JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text-Audio Sentiment Analysis Using Cross-Modal BERT

Shubham Das
M.Tech. (Data Science)
ASET, Amity University
Noida, India
shubhamdas747@gmail.com

Dr. Tanya Singh
Professor
ASET, Amity University
Noida, India
tsingh2@amity.edu

***Abstract*— Multimodal sentiment analysis is a relatively new field of research with the aim of enabling machines to perceive, analyse, and express emotion. We can learn more in-depth information about the speaker's emotional qualities through the cross-modal engagement. Bidirectional Encoder Representations from Transformers is a powerful pre-trained language representation model (BERT). Fine-tuning has provided novel, state-of-the-art results on ten natural language processing tasks, that included question-answering and natural language inference. Although the majority of earlier studies that improved BERT only used text data, it is still worthwhile to investigate how to learn a better representation by incorporating multimodal data. The Cross-Modal BERT (CM BERT), which we suggest in this research, uses the communication between text and audio modality to hone the pre-trained BERT model. Masked multimodal attention, the core aspect of CM-BERT, combines the knowledge retrieved from text and audio modalities to dynamically modify the weight of words. On the open multimodal sentiment analysis datasets CMU-MOSI and CMU-MOSEI, we test our methodology. The findings of the experiment reveal that it has greatly outperformed prior baselines and text-only finetuning of BERT in terms of performance on all criteria. In addition, by using audio modality information, we demonstrate the masked multimodal attention and demonstrate that it can appropriately modify the weight of words.**

***Keywords*—Natural Language Processing, BERT, Cross-Modal BERT (CM-BERT)**

1. Introduction

As a result of the development of communication technology and the widespread use of social media platforms like Facebook and Youtube, people generate a substantial amount of multimodal data each day that is rich in sentiment information. Emotion is essential to human interpersonal communication. Sentiment analysis, one of the essential technologies for human-computer interaction, is widely used in a range of application scenarios, such as automatic driving and human-machine conversation, and it has a substantial impact on progress in artificial intelligence [1]. Text is a fundamental kind of communication that uses words, sentences, and relationships to express emotion [28]. Text sentiment analysis has made significant advancements in recent years. For instance, TextCNN [13] outperforms the state-of-the-art on 4 out of 7 tests. TextCNN was trained on top of pre-trained word vectors for sentence-level classification tasks. The text modality has a rather constrained capacity for information. It could be challenging to appropriately discern emotion from writing in some situations. Text and aural modes are commonly blended in real life. The audio modality's available sentiment data can be identified by changes in pitch, energy, vocal effort, loudness, and other frequency-related parameters of voice quality [14]. It might be able to communicate more specific information and capture more emotional traits when text and speech are combined [3]. Figure 1 provides an example of the relationship between text and audio medium. But you know he did it" has a vague emotional meaning and shall be used to express a variety of emotions in different settings. It's challenging to understand this line's meaning just from its words. The speaker's low voice and sobs immediately after the introduction of the pertinent audio data make it clear that this line has a negative tone. Multimodal sentiment analysis addresses single-modality restrictions as the developing field of emotional computing has drawn a lot of attention [12]. Data from many modalities are combined through inter-modal interaction in multimodal fusion. The accuracy of the final result or conclusion is often increased since the combined information may include extra emotional aspects [18].

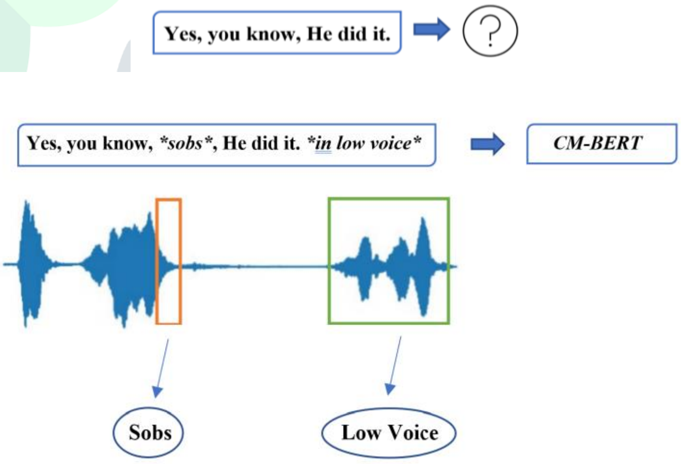
Yes, you know, He did it.
?
Yes, you know, *sobs*, He did it. *in low voice*
CM-BERT
Sobs
Low Voice

Figure 1. An example of cross-modal relationship between text and audio.

