Multimodal sentiment analysis based on multi-head attention mechanism

Chen Xi
College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
1018010624@njupt.edu.cn

Guanming Lu*
College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
lugm@njupt.edu.cn

Jingjie Yan
College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
yanjingjie@njupt.edu.cn

ABSTRACT
Multimodal sentiment analysis is still a promising area of research, which has many issues needed to be addressed. Among them, extracting reasonable unimodal features and designing a robust multimodal sentiment analysis model is the most basic problem. This paper presents some novel ways of extracting sentiment features from visual, audio and text, furthermore use these features to verify the multimodal sentiment analysis model based on multi-head attention mechanism. The proposed model is evaluated on Multimodal Opinion Utterances Dataset (MOUD) corpus and CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus for multimodal sentiment analysis. Experimental results prove the effectiveness of the proposed approach. The accuracy of the MOUD and MOSI datasets is 90.43% and 82.71%, respectively. Compared to the state-of-the-art models, the improvement of the performance are approximately 2 and 0.4 points.

CCS Concepts
• Computing methodologies → Artificial intelligence → Natural language processing → Discourse, dialogue and pragmatics

Keywords
Feature extraction; Multimodal sentiment analysis; Multi-head attention mechanism.

1. INTRODUCTION
Sentiment analysis [1-2] was originally a task of analyzing sentiment state from text, which includes reviews and opinions from various forums on products, movies, etc. However, with the prosperity of multimedia social platforms such as YouTube, Facebook, and Twitter [3-4], people are increasingly keen to upload videos containing their own opinions and reviews. As a result, the appearance of these videos enables the sentiment analysis evolved from unimodal sentiment analysis to multimodal sentiment analysis that integrates visual, audio, and text. Compared to the text-based sentiment analysis, multimodal sentiment analysis contains more visual and audio information that can help improve the effectiveness of the sentiment analysis. In recent years, research on multimodal sentiment analysis mainly on utterance-level multimodal sentiment analysis [5], which refers to the sentiment analysis of each sentence in the video with a sentiment label (rather than just assigning a unique label to the whole video). In particular, utterance-level sentiment analysis is useful to understand the sentiment dynamics of different aspects of the topics covered by the speaker throughout his/her speech [6].

The development in machine learning methods promotes the advance in multimodal sentiment analysis. Especially the deep neural network-based approaches have reached the state-of-the-art in many multimodal sentiment analysis tasks [7]. However, the development of multimodal sentiment analysis still faces many problems. One is to extract reasonable unimodal features from the raw video data. On the other hand, the lack of a robust multimodal sentiment analysis model will lead to the underutilization of unimodal features.

This paper tries to improve these two issues. As the limited number of existing tagged multimodal sentiment analysis datasets, transfer learning [8-9] is considered to extract sentiment features. The combination of pre-trained VGG16 model and LSTM is used to extract visual sentiment features, which has achieved good results in human action recognition and emotion recognition [10-11]. The pre-trained model BERT [12] is applied to extract text features and the model has refreshed the state-of-the-art in many NLP tasks. Due to the lack of transfer model, the combination of spectrogram and convolutional neural network is used to extract audio features [13-14]. Moreover, a multimodal sentiment analysis network based on multi-head attention mechanism [15] is developed. The multi-head self-attention mechanism network is used to analyze the extracted unimodal sentiment features, and the unimodal features through the multi-headed self-attention mechanism network are called multi-head self-attention unimodal features. Then use the multi-head mutual attention mechanism network to analyze the correlation between different modalities, and the bi-modal feature through the multi-head mutual attention mechanism network is called the multi-head mutual attention features. Finally, multi-modal self-attention unimodal features

ICMLSC 2020, January 17–19, 2020, Haiphong City, Viet Nam

ACM ISBN 978-1-4503-7631-0/20/01…$15.00
https://doi.org/10.1145/ 3380688.3380693

*Corresponding author

and multi-head mutual attention features are concatenated as complete multi-modal sentiment features, which is then sent to the classifier for multi-modal sentiment analysis. Experiments are conducted on two public multimodal datasets of Multimodal Opinion Utterances Dataset (MOUD) corpus and CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus. The results show that, compared to the state-of-the-art models, the proposed multi-head attention framework attains better performance for various combinations of input modalities (text, visual and audio).

