CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis

Kaicheng Yang
State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China & School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
yangkaicheng@stu.hebust.edu.cn

Hua Xu*
State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China & Beijing National Research Center for Information Science and Technology(BNRist),
Beijing 100084, China
xuhua@tsinghua.edu.cn

Kai Gao*
School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China
gaokai@hebust.edu.cn

ABSTRACT

Multimodal sentiment analysis is an emerging research field that aims to enable machines to recognize, interpret, and express emotion. Through the cross-modal interaction, we can get more comprehensive emotional characteristics of the speaker. Bidirectional Encoder Representations from Transformers (BERT) is an efficient pre-trained language representation model. Fine-tuning it has obtained new state-of-the-art results on eleven natural language processing tasks like question answering and natural language inference. However, most previous works fine-tune BERT only base on text data, how to learn a better representation by introducing the multimodal information is still worth exploring. In this paper, we propose the Cross-Modal BERT (CM-BERT), which relies on the interaction of text and audio modality to fine-tune the pre-trained BERT model. As the core unit of the CM-BERT, masked multimodal attention is designed to dynamically adjust the weight of words by combining the information of text and audio modality. We evaluate our method on the public multimodal sentiment analysis datasets CMU-MOSI and CMU-MOSEI. The experiment results show that it has significantly improved the performance on all the metrics over previous baselines and text-only finetuning of BERT. Besides, we visualize the masked multimodal attention and proves that it can reasonably adjust the weight of words by introducing audio modality information.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; *Multimedia and multimodal retrieval*; • **Computing methodologies** → Natural language processing;

KEYWORDS

Multimodal sentiment analysis; Pretrained model; Attention network

ACM Reference Format:
Kaicheng Yang, Hua Xu, and Kai Gao. 2020. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394171.3413690

1 INTRODUCTION

With the advancement of communication technology and the popularity of social platforms such as Facebook and YouTube, people produce a large number of multimodal data with rich sentiment information every day. Sentiment plays a crucial role in human interpersonal communication. Sentiment analysis as one of the critical technologies of human-computer interaction affects the development of artificial intelligence, and it has been widely used in many application scenarios, such as human-machine conversation, automatic drive, and so on [1]. Text is an essential modality in our daily life, it expresses sentiment through words, phrases, and relations [28]. In the past few years, text sentiment analysis has achieved a lot of achievements, for example, TextCNN [13] trained on top of pre-trained word vectors for sentence-level classification tasks and it improves upon the state-of-the-art on 4 out of 7 tasks.

However, the information contained in text modality is limited. In some cases, it is difficult to judge emotion accurately by text information. In daily life, text modality is often accompanied by audio modality. The sentiment information contained in audio modality is characterized by the variations in voice characteristics such as pitch, energy, vocal effort, loudness, and other frequency-related measures [14]. The interaction between text and audio modality can provide more comprehensive information and capture more emotional characteristics [3]. Figure 1 is an example of the inter-modality interaction between text and audio modality. The emotion of the sentence "But you know he did it" is ambiguous, and it can express a variety of emotions in different situations. It is challenging to determine the sentiment of this sentence according to these words. After introducing corresponding audio information, because of the speaker's low voice and sobs, it is not difficult to predict the sentiment of this sentence is negative. To make up for the disadvantage of the single modality, multimodal sentiment analysis as

[*] Corresponding author

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413690

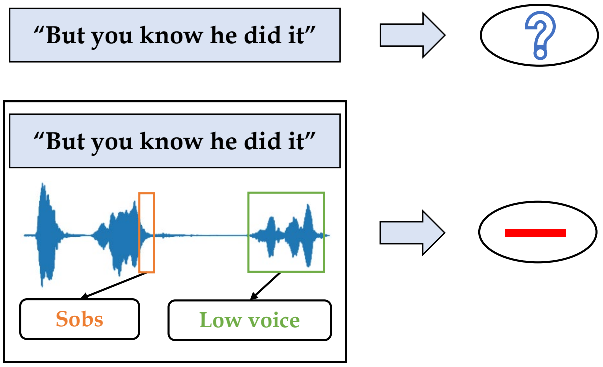


Figure 1: An example to illustrate the cross-modal interaction between text and audio modality.

an increasingly extensive field of affective computing has attracted widespread attention [12]. Multimodal fusion is to combine the information from different modalities through the inter-modality interaction. Cause the fusion information can provide more emotional characteristics, it often increases the accuracy of the overall result or decision [18].