An efficient pre-trained language model called Bidirectional Encoder Representations from Transformers (BERT) recently showed state-of-the-art performance on a range of natural language processing tasks, including inference and question answering [5]. BERT develops contextual word representations as opposed to the traditional pre-trained language model by simultaneously conditioning on both left and right context in all levels. [15].

To perform well on a range of tasks at the token and sentence levels, pre-trained BERT has been enhanced [25]. However, since the majority of fine-tune processes are only based on text modality, it is still unclear how to adapt them to multimodality and provide better representations.

In this paper, we create a Cross-Modal BERT (CM-BERT), which incorporates audio modality data to help text modality fine-tune the pre-trained BERT model. The core of the CMBERT, masked multimodal attention, seeks to dynamically alter the weight of words through cross-modal interaction. The free and open-source multimodal sentiment analysis datasets CMU-MOSI [35] and CMU-MOSEI [36] are used to assess the efficacy of our model.

2. Related Works

2.1 Multi-modal Sentiment Analysis

Multimodal sentiment analysis is a current and popular area of study in natural language processing. Multimodal fusion can capture more intense emotional qualities for sentiment analysis when the internal correlation between several modalities is taken into account [2]. Multimodal fusion is difficult since it is unclear how to effectively combine the multimodal information. The two primary kinds of fusion approaches employed thus far are feature fusion and decision fusion [9, 21]. Concatenating features from different modalities or combining them in other ways is the aim of feature fusion. Fusion traits can unquestionably enhance performance because they carry more emotional information. The information from the audio and text modes was combined by Zhou et al. [37] to produce a semi-supervised multi-path generative neural network with improved emotion recognition. Zadeh et al. [32] developed a tensor fusion network that makes use of the product of multimodal features in order to more efficiently express multimodal fusion information. The experiment's findings demonstrate that Liu et al's multimodal fusion technique[17] improves sentiment analysis performance in addition to using low-rank tensors to increase efficiency in contrast to the tensor fusion network. Given their connection, the two terms might influence one another. Poria et al. [22] have established a contextual long, short term memory network that can capture more emotional characteristics by utilising contextual data at the utterance level and considering the relationship between the utterances.

A decision vector is produced when the outcomes of the independent categorization and analysis of the features of various modalities are combined. Dobriek et al. [6] sum and weighted product rules were utilised for audio and video decision-level fusion; the experiment's findings demonstrate that the weighted product outperforms the weight sum. Along with their prominence, the role of attention processes in multimodal fusion is expanding. By using a multi-attention block, a multi-attention recurrent network, developed by Zadeh et al. [34], can be utilised to identify interactions between different modalities. A multimodal multi-utterance bi-modal attention framework was introduced by Ghosal et al. [11] to study the factors that influence multimodal representations. In their Multimodal Transformer model, Tsai et al. [26] developed a directional paired crossmodal attention that can latently adapt streams from one modality to another while listening to interactions across multimodal sequences over different time steps..

2.2 Pre-Trained Language Model

Pre-trained language models are increasingly being used because they perform better on a variety of sentence- and token-level tasks, such as named entity recognition and question-answering [7]. The Extracted features from Language Models (ELMo) was developed by Peters et al. [19] and is pre-trained on a sizable text corpus using a deep bidirectional language model. The experiment's findings show that performance on six tasks can be significantly enhanced. The Generative Pre-trained Transformer was then presented by Radford et al. [23] to teach a universal representation (GPT). In contrast to previous approaches, they applied task-aware input transformations during fine-tuning, and it can be effectively transferred with little architecture change. A masked language model that was already been trained on the unsupervised prediction tasks Masked LM and Next Sentence Prediction is called Bidirectional Encoder Representations from Transformers (BERT), which is different from ELMo and GPT. The pre-trained BERT was modified to provide results that were state-of-the-art and definitely superior to those of earlier pre-trained language models on ten natural language processing tasks [5, 10].

