MF-BERT: Multimodal Fusion in Pre-Trained BERT for Sentiment Analysis

Jiaxuan He and Haifeng Hu, *Member, IEEE*

***Abstract*—Multimodal sentiment analysis mainly concentrates on language, acoustic and visual information. Previous work based on BERT utilizes only text (language) representation to fine-tune BERT, while ignoring the importance of nonverbal information. Due to the fact that features extracted from a single modality may contain uncertainty, it is challenging for BERT to perform well in real-world applications. In this paper, we propose a multimodal fusion BERT that can explore the time-dependent interactions among different modalities. Additionally, prior BERT-based methods tend to train the models with only one optimizer to update the parameters. However, we argue that BERT has been pre-trained with a lot of corpora so it needs to be fine-tuned slightly. Therefore, an internal updating mechanism is introduced to avoid the overfitting of the model in the training process. We set two optimizers for multimodal fusion BERT and other components of the model with different learning rates, which enables the model to attain optimal parameters. The results of experiments on public datasets demonstrate that our model is superior to the baselines and achieves the state-of-the-art.**

***Index Terms*—Internal updating, multimodal sentiment analysis, multimodal fusion BERT.**

I. Introduction

MULTIMODAL sentiment analysis has earned a lot of popularity with the development of communication technology. The audience may be influenced by the sentiment expressed by actors on the videos. It is easy to find multimodal data from daily life on the Internet nowadays. In the past, to fully understand the sentiment of people with natural language processing (NLP) strategy, researchers used to concentrate on text (language) modality which was essential for people to share their feeling [1]–[4].

However, human beings prefer to express their feelings with not only words but also nonverbal (acoustic or visual) behavior. For instance, the sentence “He always goes there” lacks accurate sentiment expression. Nowadays, multimodal learning research plays an important role in artificial intelligence [5]. For example, [6] develops a novel network which transfers labeling information across heterogeneous domains. Therefore, multimodal learning method is often regarded as a better strategy to infer sentiment. Previous multimodal learning systems [7]–[11] mainly studied the inter-modality and intra-modality information for multimodal feature fusion. Both Recurrent neural network (RNN) methods [12], [13] and Convolutional neural network (CNN) [14] methods were utilized to study the context information for sentiment inference. However, RNNs cannot process the sequences in parallel and CNNs cannot process time sequences with extended length. Nowadays, transformer-based strategies [15]–[17] play an important role in sentiment analysis. Furthermore, the pre-trained model has become popular in the NLP field. Though BERT [18] can achieve superior results on some tasks, it can just learn information by generating contextual word representations. For sentiment analysis, it is better to study more than one single modality to infer sentiment. Therefore, BERT-based methods may affect the performance of the sentiment inference. How to apply BERT mechanism to multimodal learning is promising.

To address the limitation of BERT and introduce the nonverbal information to language embedding, we propose a multimodal fusion BERT (MF-BERT) for sentiment analysis. Different from the BERT which only utilizes language information to fine-tune its parameters, the multimodal fusion (MF) module fuses language embedding with nonverbal embedding before sending them to encoder layers of BERT. To avoid transferring noise to the latent fusion, similarity loss function [19] is applied to compute the distribution difference between language embedding and nonverbal embedding. Nonverbal embedding with the bigger difference is selected to combine with the language embedding in MF module. Therefore, fusion embedding that contains various modality information is obtained for the encoder layer of BERT. Furthermore, to strengthen the essential information of nonverbal modality in our framework, we apply MF module again to the output embedding and nonverbal embedding shown in Fig. 1. The acoustic embedding and visual embedding processed by 1d-convolution are sent to MF module and fuse with the MF-BERT embedding in O_2 of Fig. 1.

To avoid the overfitting of BERT and attain the optimal parameters of other components of the model during training, an internal updating strategy is proposed in our framework. We divide MF-BERT and other components as two parts of the proposed framework. Each part of the model is accompanied by an optimizer, so they are fine-tuned with different learning rates. In our model, a lower learning rate is utilized in MF-BERT fine-tuning process, while a higher one is applied to train additional components. In this way, the entire model is fine-tuned properly and performs better in the testing process.