Recently, Bidirectional Encoder Representations from Transformers (BERT) as an efficient pre-trained language model has presented state-of-the-art results on eleven natural language processing tasks, including question answering, natural language inference, and others [5]. Different from the traditional pre-trained language model, BERT generates contextual word representations by jointly conditioning on both left and right context in all layers. So, the representations of the words can describe the context content [15]. Fine-tuning pre-trained BERT has achieved efficient performance on a large number of sentence-level and token-level tasks [25]. However, most fine-tune strategies are designed only base on text modality, how to extend it from unimodal to multimodal and get better representations is still an open research problem.

In this paper, we propose a Cross-Modal BERT (CM-BERT) that introduces the information of audio modality to help text modality fine-tune the pre-trained BERT model. As the core unit of the CM-BERT, masked multimodal attention is designed to dynamically adjust the weight of words through the cross-modal interaction. To prove the effectiveness of our method, we evaluate it on the public multimodal sentiment analysis datasets CMU-MOSI [35] and CMU-MOSEI [36]. The experiment results show that the CM-BERT has significantly improved the performance on all the metrics over previous baselines and text-only finetuning of BERT. The main contributions of this paper can be summarized as follows:

- We propose a Cross-Modal BERT (CM-BERT) model that introduces the information of audio modality to help text modality fine-tune the pre-trained BERT model.
- We design a novel masked multimodal attention that can dynamically adjust the weight of words through the interaction between text and audio modality.
- We show our model only uses the data of text and audio modality to create a new state-of-the-art multimodal sentiment analysis results on the public sentiment benchmark datasets CMU-MOSI and CMU-MOSEI[1].

2 RELATED WORK

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis is a new popular research area in natural language processing. Considering the internal correlation between different modalities, multimodal fusion can capture more effective emotional characteristics for sentiment analysis [2]. The difficulty of multimodal fusion lies in how to integrate the multimodal information effectively. To date, there are mainly two types of fusion strategies: feature fusion and decision fusion [9, 21]. Feature fusion is to fuse the features of different modalities through concatenating and other ways. Because the fusion features contain more emotional information, it can improve the performance obviously. Zhou et al. [37] combined the features of text and audio modality and designed a semisupervised multi-path generative neural network to better infer emotion. To get better multimodal information representation, Zadeh et al. [32] proposed a tensor fusion network which use the product of multimodal features to represent the multimodal fusion information. Different from the tensor fusion network, Liu et al. [17] employed a low-rank multimodal fusion method which use low-rank tensors to improve efficiency, the experiment results show that not only it reduces the parameters but also enhances the sentiment analysis performance. The utterances are correlated, and they can affect each other. Considering the relationship among the utterances, Poria et al. [22] introduced a contextual long short term memory network that can utilize the utterance-level contextual information to capture more emotional characteristics. In the decision fusion process, the features of different modalities are examined and classified independently and the results of them are fused as a decision vector to obtain the final decision. Dobrišek et al. [6] employed weight sum and weighted product rule for audio and video decision-level fusion, the experiment results show that the performance of the weighted product is better than weight sum.

With the popularity of attention mechanism, it plays an increasingly important role in multimodal fusion. Zadeh et al. [34] presented a multi-attention recurrent network, which can discover the interaction between different modalities by using a multi-attention block. Ghosal et al. [11] proposed a Multimodal Multi-utterance-Bi-modal Attention framework which employed attention on multimodal representations to learn the contributing features among them. Besides, Tsai et al. [26] used a directional pairwise crossmodal attention in their Multimodal Transformer model, and it can attend to interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another.