3. Methodology

In this paper, the Cross-Modal BERT (CM-BERT), a technique for integrating data from the text and audio modalities into the pre-trained BERT model, is introduced. Its essential element is the employment of hidden multimodal attention to cross-modal interaction to change the weight of words on the fly. The problem specification is covered in the subsections i.e. followed in Section 3.1, while Section 3.2 presents the CM-BERT model's design. Disguised multimodal attention is a concept that is introduced in Section 3.3.

3.1 Problem Definition

A text sequence made up of word-piece tokens is given as T = [T1,T2,...Tn], where, n is the sequence length. The last encoder layer produces a sequence of length n+ 1 since the embedding layer of the BERT model adds a particular classification embedding ([CLS]) before the input sequence. This is indicated by the expression Xt = [E[CLS], E1, E2,...En]. The word-level alignment audio features (described in Section 4.2) are prefixed with a zero vector to be consistent with text modality, and the features of the audio are designated as Xa = [A[CLS], A1, A2,...An], where A[CLS] is a zero vector. Our approach aims to better fine-tune the pre-trained BERT model and By adjusting the weight of each word through the interplay between Xt and Xa, performance of sentiment analysis can be enhanced.

3.2 Cross-Modal BERT (CM-BERT)

Figure 2 depicts the configuration of the CM-architectural BERT. The word-piece token sequence from the text and the audio features for word-level alignment make up the two components that make up the CM-BERT model's input. After the text sequence has initially been run through the trained BERT model, the output of the final encoder layer—defined as Xt = [E[CLS], E1, E2,...En]—will be used as the text features. We utilise a 1D temporal convolutional layer to bring the text attributes Xt, which are obviously bigger in dimension than the word-level alignment audio characteristics Xa, to the same dimension:

$$\{X\hat{}\ t ,X\hat{}\ a \} = \text{Conv 1D} (\{Xt , Xa \}, k\{t,a\})$$

where kt,a are the corresponding convolutional kernel sizes for the text and audio modalities.

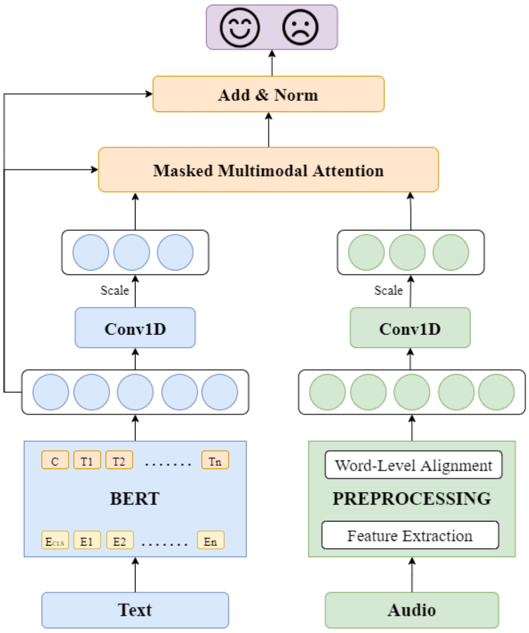

Figure 2. Overview of the Cross-Modal BERT Network's architecture.

