIS

3.1 Experimental Datasets

We evaluate our method on two multimodal sentiment analysis benchmark datasets, CMU-MOSI (CMU Multi-modal Opinion-level Sentiment Intensity) [20] and CMU-MOSEI (CMU Multi-modal Opinion Sentiment and Emotion Intensity) [15].

The CMU-MOSI dataset collects video blogs (vlogs) on YouTube mainly about movie review videos. The length of the video ranges from 2 to 5 minutes. A total of 93 videos are randomly collected. These videos are labeled as seven types of emotional tendencies ranging from -3 to +3. The emotional annotation of the data set is not the viewer's feeling, but the emotional tendency of the commenter in the annotation of the video. In this paper, in order to conduct an experimental comparison with the methods proposed in the relevant literature, we divide the emotional tendency into two categories. If it is greater than or equal to 0, it is judged as positive sentiment, and if it is less than 0, it is judged as negative sentiment.

The CMU-MOSEI dataset collects data from YouTube monologue videos and removes videos that contain too many characters. The final dataset contains 3228 videos, 23453 sentences, 1000 narrators and 250 topics, with a total time of up to 65 hours. The dataset has both sentiment annotations and emotion annotations. Sentiment labeling is an emotional labeling of 7 categories for each sentence, and the author also provides labels for 2/5/7 categories. Emotion labels are emotional labels that include happiness, sadness, anger, fear, disgust, and surprise.

3.2 Experiment

We use accuracy scores as the model's performance evaluation indicators on two multimodal sentiment analysis benchmark data sets. In terms of experimental settings, time_steps is set to 63, that is, the length of the sequence itself is set to 63. We use 3 Bi-GRUs with 300 neurons to capture the context, each Bi-GRU is followed by a Dense layer with 100 neurons, the purpose is to project the features of the three modal inputs to the same dimension. Dropout is set to 0.5 (MOSI) and 0.3 (MOSEI) for regularization to prevent over-fitting. At the same time, in order to reduce the dependence between Bi-GRU neurons, the dropout used in the Bi-GRU layer is 0.4 (MOSI) and 0.3 (MOSEI). For the Dense layer, we use the ReLu activation function. In order to ensure that Multi-Head Attention pays attention to different angles of information, we use the method provided in [17] to constrain Multi-Head Attention, and use softmax in the classification layer for final classification. For network training, the batch size is set to 32, Adam optimizer with cross-entropy loss function is used, and 50 epochs are trained, and the average accuracy of 5 trainings is taken as our final result.

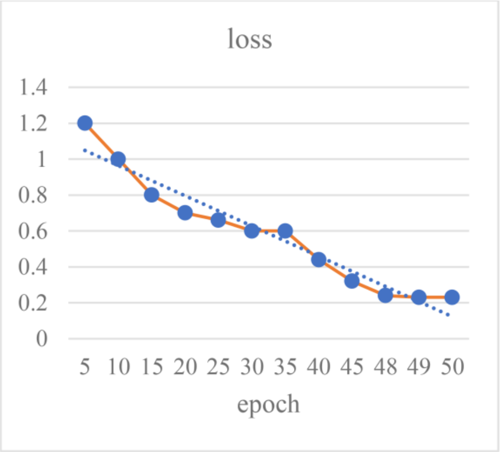

Figure 5: loss curve.

We conduct separate experiments on the three modalities of text, hearing, and vision. At the same time, any two modalities are also tested separately. Finally, the three modalities are used for experiments. The experimental results are shown in Table 1.

We also give the change of loss during the training process to prove the stability of the network, as shown in Figure 5.

At the same time, from the above introduction to the structure of the multi-head attention mechanism, it can be seen that the K value must be evenly divided by the number of dimensions of the feature vector, after passing through the Dense layer, the dimensionality of the feature vector is 100. We choose 1, 2, 4, 10, 20, and 25 to conduct experiments to select the optimal K value. It is obtained through experiments that when the K value is equal to 4, the best accuracy is obtained, as shown in Table 2.

We give the experimental results that do not restrict multi-head attention. Through the experimental results, it can be seen that when the K value is 1, 2, 4, the similarity of the information area of each head's attention is very high, which proves the effectiveness of constraining multi-head, As shown in Table 3.

Table 2: Experimental results of different K values on different datasets

K	1	2	4	10	20	25
Acc%(MOSI)	79.36	79.89	82.63	76.42	72.25	67.44
Acc%(MOSEI)	78.14	79.12	81.82	76.22	70.86	64.12

Table 3: Experimental results that do not restrict multi-head attention

K	1	2	4	10	20	25
Acc%(MOSI)	79.26	79.46	79.55	74.21	73.12	65.46
Acc%(MOSEI)	78.26	78.37	79.12	77.46	76.21	66.64

Table 4: Comparison of experimental results of different method

CMU-MOSEI						CMU-MOSI					
algorithm	[18]	[21]	[22]	[7]	our	[18]	[19]	[20]	[21]	[7]	our
Acc%	77.64	76.0	76.4	79.63	81.82	80.3	81.3	76.5	77.4	80.58	82.63

3.3 Comparative analysis of experiments

For the MOSI dataset, we compare our method with the methods proposed in [13] and [23]-[26]. Reference [13] proposed a multimodal sentiment analysis framework that captures contextual relationships through Bi-GRU and self-attention mechanism. Reference [23] proposed a model based on LSTM to enable this model to capture contextual information in the discourse environment. Reference [24] proposed a tensor-level fusion technique for combining all three modalities. Reference [25] proposed a word-level fusion model (GME-LSTM) for multimodal input with temporal attention. Reference [26] proposed to use multi-level attention mechanism to extract different modalities interaction information. For the MOSEI dataset, in addition to reference [13], [23], [26], we also compared the combination of modality fusion and coding proposed in reference [27] and the modalities are independent of each other during the coding process. The experimental results are shown in Table 4.