2.2 Pre-trained Language Model

In recent years, the pre-trained language model has been widely used in natural language processing, and it has improved performance on an extensive suite of sentence-level and token-level tasks

[1] Code released in https://github.com/thuiar/Cross-Modal-BERT

such as question answering and named entity recognition [7]. Peters et al. [19] introduced the Embeddings from Language Models (ELMo), which is pre-trained on a large text corpus by using a deep bidirectional language model. The experiment results show that it can significantly improve performance across six tasks. After that, to learn a universal representation, Radford et al. [23] presented the Generative Pre-trained Transformer (GPT). Contrast to previous approaches, they utilized task-aware input transformations during fine-tuning, and it can well transfer with minimal architecture to change. Different from ELMo and GPT, Bidirectional Encoder Representations from Transformers(BERT) is a masked language model, and it is pre-trained by using two unsupervised prediction tasks Masked LM and Next Sentence Prediction. Fine-tuning the pre-trained BERT has obviously outperformed other pre-trained language models, and it has created new state-of-the-art results on eleven natural language processing tasks [5, 10].

3 METHODOLOGY

In this paper, we propose the Cross-Modal BERT (CM-BERT), it can combine the information from text and audio modality to fine-tune the pre-trained BERT model. As the core of it, masked multimodal attention is employed to dynamically adjust the weight of words through the cross-modal interaction. In the following subsections, Section 3.1 discusses the problem definition. Section 3.2 describes the architecture of the CM-BERT model. Section 3.3 presents the principle of masked multimodal attention.

3.1 Problem Definition

Given a text sequence of word-piece tokens $T = [T_1, T_2, ...T_n]$, where n is the number of sequence length. Since the embedding layer of BERT model will append a special classification embedding ($[CLS]$) before the input sequence, the output of the last encoder layer is a $n+1$ length sequence which is denoted as $X_t = [E_{[CLS]}, E_1, E_2, ...E_n]$. To be consistent with text modality, we append a zero vector before the word-level alignment audio features (introduced in Section 4.2), and the audio features are denoted as $X_a = [A_{[CLS]}, A_1, A_2, ...A_n]$, where $A_{[CLS]}$ is a zero vector. The goal of our method is to utilize the interaction between X_t and X_a to adjust the weight of each word, so as to better fine-tune the pre-trained BERT model and improve the sentiment analysis performance.

3.2 CM-BERT: Cross-Modal BERT

The architecture of the CM-BERT is shown in Figure 2. The input of the CM-BERT model consists of two parts: the text sequence of word-piece tokens and the word-level alignment audio features. Firstly, the text sequence will pass through the pre-trained BERT model, and the output of the last encoder layer is used as the text features, which is defined as $X_t = [E_{[CLS]}, E_1, E_2, ...E_n]$. Because the dimension of the word-level alignment audio features X_a is obviously smaller than the text features X_t, following [26], we employ a 1D temporal convolutional layer to control them to the same dimension:

$$\{\hat{X_t}, \hat{X_a}\} = \text{Conv 1D}\left(\{X_t, X_a\}, k_{\{t,a\}}\right) \tag{1}$$

where $k_{\{t,a\}}$ represent the size of convolutional kernels for text and audio modality. Because the dimension of X_t is significantly higher than X_a, in the training process, the value of $\hat{X_t}$ will be larger and larger than $\hat{X_a}$. In order to prevent the dot products grow large in magnitude and push the softmax function into extremely small gradients regions, we scale the text features $\hat{X_t}$ to $\hat{X_t}'$ and audio features $\hat{X_a}$ to $\hat{X_a}'$:

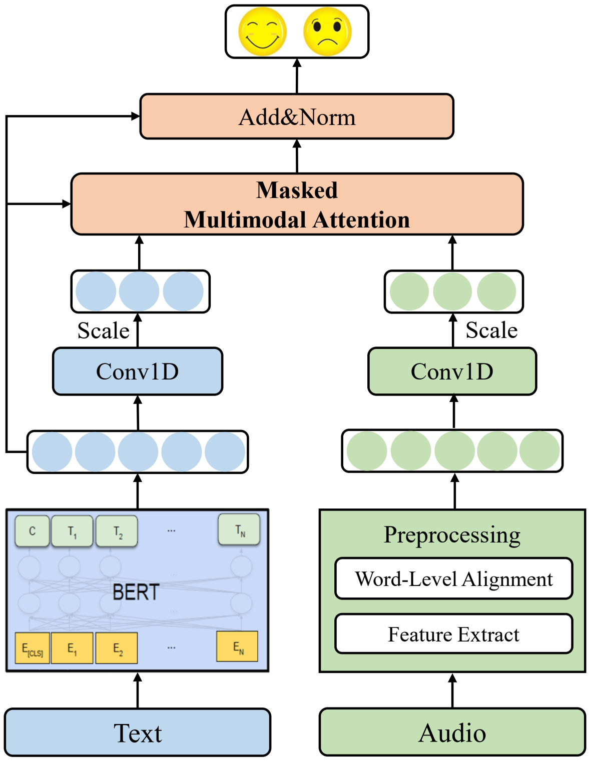


Figure 2: Overview architecture of the Cross-Modal BERT Network.