The paper is organized as follows: Section 2 reviews the literature of multimodal sentiment analysis briefly; Section 3 describes the proposed method in detail; experimental results and analysis are shown in Section 4; finally, Section 5 concludes the paper.

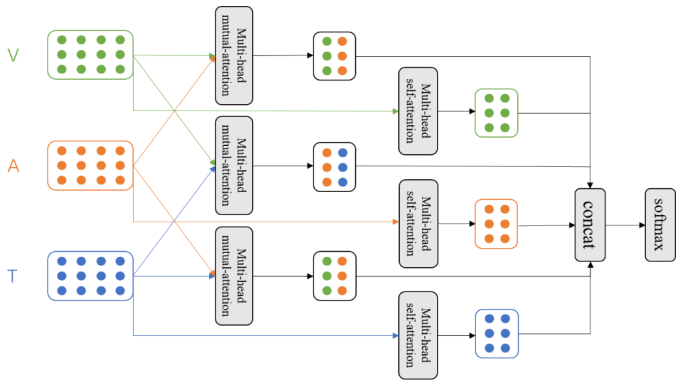

Figure 1. Multi-Modal Multi-Head Attention (MMHA) model

2. RELATED WORKS

The rise of multimedia platforms makes it easier to obtain user opinions, but it is increasingly unrealistic to manually distinguish user sentiment from huge video data. Therefore, whether for research value or commercial value, multimodal sentiment analysis has received widespread concern in both academia and industry.

The visual-based sentiment analysis is mainly to analyze the sentiment transmitted by the facial expressions of each frame in the video. In [16], the research on facial expressions shows that by analyzing universal changes of facial expressions, effective clues for distinguishing sentiment can be obtained, so the traditional method is to extract the action unit of each frame. In [5], they automatically extract 40 facial expression units for sentiment analysis using a computer expression recognition tool (CERT). Recent approaches to visual sentiment analysis use deep learning methods to automatically extract visual sentiment features. In [6] [17], they apply 3D-CNN to extract visual sentiment features from video, these features include the representation features of each frame, and also contain expression changes on the timeline.

In most of the audio sentiment research, the audio features extracted by the researchers include global features such as pitch, energy, RMS energy, spectral centroid, and tonal centroid characteristics. In [19], the research on sentiment analysis found that pitch and energy related features play a vital role in audio sentiment analysis. As these global features are relatively few and can be quickly calculated, they get widely used. In [5] [17], they separately use the OpenEAR and OpenSmile to automatically extract the global features. However, global features still have limitations. The global features are better at distinguishing the higher arousal sentiment, but they do not distinguish the sentiment with lower arousal [18]. In addition, since each utterance needs to be divided into several segments before the global features are extracted, there is no temporal correlation between these segments. As a result, recent research on audio sentiment analysis uses the combination of spectrograms and deep learning methods. In [20], they use Bi-LSTM, CNN and attention mechanism extract features from raw waveform and spectrograms, then concatenate the features as the whole audio sentiment features.

In recent years, the extraction of text sentiment features mainly uses the word vector model, mainly using pre-trained word vector models such as word2vector, Glove to extract word vectors. Then the word vector is composed into a sentence vector and send to the classifier to classify [17] [21-22]. The word vector model has a good effect on the sentiment analysis of short sentences and unambiguous sentences.

Multimodal sentiment analysis, which is a promising area of research [23]. In recent years, there have been more and more researches on utterance-level multimodal sentiment analysis for vision, audio and text. Obtaining the correlation between different modalities from multimodality and combining the characteristics of each modality can greatly improve the accuracy of sentiment analysis. In [21], they used a method called the Multi-Note Block (MAB) to extract the interrelationships between the different modalities and store them in a long-short-term repetitive component of the Memory Mix Memory (LSTHM). In [22], they used a hierarchical network structure that integrates the two modalities first and then the three modalities. In [17], they used a combination of LSTM and attention mechanisms to extract the sentiment relationship between contextual utterances in a video, and show the state-of-the-art models on CMU-MOSI. Although this paper and [17] both use similar multimodal attention structures in multimodal sentiment analysis, there are two differences (1) this paper uses different methods to extract unimodal sentiment features. (2) the multi-modal attention mechanism used in this paper can capture related information on different subspaces by counting multiple times.