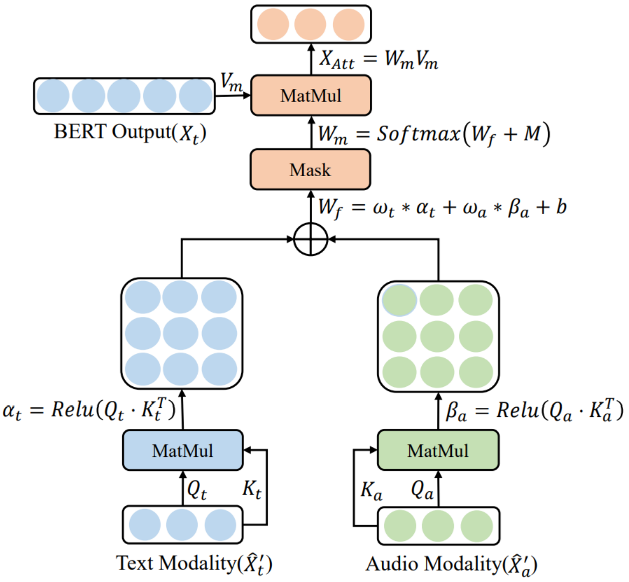
$X_{Att} = W_m V_m$
V_m
MatMul
BERT Output(X_t)
$W_m = Softmax(W_f + M)$
Mask
$W_f = \omega_t * \alpha_t + \omega_a * \beta_a + b$
$\alpha_t = Relu(Q_t \cdot K_t^T)$
$\beta_a = Relu(Q_a \cdot K_a^T)$
MatMul
MatMul
Q_t
K_t
K_a
Q_a
Text Modality($\hat{X}_t'$)
Audio Modality($\hat{X}_a'$)

Figure 3. The masked-multimodal attention architecture.

Since Xt has a much higher dimension than Xa, its worth will grow significantly throughout the course of training. To prevent the dot products from becoming exceedingly large in magnitude and forcing the softmax function into incredibly tight gradient areas, we scale the text features X t to X t′ and the audio characteristics X a to X a′ :

$$X\hat{}t\,' = X\hat{}t\ /\mathrm{q}\ X\hat{}t\ 2$$

$$X\hat{}a\,' = X\hat{}a\ /\mathrm{q}\ X\hat{}a\ 2$$

As soon as we have Xt, Xt′ , and Xa, we enter them into the masked multimodal attention, which can modify the weight of words by fusing the way the words are performed across several modalities., to entirely interactively combine text and auditory information. In order to preserve the original structure of the data, we apply a residual connection on Xt and XAtt after receiving the result of the multimodal attention masked XAtt. After that, a normalisation layer and a linear layer will be applied. The output of the last linear layer may now be accessed, Yl = [L[CLS], L1, L2,...Ln.

3.3 Masked Multi-modal Attention

The core aspect of the CM-BERT, known as masked multimodal attention, modifies the pre-trained BERT model and the text modality with the aid of audio input. The organisational structure of the veiled multimodal attention is shown in Figure 3. We begin by evaluating each word's significance in several modalities. The text modality's query Qt and key Kt are determined by the scaled text features, or Xt′, with Qt = Kt = Xt′. The Query Qa and the Key Ka of the audio modality are defined by the scaled word-level alignment audio characteristics, or Xa′.

The text attention matrix t, and the audio attention matrix an are defined as follows:

$$\alpha t = \mathrm{Relu}(QtKt')$$

$$\beta a = \mathrm{Relu}(QaKa')$$

The weighted fusion attention matrix Wf is calculated as follows: to adjust each word's emphasis through the summation of the auditory and textual modalities, we weight sum the text attention the audio attention matrix a matrix t.

$$Wf = wt * \alpha t + wa * \beta a + \mathrm{b}$$

where wt and wa stand for the next weights of the text and audio modalities, and b represents the bias. We develop a mask matrix M that uses 0 to represent the token position and $-\infty$ utilises to represent the padding position in order to lessen the impact of the padding sequence (after softmax function the attention score of padding position will be 0). The multimodal attention matrix Wm is therefore described as follows:

$$\mathrm{Wm} = \mathrm{Softmax}\ (\ Wf + M\ )$$

To get the result of the multimodal attention XAtt matrix, we multiply Wm by the value of the multimodal attention that has been masked, Vm:

$$XAtt = WmVm$$

where Vm is the final encoder layer output of the BERT, defined as Vm = Xt.

4. Experimental Methodology

On the open multimodal sentiment analysis datasets CMUMOSI and CMU-MOSEI, we assess the Cross-Modal BERT's performance in this section. We will discuss our experiments from the subsequent angles. We will start by discussing the details of the datasets and the experimental environment. Following that, we'll discuss the audio features and multimodal alignment. Finally, we will provide the evaluation criteria and baselines that we employed in our study.