As shown in the table, on the MOSEI and MOSI datasets, the methods proposed by reference [23] achieved accuracy of 77.64% and 80.3% respectively, and reference [26] achieved accuracy of 76.0% and 77.4% respectively. The methods proposed in [24] and [25] achieved accuracy of 81.3% and 76.5% on the MOSI dataset, respectively. Reference [27] achieved 76.4% accuracy on the MOSEI dataset. By comparing with the method proposed in [13], the performance of our proposed method on the two datasets is 2%-3% higher, it shows that the performance of the constrained multi-head attention is better than that of a single self-attention.

4 CONCLUSION

This paper proposes a multimodal sentiment analysis method based on the multi-head attention mechanism, which uses Bi-GRU to capture contextual information. In addition, the method focuses on learning the feature representations of different subspaces in different modalities through the constrained multi-head attention mechanism, and learns the contribution features among them. Experiments show that Bi-GRU can effectively capture contextual information, and the constrained multi-head attention mechanism can focus on more useful emotional feature information and avoid the situation that multiple heads pay attention to the same area and cause information redundancy. It has achieved better results than a single self-attention mechanism on the MOSEI and MOSI datasets.

REFERENCES

[1] Yequan Wang, Aixin Sun, Jialong Han, Ying Liu and Xiaoyan Zhu. 2018. Sentiment analysis by capsules. Proceedings of the 2018 world wide web conference. 1165-1174.

[2] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, Jianxin Liao, Tong Xu and Ming Liu. 2019. Capsule network with interactive attention for aspect-level sentiment classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 5492-5501.

[3] Qingbiao Li, Chunhua Wu, Zhe Wang and Kangfeng Zhen. 2020. Hierarchical transformer network for utterance-level emotion recognition. Applied Sciences 10, 13 (June 2020), 4447-4460. https://doi.org/10.3390/app10134447

[4] Xinxi Wu, Songxiang Liu, Yuewen Cao, Xu Li, Jianwei Yu, Dongyan Dai, *et al.* 2019. Speech emotion recognition using capsule networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brighton, UK, 6695-6699.

[5] Haiyong Wang and Hongzhu Liang. 2020. Local occlusion facial expression recognition based on improved GAN. Computer Engineering and Application 56, 5 (2020), 141-146.

[6] Naiming Yao, Qingpei Guo and Fengchun Qiao. 2018. Robust facial expression recognition based on Generative adversary Network. Acta Automatica Sinica 44, 5 (2018), 865-877.

[7] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60, 6 (2017), 84-90.

[8] Yuhai Yu, Hongfei Lin, Jiana Meng and Zhehuan Zhao. 2016. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms 9, 2 (2016), 0-11.

[9] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria and Louis-Philippe Analysis. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250(2017).

[10] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. Proceedings of the 55th annual meeting of the association for computational linguistics. 873-883.

[11] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. Proceedings of the AAAI Conference on Artificial Intelligence. 6818-6825.

[12] Zheng Lian, Bin Liu and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021), 985-1000.

[13] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. Proceedings of the 2018 conference on empirical methods in natural language processing. 3454-3466.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Lukasz Kaiser, *et al.* 2017. Attention is all you need. Advances in neural information processing systems. 5998-6008.

[15] Jeffrey Pennington, Richard Socher and Christopher Manning. 2014. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 1532-1543.

[16] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Florence, Italy, 960-964.

[17] Zachary C. Lipton, John Berkowitz and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019(2015).

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (November 1997), 1735-1780.

[19] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078(2014).

[20] Amir Zadeh, Rowan Zellers, Eli Pincus and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems 31, 6 (November 2016), 82-88.

[21] Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, *et al.* 2018. 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. 2236-2246.

[22] Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu and Tong Zhang. 2018. Multi-head attention with disagreement regularization. arXiv preprint arXiv: 1810.10183(2018).

[23] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. Proceedings of the 55th annual meeting of the association for computational linguistics. 873-883.

[24] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. IEEE International Conference on Data Mining (ICDM). IEEE, New Orleans, LA, USA, 1033-1038.

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

[26] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. Thirty-Second AAAI Conference on Artificial Intelligence. 5642-5649.

[27] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanva Poria, Erik Cambria and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. Proceedings of the AAAI Conference on Artificial Intelligence. 5634-5641.

[28] Shuiwang Ji, Wei Xu, Ming Yang and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 35, 1 (March 2012), 221-231.

[29] Md Shad AKhtar, Dushyant Singh Chauhan, Asif Ekbal. 2020. A deep multi-task contextual attention framework for multi-modal affect analysis. ACM Transactions on Knowledge Discovery from Data (TKDD) 14, 3 (May 2020), 1-27.