3. PROPOSED METHOD

In the proposed framework, the multi-modal information is applied to predict the sentiment of utterance. Firstly, different methods are employed to extract the visual, audio and text features. Then multimodal multi-head attention is applied on the unimodal features. The objective is (1) learn the dependence of unimodal features, capture the internal structure of unimodal features, and (2) learn the relationship between the multiple modalities, put more attention on the contributing features. In particular, this paper uses the unimodal multi-head self-attention network and the bimodal multi-head mutual-attention network, respectively, in which the attention function is applied to unimodal visual, audio, text representation, and the representations of the combination of modalities (visual-text, text-audio and audio-visual). Finally, the outputs of unimodal representations along with the bimodal representations are concatenated and passed to the softmax layer for classification. The proposed method is called the Multi-Modal Multi-Head Attention (MMHA) framework. The overall architecture of the MMHA framework is shown in Figure 1. The multi-head self-

attention mechanism network and mutual-attention mechanism network are illustrated in Figure 2 and Figure 3 (take unimodal visual and bimodal visual-audio as examples), respectively.

3.1 Extracting Unimodal Features

Similar to [6], the sentiment features of visual, audio and text are extracted from each utterance separately. The methods of feature extraction are as follows.

3.1.1 Visual Features Extraction Based on VGG16-LSTM

This paper uses the pre-trained VGG16 [24] model on the ImageNet dataset and Long-short Term Memory network (LSTM) [25] combination to extract visual sentiment features. In the past, the combination of the VGG16 network and LSTM has been successfully applied to human action recognition and emotion recognition [10-11] [26], and achieve the state-of-the-art result. This motivates us to believe this network can be used to extract visual features. This combination helps us extract either the visual presentation features of each frame and the time information that can dynamically reflect the changes between the adjacent video frames.

Let $vid \in R^{n\times f\times h\times w}$ represent a video, where n = number of channels in a frame (as this paper only considers RGB images, the $n=3$ in the model), f = number of frames in a video, h = height of the frames and w =width of the frames (consistent with the input size of VGG16 network, the height and width of the frames are both 224). Besides, let c = the hidden size of LSTM.

Since the number of frames in per video is different, the fixed number of frames is f, the frames that exceed f are truncated, and the video with less than f frames performs the zero-padding operation. The pre-trained VGG16 network is followed by the global average pooling layer to get the visual presentation features, and then send the features to the LSTM. Subsequently, a dense layer of size 1024 and sigmoid are followed. The activation values of the dense layer are finally used as the visual features for each utterance. In the work, $f=100$ and $c=256$ gives the best results.

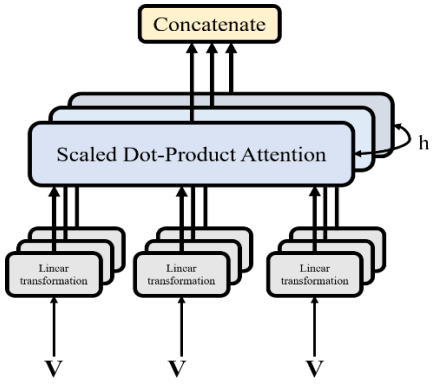

Figure 2. Multi-head self-attention mechanism network (take visual features as an example)

3.1.2 Audio Features Extraction Based on Spectrograms-CNN

In the paper, spectrograms and convolutional neural network are used to extract audio sentiment features. The specgram function in the python drawing library matplotlib is used to automatically extract the spectrogram, where the frame length is set to 25ms and the sampling frequency is set to 1024Hz. Furthermore, the size of the saved spectrograms is 256×256, and then send the spectrograms to the designed convolutional neural network.