4.1 Dataset & Experimental Environment

We evaluate the efficacy of our methodology using the CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) datasets. CMU-MOSI consists of 93 YouTube movie reviews and opinion videos. The videos cover over 2199 utterances. Five different workers classify each statement on a scale from -3 to +3, with -3 meaning extremely negative and 3 denoting highly positive. We split the 1284, 229, and 686 utterances from 52, 10, and 31 films into training, validation, and test sets, respectively, keeping in mind that the speaker shouldn't appear in both the testing and training sets also, the balance of the positive and negative data. CMU-MOSEI is a emotion analysis dataset and multimodal sentiment that, like CMU-MOSI, is composed of 23,454 YouTube movie review video clips. The strategy we employ is consistent with previously published publications [26, 30].

Our proposed CM-BERT employs the pre-trained BERT model's uncased BERTBASE variant, which contains 12 transformer blocks. In order to prevent overfitting, we set the learning rates of the encoder layers to 0.01 and the remaining layers to 2e-5. To enhance performance, we freeze the embedding layer's parameters. For training the CM-BERT model, we set the batch size, maximum sequence length, and number of epochs to 24 and 50, respectively. We used the Adam optimizer with mean-square error as the loss function.

Table 1. Experimental results on CMU-MOSI dataset.

Model	Modality	Acc7(h)	Acc2(h)	F1 Score	MAE(h)	Corr(h)
LMF [17]	T+A+V	32.7	76.3	75.5	0.913	0.666
MFN [33]	T + A + V	34.2	77.3	77.4	0.964	0.633
MARN [34]	T + A + V	34.5	77.2	77.1	0.966	0.634
RMFN [16]	T + A + V	38.2	78.3	78.1	0.923	0.682
MFM [27]	T + A + V	36.1	78.2	78.3	0.952	0.662
MCTN [20]	T + A + V	36.5	79.2	79.3	0.908	0.675
MulT [26]	T + A + V	40.1	83.1	82.7	0.872	0.697
T-BERT [5]	T	41.4	83.1	83.3	0.783	0,775
CM-BERT (ours)	**T + A**	**44.8**	**84.4**	**84.5**	**0.730**	**0.792**

4.2 Audio Features & Multi-Modal Alignment

COVAREP [4] is used in this study to extract audio features. Each segment is represented by a 74-dimensional feature vector, which contains twelve Mel-frequency cepstral coefficients (MFCCs), pitch and segmenting characteristics, glottal source parameters, peak slope parameters, and maxima dispersion quotients. Using P2FA [31] in the wake of [26], we acquire the timesteps for each word in order to extract the word-level alignment characteristics. Following that, the audio elements contained inside the pertinent word timesteps are averaged. The audio sequences are padded with zero vectors so that they match the sequence length of the text mode.

4.3 Evaluation Metrics

In line with prior research [30], the performance of our model and the baselines using the same assessment metrics in our experiment. The 7-class accuracy (Acc7) is used in the sentiment score classification task, the 2-class accuracy (Acc2) and the F1 score (F1) are used in the binary-sentiment classification task, and the mean absolute error (MAE) and correlation (Corr) of model predictions with true labels are used in the regression task. In addition to MAE, the model will perform better when the other measures have higher values. We select five random seeds at random and average the results of five runs to increase the validity of the experiment results.

4.3 Baseline Models

We evaluate CM-multimodal BERT's sentiment analysis performance in comparison to earlier models. The models that we compared are as follows:

LMF [17] Low-rank weight tensors are used in a method known as low-rank multimodal fusion, or LMF, to enhance performance without reducing efficiency. Performance is much improved, and computational complexity is significantly decreased.

MFN [33] The Memory Fusion Network (MFN) is mainly composed of System of LSTMs, Delta-memory Attention Network, and Multi-view Gated Memory, which explicitly accounts for both interactions in neural architecture and continuously simulates them.