$$\hat{X_t}' = \frac{\hat{X_t}}{\sqrt{\|\hat{X_t}\|_2}} \tag{2}$$

$$\hat{X_a}' = \frac{\hat{X_a}}{\sqrt{\|\hat{X_a}\|_2}} \tag{3}$$

After getting X_t, $\hat{X_t}'$, and $\hat{X_a}'$, to make text and audio information fully interactive, we input them into the masked multimodal attention which can adjust the weight of words by combining the performance of the words in different modalities. After getting the output of the masked multimodal attention X_{Att}, following priors works [8, 29], we employ a residual connection on X_t and X_{Att} to keep the original structure of the data. Then it will pass through a linear layer and a normalization layer. Finally, we can get the output of the last linear layer $Y_l = [L_{[CLS]}, L_1, L_2, ...L_n]$. Because the representation of the first token $L_{[CLS]}$ is learned according to the information of the other tokens, we use it as the aggregate representation and input into a linear layer to produce the final predict results.

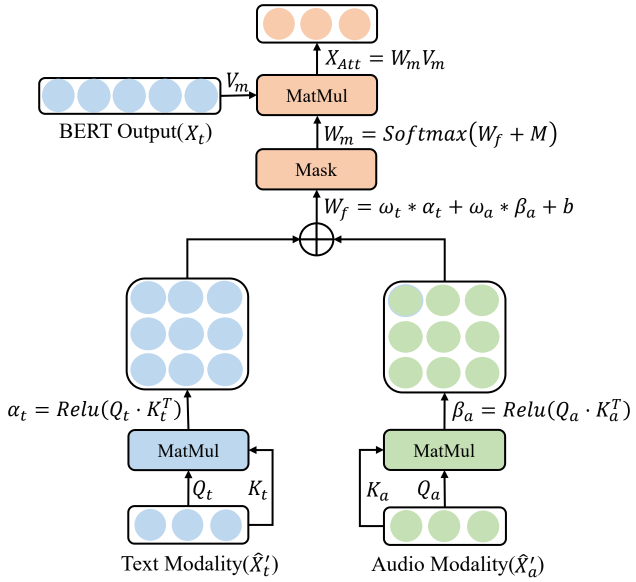

$X_{Att} = W_m V_m$
V_m
MatMul
BERT Output(X_t)
$W_m = Softmax(W_f + M)$
Mask
$W_f = \omega_t * \alpha_t + \omega_a * \beta_a + b$
$\alpha_t = Relu(Q_t \cdot K_t^T)$
$\beta_a = Relu(Q_a \cdot K_a^T)$
MatMul
MatMul
Q_t
K_t
K_a
Q_a
Text Modality($\hat{X}_t'$)
Audio Modality($\hat{X}_a'$)

Figure 3: The architecture of the masked multimodal attention.

3.3 Masked Multimodal Attention

The masked multimodal attention as the core of the CM-BERT is designed to utilize the information of audio modality to help text modality adjust the weight of words and fine-tune the pre-trained BERT model. The structure of the masked multimodal attention is shown in Figure 3. Firstly, we evaluate the weight of each word in different modalities. The Query Q_t and the Key K_t of text modality are defined as $Q_t = K_t = \hat{X_t}'$, where $\hat{X_t}'$ is the scaled text features. The Query Q_a and the Key K_a of audio modality are defined as $Q_a = K_a = \hat{X_a}'$, where $\hat{X_a}'$ is the scaled word-level alignment audio features. Then the text attention matrix α_t and the audio attention matrix β_a are defined as:

$$\alpha_t = \text{Relu}\left(Q_t K_t^\top\right) \tag{4}$$

$$\beta_a = \text{Relu}\left(Q_a K_a^\top\right) \tag{5}$$

To adjust the weight of each word through the interaction between text and audio modality, we weight sum the text attention matrix α_t and the audio attention matrix β_a, the weighted fusion attention matrix W_f is computed as:

$$W_f = w_t * \alpha_t + w_a * \beta_a + b \tag{6}$$

where w_t and w_a represent the weight of text and audio modality respectively, and b is the bias. To reduce the influence of padding sequence, we introduce a mask matrix M, which uses 0 to represent the token position and uses $-\infty$ to represent the padding position (after softmax function the attention score of padding position will be 0). Then the multimodal attention matrix W_m is defined as:

$$W_m = \text{Softmax}\left(W_f + M\right) \tag{7}$$

After obtaining the multimodal attention matrix, we multiply W_m with the Value of the masked multimodal attention V_m to get the output of the attention X_{Att}:

$$X_{Att} = W_m V_m \tag{8}$$

where V_m is the output of the BERT's last encoder layer, which is defined as $V_m = X_t$.

4 EXPERIMENTAL METHODOLOGY

In this section, we evaluate the performance of the Cross-Modal BERT on the public multimodal sentiment analysis datasets CMU-MOSI and CMU-MOSEI. We will introduce our experiments from the following aspects. Firstly, we will show the information about datasets and experimental setting. Then, we will present the audio features and multimodal alignment. Finally, we will introduce the evaluation metrics and baselines used in our experiment.

4.1 Datasets and Experimental Settings

We evaluate our approach on the CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) [35] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [36] datasets. CMU-MOSI consists of 93 opinion videos from YouTube movie reviews. The videos are spanning over 2199 utterances. The label of each utterance is annotated by 5 different workers and is in a continuous range of -3 to +3, where -3 indicates highly negative and 3 indicates highly positive. Considering the speaker should not appear in both training and testing sets and the balance of the positive and negative data, we split 52, 10, 31 videos in training, validation and test set accounting for 1284, 229, and 686 utterances. Similar to CMU-MOSI, CMU-MOSEI is a multimodal sentiment and emotion analysis dataset which is made up of 23,454 movie review video clips taken from YouTube. The strategy we adopt is consistent with the previously published works [26, 30].

The pre-trained BERT model used in our proposed CM-BERT is the uncased BERT$_{\text{BASE}}$ version, which consists of 12 transformer blocks. To prevent overfitting, we set the learning rate of the encoder layers to 0.01 and set the learning rate of the rest layers to 2e-5. To get better performance, we freeze the parameters of the embedding layer. For training the CM-BERT model, we set the batch size and the max sequences length to 24 and 50 respectively, the number of the epoch is set to 3. Besides, we use *Adam* optimizer with $mean - squareerror$ loss function.

4.2 Audio Features and Multimodal Alignment

In this work, we use COVAREP [4] to extract audio features. Each segment is represented as a 74-dimensional feature vector including 12 Mel-frequency cepstral coefficients (MFCCs), pitch and segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients. To get the word-level alignment features, following [26], we use P2FA [31] to get the timesteps of each word. Then we perform averaging on the audio features within the corresponding word timesteps. In order to keep consistent with the sequence length of text modality, zero vectors are used to pad audio sequences.

Table 1: Experimental results on CMU-MOSI dataset. The best results are highlighted in bold. ^{h} means higher is better and ^{l} means lower is better. T:text,A:audio,V:video.

Model	Modality	Acc_7^h	Acc_2^h	F_1^h	MAE^l	$Corr^h$
EF-LSTM	T+A+V	33.7	75.3	75.2	1.023	0.608
LMF [17]	T+A+V	32.8	76.4	75.7	0.912	0.668
MFN [33]	T+A+V	34.1	77.4	77.3	0.965	0.632
MARN [34]	T+A+V	34.7	77.1	77.0	0.968	0.625
RMFN [16]	T+A+V	38.3	78.4	78.0	0.922	0.681
MFM [27]	T+A+V	36.2	78.1	78.1	0.951	0.662
MCTN [20]	T+A+V	35.6	79.3	79.1	0.909	0.676
MulT [26]	T+A+V	40.0	83.0	82.8	0.871	0.698
T-BERT [5]	T	41.5	83.2	83.2	0.784	0.774
CM-BERT(ours)	T+A	**44.9**	**84.5**	**84.5**	**0.729**	**0.791**