The convolutional neural network designed in this paper consists of two convolution layers. The first layer has a kernel of size 10×8, with 6 kernels and the second layer has a kernel of size 8×6 with 12 kernels. The convolution layers are interleaved with max-pooling layers of the window, separately. Then this is followed by a dense layer with a size of 300 and sigmoid layer. The rectified linear unit (ReLU) is used as the activation function and the output of the activation values of the dense layer is the audio sentiment features for each utterance.

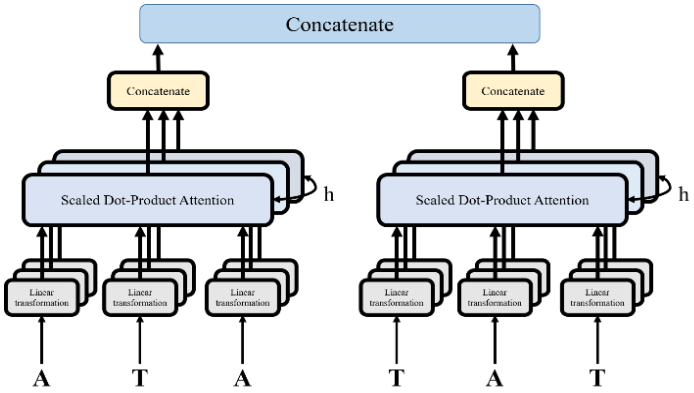

Figure 3. Multi-head mutual-attention mechanism network (take visual and audio features as an example)

3.1.3 Text Features Extraction Based on Bert-As-Service

As discussed in section 2, the word vector model has a good effect on the sentiment analysis of short sentences and unambiguous sentences. However, the sentences processed in reality are not that simple. It needs to fully consider the relationship between the preceding words and the following words in the utterance and to eliminate the problem of polysemy. The BERT [12] model is a new language characterization model proposed by Google. It uses Transformer's two-way encoder representation. The difference from other language models is that the model pre-trains deep two-way representation by jointly adjusting the context in all layers.

Based on the BERT-Base, Uncased pre-training model released by Google Research, this paper uses the Bert-as-service[1] toolkit to extract sentence-level vectors. It gets 768 features for each utterance.

3.2 Multi-Modal Multi-Head Attention (MMHA) framework

This paper separately denotes the extracted visual, audio, text features for each utterance in Section 3.2 as $V\in R^{1024}$, $A\in R^{300}$, $T\in R^{768}$. In order to send the features into the multi-head network, the size of visual, audio, and text features should be reshaped as $V\in R^{16\times64}$, $A\in R^{10\times30}$, $T\in R^{8\times96}$.

[1] https://github.com/hanxiao/bert-as-service

Let the corresponding multiple linear transformation weights as $W_V^h \in R^{64\times d}$, $W_A^h \in R^{30\times d}$, $W_T^h \in R^{96\times d}$, where h represents the serial number of the linear transformation, d is the output size of each linear transformation. Then the features through the linear transformation become $V \in R^{16\times d}$, $A \in R^{10\times d}$, $T \in R^{8\times d}$.

Multi-Head Attention Mechanism. As shown in Figure 2 and 3, the multi-head network is to learn different features from different subspaces using multiple linear transformations. The attention mechanism used in this paper is the same as [15], which is called scaled dot-product attention mechanism. Next, the principle is explained that this attention mechanism can be used on the multimodal sentiment analysis.

Multi-Head Self-Attention Mechanism. In particular, the scaled dot-product self-attention of V is computed as follows,

$$O_V^h = soft\max(\frac{VV^T}{\sqrt{d}})V, h = 1, ..., m \quad (1)$$

When the number of the linear transformation set as m , the multi-head self-attention features can be computed by concatenating the O_V^h as

$$O_V = Concat(O_V^1, ...O_V^m) \quad (2)$$

Multi-Head Mutual-Attention Mechanism. The scaled dot-product mutual-attention between V and A are computed as follows,

$$\begin{aligned} O_{VA}^h &= soft\max(\frac{VA^T}{\sqrt{d}})V, h = 1, ..., m \\ O_{AV}^h &= soft\max(\frac{AV^T}{\sqrt{d}})A, h = 1, ..., m \\ O_{A\_V}^h &= Concat(O_{VA}^h, O_{AV}^h) \end{aligned} \quad (3)$$

the multi-head mutual-attention features can be computed by concatenating the $O_{A\_V}^h$ as

$$O_{A\_V} = Concat(O_{A\_V}^1, ...O_{A\_V}^m) \quad (4)$$

Concatenation. The overall feature computed by concatenating all multi-head self-attention features and mutual-features as

$$O = Concat(O_V, O_A, O_T, O_{V\_A}, O_{V\_T}, O_{A\_T}) \quad (5)$$

This concatenated features are then used for final classification.