Manuscript received October 1, 2021; accepted December 22, 2021. Date of publication December 31, 2021; date of current version January 28, 2022. This work was supported by the National Natural Science Foundation of China under Grant 62076262. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniele Giacobello. *(Corresponding author: Haifeng Hu.)*

The authors are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China (e-mail: hejx25@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/LSP.2021.3139856

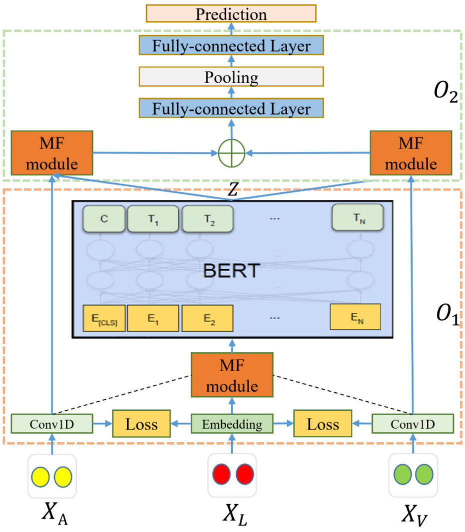

Fig. 1. The overall architecture of the model. O_1 denotes the MF-BERT with an optimizer and O_2 denotes other components of the model with another optimizer. MF module denotes the multimodal fusion module shown in Fig. 2.

In general, the contributions of this paper are as follows: 1) proposing multimodal fusion BERT to analyze sentiment from more than just language modality; 2) proposing a late fusion strategy and an internal updating strategy for model fine-tuning; 3) conducting extensive experiments to verify the superiority of our proposed multimodal learning system.

II. Proposed Method

As shown in Fig. 1, our proposed framework contains two parts, including multimodal fusion BERT (MF-BERT), late fusion and internal updating. The detailed structure of the model will be introduced in the following subsections:

A. Multimodal Fusion BERT (MF-BERT)

Language information is often considered to be the direct information to express the feelings of people in their daily life. However, models trained with only language modality may perform poorer than those trained with multimodal information. The reason is that unimodal information may sometimes have trouble expressing the sentiment clearly with ambiguous words. In addition, since BERT [18] can only process the text (language) embedding, it may have difficulty extracting useful information from language modality. To enable BERT to fully understand the interaction of multimodal information, we propose a multimodal fusion (MF) subnetwork which fuses nonverbal embedding with the language embedding with a similar loss function [19]. With such fusion operation, BERT is able to learn multimodal information and thereby provides a more accurate understanding of sentiment.

As shown in Fig. 1, the input of the MF-BERT consists of three parts: language, acoustic and visual modality, respectively. Firstly word-piece tokens of language modality $X_L = [L_1, L_2, \ldots L_n]$ are sent to the embedding layer of the BERT to attain the language embedding $X_T = [CLS, L_1, L_2, \ldots L_n]$, where $[CLS]$ denotes the special classification embedding. Meanwhile, acoustic and visual raw embeddings extracted from the dataset are concatenated with padding 0 to match $[CLS]$ embedding. Then the processed nonverbal embeddings after the convolution operation are sent to the MF module in MF-BERT for the latent fusion. Here the acoustic and visual raw embeddings are defined as X_A and X_V separately.

To enable language embedding to combine with nonverbal embedding properly in the embedding creating process, X_T, X_A and X_V are combined to obtain fusion embedding as the input for the first encoder layer of BERT. Given the unimodal embedding $X_m \in \mathbb{R}^{T \times d_m}$, where T is the time sequence length (which is set to 50), $m \in \{A,\ V\}$, 1D temporal convolution (Conv1D) is applied to reshape the feature dimension d_m to d:

$$\hat{X}_m = \text{Conv 1D}\left(X_m, K_m\right) \in \mathbb{R}^{T \times d},\ m \in \{A,\ V\} \tag{1}$$

where K_m denotes the kernel size of the convolution, $\hat{X}_m$ is the processed embedding for modality m. A and V denote the acoustic and visual modality respectively.