4.3 Evaluation Metrics

In our experiment, consistent with previous work [30], we use the same evaluation metrics to evaluate the performance of the baselines and our model. 7-class accuracy (Acc_7) is used in the sentiment score classification task, 2-class accuracy (Acc_2) and F1 score (F_1) are used in the binary sentiment classification task, mean absolute error (MAE) and the correlation ($Corr$) of model predictions with true labels are used in the regression task. Besides MAE, the higher value of the metrics means the better performance of the model. To make the experiment results more convincing, we randomly select five random seeds and take the average result of 5 runs as the final experiment results.

4.4 Baselines

We compare the performance of CM-BERT with previous models on the multimodal sentiment analysis task. The models we compared are as follows:

EF-LSTM Early Fusion LSTM (EF-LSTM) concatenates multimodal inputs and uses a single LSTM to learn the contextual information.

LMF [17] Low-rank Multimodal Fusion (LMF) is a method that leveraging low-rank weight tensors to make multimodal fusion efficient without compromising on performance. It not only drastically reduces computational complexity but also significantly improves performance.

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

MARN [34] Multi-attention Recurrent Network (MARN) uses the Multi-attention Block and the Long-short Term Hybrid Memory to discover the interactions between different modalities.

RMFN [16] Recurrent Multistage Fusion Network (RMFN) integrates the multistage fusion process with the recurrent neural networks to model temporal and intra-modal interactions.

MFM [27] Multimodal Factorization Model (MFM) can factorize the multimodal representations into multimodal discriminative factors and modality-specific generative factors, it can help each factor focus on learning from a subset of the joint information across multimodal data and labels.

MCTN [20] Multimodal Cyclic Translation Network (MCTN) is designed to learn robust joint representations by translating between different modalities, it can only use text modality data in the test process and create a new state-of-the-art result.

MulT [26] Multimodal Transformer (MulT) uses the directional pairwise crossmodal attention to interactions between multimodal sequences across distinct time steps and latently adapts streams from one modality to another, and it is the current state-of-the-art method on MOSI dataset.

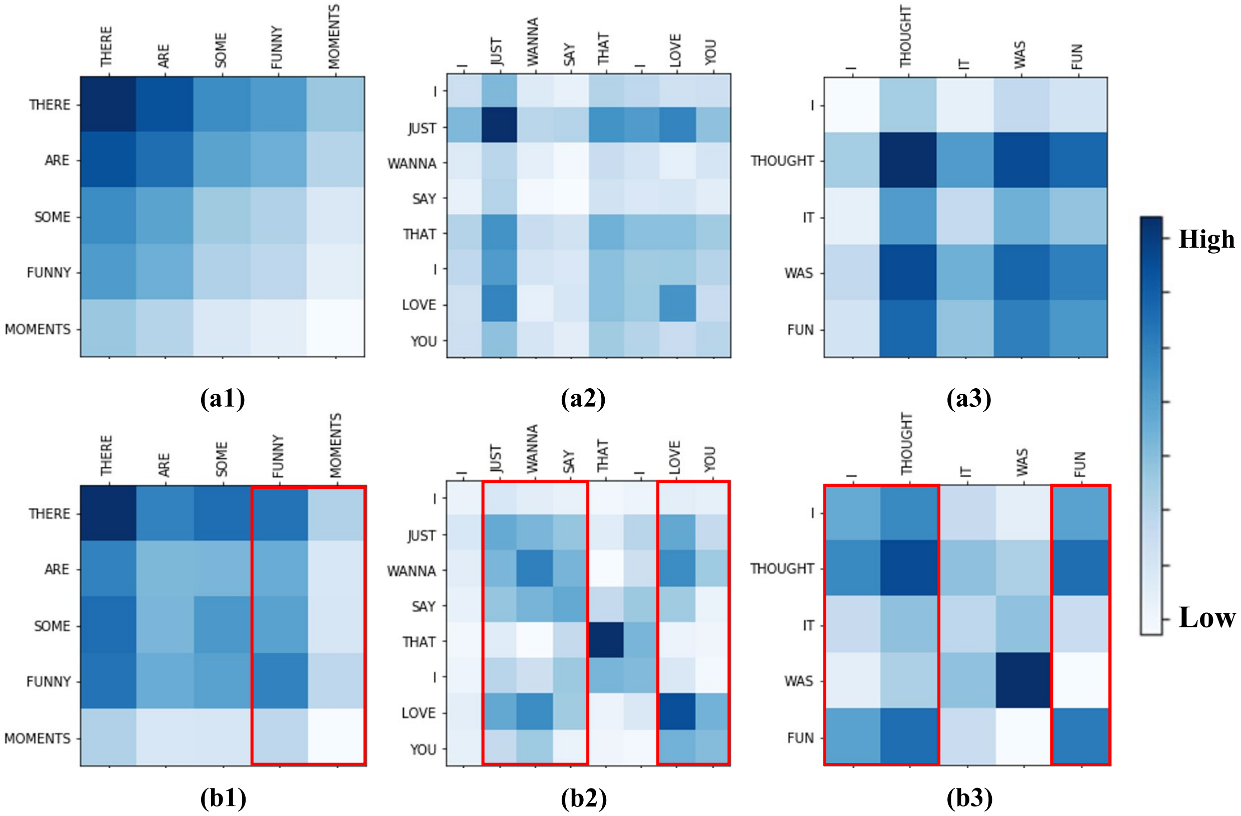
T-BERT [5] Bidirectional Encoder Representations from Transformers (BERT), which is fine-tuned only using text modality information.