4. EXPERIMENTS

In this section, the datasets used for the experiments are introduced first, and then give the experimental results along with the necessary analysis.

4.1 Datasets

The proposed model is evaluated on Multimodal Opinion Utterances Dataset (MOUD) corpus [5] and CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) corpus [27] for multimodal sentiment analysis.

The MOUD dataset contains evaluation videos from different products for a total of 55 people. Since these reviews are all in Spanish, and the text pre-trained model used in this paper is based on English corpus, this paper employs the same strategy as [5] to call the Google Translate Interface[2] to obtain English text data. This dataset contains three labels: positive, negative and neutral, but this paper only considers the positive and negative data. There are 79 videos in the dataset, of which 59 videos are in the training and validation sets.

The CMU-MOSI dataset contains a total of 93 review videos for different topics. Each utterance in this dataset corresponds to an integer label between -3 and 3, indicating the emotional intensity. As only two classes (positive and negative) are considered in this paper, the data with the label greater than or equal to 0 are marked as positive, and the label less than 0 is negative. The training and validation set contains 1447 utterances out of 62 videos, and the test set contains 752 utterances out of the remaining 31 videos. Table 1 shows the details of train/test split in MOUD and CMU-MOSI dataset. What's more, the CMU-MOSI and MOUD dataset are used as the training and test sets separately to evaluate the robustness of our model.

Table 1. Train/Test split details of each dataset. Legend: X→Y represents train: X and test: Y; Validation set is split from the shuffled training sets with 80/20% training/validation ratio

Dataset	Train		Test	
	utterance	video	utterance	video
MOUD	322	59	115	20
MOSI	1447	62	752	31
MOSI→MOUD	2199	93	437	79

4.2 Performance of the Model

The proposed approach is evaluated on MOUD & CMU-MOSI. The score of the accuracy is used as the evaluation metric.

In the experiment, the number of the linear transformation matrix set in multi-head attention mechanism is 8 (that is, 8 heads), the output size of each multi-head attention is 16. ReLU activation function is employed in each multi-head attention layer and softmax activation in the final classification layer. In addition, the multi-head attention layer is followed by the global average pooling layer and dropout layer, this paper sets the dropout=0.3 (MOUD) & dropout=0.45(MOSI). For training the model, the batch size is 16 and the Adam optimizer is used along with cross-entropy loss function to train 80 epochs. The average accuracy scores of five runs is recorded for the experiment.

This model is used to evaluate all valid combinations between modalities. Including unimodal sentiment analysis (apply each multi-head self-attention features to the classification layer), bimodal sentiment analysis (concatenate the bimodal multi-head self-attention features and mutual-attention features as bimodal sentiment features and send to the classification layer), tri-modal sentiment analysis (apply all the concatenated multi-head attention features to the classification layer). Table 2 shows the results of the sentiment analysis.

[2] http://translate.google.com

For MOUD dataset, compared to the state-of-the-art, the unimodal accuracy score has increased by 3.71% to 11.45% and obtains better performance in visual. At the same time, the accuracy scores of bimodal have increased by 0.23% to 5.01% and the tri-modal by 1.83%. The combination of visual and audio modality accuracy score performs better in bimodal sentiment analysis. For MOSI dataset, the visual and audio modalities increased by 8.42% and 3.72%. However, the accuracy score of text modality is 1.99% lower than the state-of-the-art. We surveyed the method of the state-of-the-art and found that they leverage the multi-modal and contextual information for predicting the sentiment of an utterance. This idea has greatly improved the accuracy score of text sentiment analysis, while the improvement of visual and audio modality is limited. Due to the limitations of the text modality, our bimodal and tri-modal accuracy scores are not improved well. In particular, the combination of audio and text accuracy score is lower than the state-of-the-art. The accuracy score of the tri-modal only improved by 0.4%.