Since feature distribution of various modality embeddings may be similar, the embedding fusion of language and nonverbal embeddings is possibly redundant. Moreover, different modalities often contain the same sentiment of the speaker, which means that fusing of all modalities in each training process is not of necessity. Therefore, we apply the similarity loss function [19] to calculate the correlation of different modalities in our framework. With the similarity loss function, language embedding X_T is compared with nonverbal embedding $\hat{X}_m$ to attain the distribution difference. Firstly, given the processed language embedding and nonverbal embedding, we apply the mean operation to the feature dimension of the language embedding and attain the matrix $X_{TT} \in \mathbb{R}^{b \times T \times 1}$, where b denotes the batch size. Secondly, the nonverbal matrix $\hat{X}_{mm} \in \mathbb{R}^{b \times T \times 1}$ is obtained from the last dimension of the embedding. Finally, the language matrix is compared with acoustic and visual matrices respectively by using the following equation:

$$L_{CC} = \frac{1}{n^2} \sum \left(\varphi\left(\boldsymbol{X}_{TT}, \boldsymbol{X}_{TT}^{t}\right) - \varphi\left(\hat{\boldsymbol{X}}_{mm}, \hat{\boldsymbol{X}}_{mm}^{t}\right) \right)^2 \tag{2}$$

where L_{CC} denotes the difference between language embedding and nonverbal embedding, φ denotes the matrix multiplication operation, t denotes the transpose of the matrix and n denotes the number of the vectors in the matrix. With (2), nonverbal embedding that contains more differences is selected for the next embedding fusion stage.

To enable BERT to learn multimodal information, we propose a multimodal fusion mechanism to fuse the language embedding and nonverbal embedding in BERT. As shown in Fig. 2, the multimodal fusion module utilizes the information of nonverbal embedding to influence the word (language) embedding by utilizing the following equation:

$$H_m = ReLU\left(W_{TM}\left[X_T; X_m\right] + b_{TM}\right),\ m \in \{A,\ V\} \tag{3}$$

where $H_m \in \mathbb{R}^{T \times n}$ denotes the fusion embedding of two modality embeddings, $ReLU$ denotes ReLU activation function, W_{TM} denotes the weight matrix, $[;]$ denotes the concatenation of the sequences on feature dimension and b_{TM} denotes the scalar bias vector. By utilizing (3), language embedding and nonverbal embedding are combined, hence the correlation of the embeddings is strengthened. Then H_m is multiplied with

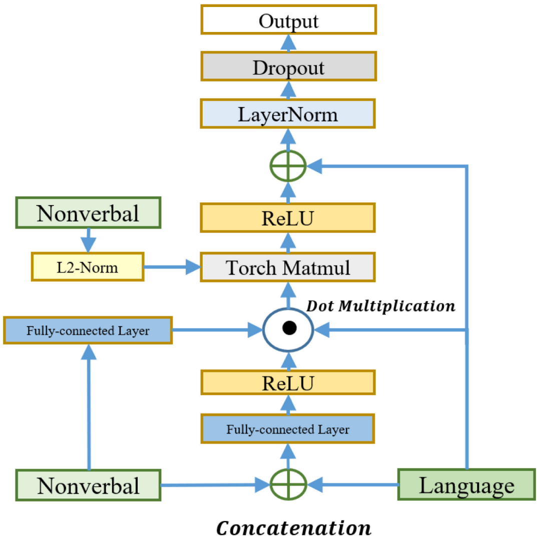

Fig. 2. The multimodal fusion of the model. Nonverbal denotes nonverbal embedding and Language denotes language embedding.

the language embedding and nonverbal embedding:

$$\hat{X}_T = H_m * X_T * FC(\hat{X_m}),\ m \in \{A,\ V\} \tag{4}$$

where $*$ denotes the dot multiplication operation, and FC denotes the fully-connected layer. Since $\hat{X}_T$ may leak some important messages of the embeddings, we apply matrix multiplication to $\hat{X}_T$ and $\hat{X}_m$, which enhances the diversity of the modality information. However, if $\hat{X}_T$ and $\hat{X}_m$ are calculated directly, the large differences of the elements in the matrices will reduce computational efficiency. Therefore, the l-2 norm is applied to process the nonverbal embedding. The embeddings are processed by the following equations:

$$\hat{X} = ReLU\left(\hat{X}_T \otimes \frac{\hat{X}_m}{\|\hat{X}_m\|_2}\right),\ m \in \{A,\ V\} \tag{5}$$

where $\otimes$ denotes the multiplication operation. Then the outputs are sent to the dropout layer (Drop) and layer-norm layer (LN) to obtain the final fusion embedding for the first encoder layer of BERT:

$$X = Drop(LN(\hat{X} + X_T)) \tag{6}$$

the X denotes the fusion embedding which will be sent to the first encoder layer of BERT.