MFM [27] The Multimodal Factorization Model (MFM) can factorise the multimodal representations into multimodal discriminative factors and modality-specific generative factors, helping each factor focus on learning from a portion of the joint knowledge across multimodal data and labels.

MCTN [20] Multimodal Cyclic Translation Network (MCTN), which solely works with text modality data during testing, generates brand-new, cutting-edge output. It is intended to translate between several modalities to learn reliable joint representations.

MulT [26] The most advanced technique utilised on the MOSI dataset is Multimodal Transformer (MulT), a state-of-the-art technique that latently changes streams of one modality to another. MulT employs directed paired crossmodal attention to examine interactions among multimodal sequences over a number of time steps.

A Bidirectional Encoder Representations from Transformers (BERT) called **T-BERT [5]** only employs text modality information for fine-tuning.

5. Results & Discussion

The results of our test of the CM-BERT model using the CMU-MOSI dataset are shown in Table 1. It is obvious that the CM-BERT model produces a brand-new, brand new result on the MOSI dataset and enhances performance across the board. The binary-sentiment classification task yields an Acch2 score of 84.4 percent for the CM-BERT model, which is around 1.5 percent -9.3 percent better than baselines. Our model raises F 1 from 8.6 to 9.2 percent, similar to Acch2. The sentiment score categorization task exhibits the strongest influence of the CM-BERT model on improvement. Our model outperforms the baselines by 4.8 to 12.0 percentage points on Acch7, scoring 44.9 percent. In the regression task, the CM-BERT increases 0.093-0.183 on Corrh while decreasing 0.142-0.294 on MAEl. It is interesting that the p-value for the student t-test comparing CM-BERT with T-BERT in Table 1 is significantly less than 0.05 on all measures.

T-BERT is the only baseline that utilises data from text, audio, and video modalities; nevertheless, our model only employs this data to produce a novel, cutting-edge result. According to the experimental results, the MulT model performs significantly better than the other baselines. The main reason for this is that the MulT extends transformer to multimodal situations and latently adapts elements across modalities using attention. The MulT model outperforms the T-BERT model in terms of performance since it can enhance the representations of the pre-trained BERT model. The pre-trained BERT model is extended from unimodal to multimodal in the CM-BERT model we developed, in contrast to the T-BERT model, and audio modality information is added to help text modality effectively modify the weight of words. Because the CM-BERT model can more correctly reflect the speaker's emotional state and can capture more emotive elements through the relation between textual and audio modalities, its performance on all assessment criteria is significantly enhanced.

To demonstrate the applicability of our methodology to other multimodal language datasets, we also conduct tests on the CMU-MOSEI dataset. Following earlier research [24], we compare the Acch 2 and F 1 for the top 3 models in Table 1 for the sake of convenience. First off, the MulT scores an 82.7 on Acch 2 and an 82.3 on F 1. When compared to MulT, T-BERT performs better, scoring 83.3 percent on Acch 2 and 83.1 percent on F 1. Additionally, CM-BERT succeeds at 84.5 percent on Acch 2 and 84.4 percent on F 1. Our model improves on Acc 2 and F 1 by roughly 1.3 percent and 1.2 percent, respectively, as compared to the MulT and T-BERT. As a result, our suggested method's higher performance on the CMU-MOSEI dataset also supports its generalizability.

6. Conclusion

In this paper, we provide a Cross-Modal BERT multimodal sentiment analysis model (CM-BERT). In contrast to past efforts, we change the pre-trained BERT model from unimodal to multimodal. We introduce the audio modality information in order to enhance representations and aid text modality BERT. Through the interaction of the textual and auditory modalities, the central component of CM-BERT, camouflaged multimodal attention, tries to dynamically modify the weight of words. The results show that, on the CMU-MOSI and CMU-MOSEI datasets, CM-BERT performs significantly better than earlier baselines and text-only finetuning of BERT. In fact, CM-BERT can be used to integrate more than two modalities and is appropriate for both text and video. The main focus of future research will be on using neural networks to align different data modalities using pre-trained models to extract more accurate representations from multimodal data that is generally not aligned. This is due to the fact that multimodal data is rarely aligned in the real world.