Table 2. Accuracy (%) on text (T), visual (V), audio (A) modality and comparison with the state of the art

Modality	MOUD		MOSI	
	Our Method	feature-level fusion [13]	Our Method	MMMU-BA [17]
V	87.83	76.38	72.12	63.70
A	82.61	74.22	65.82	62.10
T	83.48	79.77	78.19	80.18
V+A	88.70	83.69	75.13	65.16
V+T	89.57	85.46	82.05	81.51
A+T	84.35	84.12	80.25	80.58
V+A+T	90.43	88.60	82.71	82.31

In order to examine the generalization of our model, the model is trained with the complete MOSI dataset and test on MOUD dataset (Table 3). Compared to the state-of-the-art, the performance of our model is improved by 25.61%.

Table 3. Train the model with the complete MOSI dataset and test on MOUD dataset (the classification accuracy (%) on MOUD dataset and comparison with the state of the art)

Modality	Our Method	bc-LSTM [6]
V	61.60	49.60
A	61.64	47.20
T	76.48	46.90
V+A	63.17	49.60
V+T	77.17	49.80
A+T	77.63	51.30
V+A+T	78.31	52.70

5. CONCLUSIONS

For multimodal sentiment analysis, reasonable unimodal features along with a robust model are essential. This paper uses the relatively novel methods at this stage to extract unimodal sentiment features. Meanwhile, a multi-head attention based network is developed to predict the sentiment of each utterance. The proposed method has outperformed the state of the art in MOUD and MOSI datasets.

As for future work, we plan to extract contextual features from the utterances of a video for multimodal sentiment analysis. In addition, an algorithm will be proposed for calculating the correlation between different modalities to improve the robustness of the multimodal sentiment analysis model.

6. ACKNOWLEDGMENTS

This work was partly supported by the Key Research and Development Program of Jiangsu Province (Grant No. BE2016775), the National Natural Science Foundation of China (NSFC) under Grants 61971236, the Project funded by China Postdoctoral Science Foundation under Grant 2018M632348, the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX19_0954).

7. REFERENCES

[1] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115-124). Association for Computational Linguistics.

[2] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

[3] Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., & Morency, L. P. (2017, November). Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 163-171). ACM.

[4] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

[5] Pérez-Rosas, V., Mihalcea, R., & Morency, L. P. (2013, August). Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 973-982).

[6] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017, July). 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)* (pp. 873-883).

[7] Barnes, J., Klinger, R., & Walde, S. S. I. (2017). Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. *arXiv preprint arXiv:1709.04219*.

[8] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

[9] Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A Survey of Sentiment Analysis Based on Transfer Learning. *IEEE Access*, *7*, 85401-85412..

[10] Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2017, March). Two stream lstm: A deep fusion framework for human action recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 177-186). IEEE.

[11] Fan, Y., Lu, X., Li, D., & Liu, Y. (2016, October). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 445-450). ACM.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[13] Poria, S., Cambria, E., & Gelbukh, A. (2015, September). 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* (pp. 2539-2544).

[14] Metallinou, A., Lee, S., & Narayanan, S. (2008, December). Audio-visual emotion recognition using gaussian mixture models for face and voice. In *2008 Tenth IEEE International Symposium on Multimedia* (pp. 250-257). IEEE.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*(pp. 5998-6008).

[16] Ekman, P., & Keltner, D. (1997). Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27-46.

[17] Ghosal, D., Akhtar, M. S., Chauhan, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2018). Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3454-3466).

[18] Luo, Z., Xu, H., & Chen, F. (2019). Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network. In *AffCon@ AAAI* (pp. 80-87).

[19] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98-125.

[20] Chen, F., & Luo, Z. (2019). Sentiment Analysis using Deep Robust Complementary Fusion of Multi-Features and Multi-Modalities. *CoRR*.

[21] Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P. (2018, April). Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[22] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. (2018). Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, *161*, 124-133..

[23] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443.

[24] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[25] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[26] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).

[27] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, *31*(6), 82-88.