B. Late Fusion and Internal Updating

1) Late Fusion: Different from previous BERT-based methods, the BERT utilized in our model is inserted with multimodal fusion module to learn interactions among modalities. Though the output of the MF-BERT includes the multimodal information fusion of three kinds of modalities, it may still lose some essential unimodal information. Meanwhile, the attention mechanism of the transformer may also limit the fusion ability of MF-BERT during training. To this end, we apply MF module to the MF-BERT output embedding Z and nonverbal embedding again, which helps Z to supplement unimodal information. As shown in Fig. 1, nonverbal embeddings processed by 1d-convolution operation are combined with the MF-BERT output embedding Z:

$$\hat{Z}_m = MF(Z, X_m),\ m \in \{A,\ V\} \tag{7}$$

where MF is the multimodal fusion module and $\hat{Z}_m$ is the fusion embedding of nonverbal and language modalities.

To obtain the final sentiment prediction, we first concatenate $\hat{Z}_A$ and $\hat{Z}_V$ on feature dimension:

$$\hat{Z} = \hat{Z}_A \oplus \hat{Z}_V \in \mathbb{R}^{T \times 2d} \tag{8}$$

to fuse the nonverbal features of acoustic and visual modalities, we apply a fully-connected layer to $\hat{Z}$:

$$\hat{Z}_P = w \cdot \hat{Z} + b \tag{9}$$

where w denotes the weight matrix and b denotes the bias vector. With (9), different sequences are further fused to generate sequence combinations.

Both pool layer and fully-connected layer are used on $\hat{Z}_P$ to attain the sentiment prediction Z_P:

$$Z_P = FC(Pooler(\hat{Z}_P)) \tag{10}$$

where $Pooler$ denotes the pooling operation.

2) Internal Updating: Since BERT has been pre-trained with a large number of corpora, its performance may become worse if it is fine-tuned with a high learning rate. To prevent this, we divide our proposed framework into two main parts: the former is the MF-BERT which mainly contains multimodal fusion module and BERT; the latter contains other components of the proposed framework. Each part is attached with an AdamW optimizer with an independent learning rate. O_1, O_2 shown in Fig. 1 denotes the MF-BERT and other components of the framework, respectively. MF-BERT is fine-tuned with a lower learning rate (3e-5 for both MOSI and MOSEI) while others are trained with a higher one (1e-4 for MOSI and 5e-5 for MOSEI). To obtain the optimal parameters of the model, the MSE loss function is applied to calculate the difference between the prediction Z_P and the ground truth label.

III. Results and Analysis

1) Datasets: CMU-MOSEI (MOSEI) [8] and CMU-MOSI (MOSI) [20] are utilized in our experiments. MOSEI contains both sentiment and emotion labels with three kinds of modality information. MOSI is a multimodal sentiment analysis dataset that consists of 2,199 short video clips.

2) Evaluation Metric: To evaluate the performance of the model and compare our method with other baselines, the following evaluation metrics are used in our experiments: (1) Acc2: binary score; (2) F1-score; (3) MAE: mean absolute error; (4) Corr: the correlation of the prediction.

3) Baselines: The proposed model is compared with following baselines: (1) RNN-based methods: Early Fusion LSTM (EF-LSTM), Late Fusion LSTM (LF-LSTM) and MFN [21]; (2) Tensor fusion-based methods: TFN [22] and LMF [23]; (3) transformer-based methods: MULT [16], IMR [17], MISA [24], ICCN [25] and MAG-BERT [26]; (4) Others: RAVEN [27] and QMF [28]. For a fair comparison, each baseline is fine-tuned by conducting a fifty-times random grid search to obtain the optimal

TABLE I
THE COMPARISON WITH BASELINES ON MOSI

	Acc2	F1	MAE	Corr
EF-LSTM	75.8	75.6	1.053	0.613
LF-LSTM	76.4	75.4	1.037	0.620
TFN [22]	76.4	76.3	1.017	0.604
LMF [23]	73.8	73.7	1.026	0.602
MFN [21]	78.0	76.0	1.010	0.635
RAVEN [27]	78.8	76.9	0.968	0.667
MULT [16]	79.3	78.3	1.009	0.667
QMF [28]	79.7	79.6	0.915	0.696
ICCN [25]	83.1	83.0	0.862	0.714
MISA [24]	83.4	83.6	0.783	0.761
MAG-BERT [26]	83.5	83.5	0.769	0.790
Without Early Fusion	84.6	84.5	0.725	**0.795**
Without Late Fusion	85.0	85.0	0.749	0.790
Only Acoustic	84.3	84.3	0.740	0.788
Only Visual	84.1	84.1	0.739	0.789
With Same Rate	83.5	83.6	0.773	0.785
Ours	**85.6**	**85.5**	**0.721**	0.793