5 RESULTS AND DISCUSSION

In this section, we present our experimental results and discuss the differences between our approach and previous works. In addition, we visualize the masked multimodal attention and discuss the change of the attention matrix after introducing audio modality information.

5.1 Comparison with Baseline

We evaluate the CM-BERT model on the CMU-MOSI dataset, Tabel 1 shows the experiment results. It is not difficult to see that the CM-BERT model creates a new state-of-the-art result on the MOSI dataset and it improves the performance on all the evaluation metrics. In the binary sentiment classification task, the CM-BERT model achieves 84.5% on Acc_2^h, which is about 1.5%-9.2% improvement compared with baselines. Similar to Acc_2^h, our model achieves a 1.7%-9.3% improvement on $F1$. In the sentiment score classification task, the improvement effect of the CM-BERT model is more obvious. Our model achieves 44.9% on Acc_7^h which is about 4.9 to 12.1 percentage points higher than the baselines. In the regression task, the CM-BERT reduces about 0.142-0.294 on MAE^l and improves about 0.093-0.183 on $Corr^h$. Notably, the p-value for student t-test between CM-BERT and T-BERT in Table 1 is far lower than 0.05 on all the metrics. What's more, all the baselines except T-BERT are using the information from text, audio, and video, but our model



(a1) (a2) (a3)

(b1) (b2) (b3)

Figure 4: Visualization of the attention matrices. (a1), (a2), and (a3) are the text attention matrices. (b1), (b2), and (b3) are the multimodal attention matrices. (a1) and (b1) are the attention matrices of the sentence "THERE ARE SOME FUNNY MOMENTS", (a2) and (b2) are the attention matrices of the sentence "I JUST WANNA SAY THAT I LOVE YOU", (a3) and (b3) are the attention matrices of the sentence "I THOUGHT IT WAS FUN". In these three examples, we use red boxes to emphasize the most important change in the weight of words.

only uses text and audio modality information to create a new state-of-the-art result.

It can be seen from the experimental results that the performance of the MulT model is obviously better than the other baselines. The main reason is that the MulT extends transformer to the multimodal setting and latently adapts elements across modalities via the attention. However, comparing the MulT model with the T-BERT model, because the latter can get better representations by fine-tuning the pre-trained BERT model, it gets better performance than the former. Different from the T-BERT model, the CM-BERT model we proposed extends the pre-trained BERT model from unimodal to multimodal, and it introduces the information of audio modality to help text modality effectively adjust the weight of words. Because the CM-BERT model can reflect the emotional state of the speaker more comprehensively and it can capture more sentiment characteristics through the interaction between text and audio modality, it significantly improves the performance on all the evaluation metrics.