7. References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multi modal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2 (2018), 423–443.

[2] Linqin Cai, Yaxin Hu, Jiangong Dong, and Sitong Zhou. 2019. Audio-Textual Emo tion Recognition Based on Improved Neural Networks. Mathematical Problems in Engineering 2019 (2019).

[3] 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. In Proceedings of the 19th ACM International Conference on Multimodal Interaction. 163–171.

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technolo gies. In 2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 960–964.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171–4186.

[6] Simon Dobrišek, Rok Gajšek, France Mihelič, Nikola Pavešić, and Vitomir Štruc. 2013. Towards efficient multi-modal emotion recognition. International Journal of Advanced Robotic Systems 10, 1 (2013), 53.

[7] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre training for natural language understanding and generation. In Advances in Neural Information Processing Systems. 13042–13054.

[8] Shen Gao, Xiuying Chen, Piji Li, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2019. How to Write Summaries with Patterns? Learning towards Abstractive Summa rization through Prototype Editing. In Proceedings of the 2019 Conference on Empirical Methods in Natural L anguage Processing and the 9th International Joint Conference on Natural Language Pro cessing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3741–3751.

[9] Shen Gao, Xiuying Chen, Chang Liu, Li Liu, and Rui Zhao, Dongyan an d Yan. 2020. Learning to Respond with Stickers: A Framework of Unifying Multi Modality in Multi-Turn Dialog. In Proceedings of The Web Conference 2020 (Taipei, Taiwan) (WWW '20). Association for Computing Machinery, New York, NY, USA, 1138–1148. https://doi.org/10.1145/3366423.3380191

[10] Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Ya n. 2020. From Standard Summarization to New Tasks and Beyond: Summarization wit h Manifold Information. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, Christian Bessiere (Ed.). Interna tional Joint Conferences on Artificial Intelligence Organizatio n, 4854–4860. https://doi.org/10.24963/ijcai.2020/676 Survey track.

[11] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for

multi-modal sentiment analysis. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 3454–3466.

[12] Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. Trends in Cognitive Sciences (2019).

[13] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 1746–1751.

[14] Runnan Li, Zhiyong Wu, Jia Jia, Yaohua Bu, Sheng Zhao, and Helen Meng. 2019. Towards discriminative representation learning for speech emotion recognition. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). 5060–5066.

[15] Xinlong Li, Xingyu Fu, Guangluan Xu, Yang Yang, Jiuniu Wang, Li Jin, Qing Liu, and Tianyuan Xiang. 2020. Enhancing BERT Representation With Context Aware Embedding for Aspect-Based Sentiment Analysis. IEEE Access 8 (2020), 46868–46876.

[16] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 150–161.

[17] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. 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). 2247–2256.

[18] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Systems 161 (2018), 124–133.

[19] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of NAACL-HLT. 2227–2237.

[20] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barn abás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 6892–6899.

[21] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37 (2017), 98–125.

[22] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment anal ysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 873–883.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by genera tive pre-training. URL https://s3-us-west-2. amazonaws.
com/openai assets/researchcovers/languageunsupervised/language understanding paper. pdf (2018).

[24] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2359–2369.

[25] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification?. In China National Conference on Chinese Computational Linguistics. Springer, 194–206.

[26] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

[27] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In International Conference on Representation Learning.

[28] Matthew Turk. 2014. Multimodal interaction: A review. Pattern Recognition Letters 36 (2014), 189–195.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. 5998–6008.

[30] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis Philippe Morency. 2019. Words can shift: Dynamically adjusting word repre sentations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 7216–7223.

[31] Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. Journal of the Acoustical Society of America 123, 5 (2008), 3878.

[32] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 1103–1114.

[33] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In Thirty-Second AAAI Conference on Artificial Intelligence.

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

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

[36] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2236–2246.

[37] Suping Zhou, Jia Jia, Qi Wang, Yufei Dong, Yufeng Yin, and Kehua Lei. 2018. Inferring emotion from conversational voice data: A semi-supervised multi path generative neural network approach. In Thirty-Second AAAI Conference on Artificial Intelligence