TABLE II
THE COMPARISON WITH BASELINES ON MOSEI

	Acc2	F1	MAE	Corr
EF-LSTM	79.1	78.8	0.665	0.621
LF-LSTM	79.4	80.0	0.625	0.655
TFN [22]	79.4	79.7	0.610	0.671
LMF [23]	80.6	81.0	0.608	0.677
MFN [21]	79.6	80.6	0.618	0.670
RAVEN [27]	79.0	79.4	0.605	0.680
MULT [16]	80.2	80.5	0.638	0.659
IMR [17]	80.6	81.0	-	-
QMF [28]	80.7	79.8	0.640	0.658
ICCN [25]	84.2	84.2	0.565	0.782
MISA [24]	85.5	85.3	**0.555**	0.756
MAG-BERT [26]	85.0	85.0	0.778	0.602
Without Early Fusion	85.2	85.2	0.581	0.788
Without Late Fusion	84.3	84.3	0.594	0.785
Only Acoustic	84.6	84.7	0.594	0.788
Only Visual	84.9	85.0	0.594	0.786
With Same Rate	84.2	84.4	0.598	0.788
Ours	**85.6**	**85.5**	0.588	**0.788**

hyper-parameters. Then, we train each baseline again with the optimal hyper-parameters to achieve the final result. QMF [29], ICCN [25] and MISA [24] do not provide the open-source codes so we present the results of their original papers.

A. Comparison With Baselines

The results on Tables I and II demonstrate that our proposed method outperforms all the baselines on the majority of the evaluation metrics both on MOSI and MOSEI datasets. For RNN-based methods, EF-LSTM attains the worst performance, which implies that simple concatenation on feature dimension is weak in information fusion. Instead, MF-BERT utilizes MF module to fuse different modalities properly and attain useful information for sentiment inference.

For tensor fusion-based models, they achieve various performance on the same dataset. LMF performs better than TFN, which is because LMF utilizes the low-rank strategy to improve the efficiency of the model and reduce the complexity in high dimensional weights of the generated feature information.

For transformer-based baselines, our method outperforms all the baselines on most evaluation metrics. MULT and IMR utilize transformers to learn time sequences and perform fairly on MOSEI. Though transformer-based methods are good at learning time sequences, their performance is dependent on the input extracted from the training dataset. BERT-based methods ICCN, MISA and MAG-BERT perform better than those that only utilize transformer mechanism in their frameworks, for which the BERT has been pre-trained by a large number of corpora. However, the weakness that BERT only accepts language embedding as the input reduces the performance of ICCN and MISA. MAG-BERT combines nonverbal modality embeddings with the language embedding and enables the BERT to be fine-tuned with multimodal information. Unlike MAG-BERT that may introduce too much redundant information, MF-BERT utilized information of two kinds of modalities in the embedding preparing process with the help of a similarity loss function. Under such circumstances, the MF module helps BERT to reduce the unnecessary redundant information and combine various modality embeddings properly, which improves the performance of our model.

To highlight the advantages of our method, we extract a video clip from MOSI for a fair comparison. MF-BERT and MAG-BERT are utilized to infer sentiment from the video. Part of the transcript of the video is "I DONT KNOW IF FANS OF THE COMIC WOULD BE AS INCLINED TO SAY THE SAME BECAUSE I DONT KNOW HOW MUCH IT DIFFERS FROM THE COMIC OR RADIO PROGRAM THAT PEOPLE LIKE SO MUCH" and its corresponding label is closer to "-2.4". Our model abandoned the visual information during testing and got the right prediction. However, MAG-BERT applied multimodal adaption gate to fuse three kinds of modalities and its prediction is about "-3.1," which is different from the ground truth label of the video clip.

B. Ablation Results

Ablation experiments are conducted on our method, and the results are shown on Tables I and II. We study the influence of multimodal fusion module, similarity loss function (L_{CC}) and parameter updating strategy respectively. For the multimodal fusion module, we take it away from BERT and use only language modality in BERT fine-tuning process ("Without Early fusion"), and we also remove the late fusion operation("Without Late Fusion"). Compare to the original model, a significant drop in performance indicates that the multimodal fusion module can combine the multimodal information more efficiently both inside and outside BERT.