We also perform experiments on the CMU-MOSEI dataset to prove the generalization of our method to other multimodal language datasets. For the convenience of comparison, following the previous work [24], we compare the Acc_2^h and $F1$ for the top 3 models in Table 1. Firstly, the MulT achieves 82.5% on Acc_2^h and 82.3% on $F1$. Compared with MulT, T-BERT shows better performance, and it achieves 83.0% on Acc_2^h and 82.7% on $F1$. What's more, CM-BERT achieves 83.6% on Acc_2^h and 83.6% on $F1$. Compared with the MulT and T-BERT, our model improvements about 0.6%-1.1% on Acc_2^h and 0.9%-1.3% on $F1$. Therefore, the superior performance on the CMU-MOSEI dataset also proves the generalization of our proposed method.

5.2 Visualization of the Masked Multimodal Attention

To prove the efficiency of the masked multimodal attention, we visualize the text attention matrix α_t and the multimodal attention matrix W_m respectively. By observing the difference in the weight of words, we can prove that after introducing audio modality information, the masked multimodal attention can adjust the word weight reasonably. We choice three sentences from the MOSI dataset as examples, the text attention matrices and the multimodal attention matrices of these sentences are shown in Figure 4. The color gradients represent the importance of words.

The first example is the sentence “THERE ARE SOME FUNNY MOMENT”, **(a1)** and **(b1)** are the corresponding attention matrices. It is obvious to see that there are many differences between **(a1)** and **(b1)**. For example, in **(a1)**, the word “FUNNY” gets a high attention score on the word “ARE”. However, it is meaningless, and we get nothing useful information from it. After introducing audio information, the masked multimodal attention reduces the score of “ARE”. In contrast, it distributes more attention on the words “SOME” and “MOMENTS”. The second example is the sentence “I JUST WANNA SAY THAT I LOVE YOU”, **(a2)** and **(b2)** are the corresponding attention matrices. It is not hard to see from **(a2)** and **(b2)** that the masked multimodal attention can improve the weight of related words and reduce the weight of irrelevant words. For example, in **(b2)**, the weight between the words “LOVE” and “YOU” has been improved and the weight between the words “JUST” and “THAT” has been reduced. These changes are in line with human logic. By giving more weight to relevant words, we can capture more rich emotional information and reduce the impact of noise information. The last example is the sentence “I THOUGHT IT WAS FUN”, the corresponding attention matrices are shown in **(a3)** and **(b3)**. Similar to the examples above, the weight of words in the sentence has been adjusted reasonably. For example, the weights between the word “I” and the words “THOUGHT”, “FUN” are all improved. Meanwhile, these words contain rich emotional information and it is important to correctly predict the sentiment of the speaker. From the above three examples, we can get the conclusion that the masked multimodal attention can adjust the weight of words reasonably, and it can capture the most important information through the interaction between text and audio modality.

6 CONCLUSION

In this paper, we propose a novel multimodal sentiment analysis model which is named Cross-Modal BERT (CM-BERT). Different from the previous works, we extend the pre-trained BERT model from unimodal to multimodal. We introduce the audio modality information to help text modality fine-tune BERT and to get better representations. As the core unit of CM-BERT, the masked multimodal attention is designed to dynamically adjust the weight of words through the inter-modality interaction between text and audio modality. The experiment results show that CM-BERT has significantly improved the performance on the CMU-MOSI and CMU-MOSEI datasets over previous baselines and text-only fine-tuning of BERT. Additionally, we visualize the attention matrices, which can clearly show the masked multimodal attention can adjust the word weight reasonably after introducing the audio modality. In fact, CM-BERT is also suitable for text and video modality, and it can be flexibly applied to more than two modalities. In the future, because most of the multimodal data in the real world are usually unaligned, we prefer to explore how to align different modalities data by using neural networks and how to use the pre-trained model to learn a better representation from unaligned multimodal data.

7 ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China, Grant No.61673235, 61772075, National Key R&D Program Projects of China, Grant No.2018YFC1707605.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal 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 Emotion 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 technologies. 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 pretraining 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.). International 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 Barnabás 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 analysis 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 generative 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 representations 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. Multimodal 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*.