For L_{CC}, we remove the modal embedding comparison and utilize one nonverbal embedding to fuse language embedding directly ("Only Acoustic" and "Only Visual"). The results show that without the embedding selection, a direct fusion of different embeddings will introduce much noise to the model during training, which will lead to worse performance.

For the parameter updating strategy, we utilize one optimizer with the same learning rate to fine-tune the overall model ("With Same Rate"). The result implies that the parameters of the proposed framework are not optimal for testing. Therefore, it is necessary to set different optimizers with different learning rates for the proposed framework.

IV. CONCLUSION

We propose a multimodal fusion BERT with an internal updating strategy for multimodal sentiment analysis. BERT with multimodal fusion module enables language information to combine with nonverbal information, thus enhancing the interaction of the multimodal information. The late fusion and internal updating strategies help fine-tune MF-BERT more effectively and strengthen the importance of nonverbal information. Extensive experiments on public datasets demonstrate that our method achieves the state-of-the-art performance.

REFERENCES

[1] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, “Autoencoder-based unsupervised domain adaptation for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[2] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, “Universum autoencoder-based domain adaptation for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[3] W. Zheng, M. Xin, X. Wang, and B. Wang, “A novel speech emotion recognition method via incomplete sparse least square regression,” *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 569–572, May 2014.

[4] L. Zão, D. Cavalcante, and R. Coelho, “Time-frequency feature and AMS-GMM mask for acoustic emotion classification,” *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 620–624, May 2014.

[5] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, “Generalized deep transfer networks for knowledge propagation in heterogeneous domains,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 4s, Nov. 2016. [Online]. Available: https://doi.org/10.1145/2998574

[6] X. Shu, G.-J. Qi, J. Tang, and J. Wang, “Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation,” in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 35–44. [Online]. Available: https://doi.org/10.1145/2733373.2806216

[7] S. Mai, H. Hu, and S. Xing, “Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing,” in *Proc. Anterior Cruciate Ligament*, Jul. 2019, pp. 481–492.

[8] A. Zadeh *et al.*, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proc. Anterior Cruciate Ligament*, 2018, pp. 2236–2246.

[9] S. Mai, H. Hu, and S. Xing, “Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 164–172.

[10] H. Pham, P. P. Liang, T. Manzini, L. P. Morency, and P. Barnabǎs, “Found in translation: Learning robust joint representations by cyclic translations between modalities,” in *Proc. Conf. Artif. Intell.*, 2019, pp. 6892–6899.

[11] S. Mai, S. Xing, and H. Hu, “Locally confined modality fusion network with a global perspective for multimodal human affective computing,” *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 122–137, Jan. 2020.

[12] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[13] K. Cho *et al.*, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[14] G. Chen and X. Zeng, “Multi-modal emotion recognition by fusing correlation features of speech-visual,” *IEEE Signal Process. Lett.*, vol. 28, pp. 533–537, 2021.

[15] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[16] 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 *Proc. Anterior Cruciate Ligament*, Jul. 2019, pp. 6558–6569.

[17] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, “Multimodal routing: Improving local and global interpretability of multimodal language analysis,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1823–1833.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*.

[19] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, *arXiv:1503.02531*.

[20] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, “MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, 2016.

[21] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, “Memory fusion network for multi-view sequential learning,” in *Proc. Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[22] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1114–1125.

[23] Z. Liu, Y. Shen, P. P. Liang, A. Zadeh, and L. P. Morency, “Efficient low-rank multimodal fusion with modality-specific factors,” in *Proc. Anterior Cruciate Ligament*, 2018, pp. 2247–2256.

[24] D. Hazarika, R. Zimmermann, and S. Poria, “MISA: Modality-invariant and -specific representations for multimodal sentiment analysis,” in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131. [Online]. Available: https://doi.org/10.1145/3394171.3413678

[25] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, “Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.

[26] W. Rahman *et al.*, “Integrating multimodal information in large pretrained transformers,” in *Proc. Annual Meeting Assoc. Comput. Linguist.*, 2020, pp. 2359–2369.

[27] 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 *Proc. Conf. Artif. Intell.*, vol. 33, 2019, pp. 7216–7223.

[28] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, “Quantum-inspired multimodal fusion for video sentiment analysis,” *Inf. Fusion*, vol. 65, pp. 58–71, 2021. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253520303365

[29] A. A. Ismail, M. Hasan, and F. Ishtiaq, “Improving multimodal accuracy through modality pre-training and attention,” 2020, *arXiv:2011.06102*.