Sense-aware BERT and Multi-task Fine-tuning for Multimodal Sentiment Analysis

Lingyong Fang[1], Gongshen Liu[1,*], Ru Zhang[2]
[1]*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China*
[2]*School of Cyberspace Security, Beijing University Posts and Telecommunications, Beijing, China*
Email: {fangly, lgshen}@sjtu.edu.cn, zhangru@bupt.edu.cn

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892116

***Abstract*—Humans convey emotions through verbal and non-verbal signals when communicating face-to-face. Pre-trained language model such as BERT can be fine-tuned to improve the performance of various downstream tasks including sentiment analysis. However, most prior works about BERT fine-tuning contains only textual unimodal data and lacks information from sense organs, such as audio and visual signals, which are crucial for sentiment analysis. In this paper, we propose Sense-aware BERT (SenBERT) which allows sense information integrated with BERT during fine-tuning. In particular, we exploit multimodal multi-head attention to capture the interaction between unaligned multimodal data. Additionally, due to the variable information richness of different modalities, multimodal network may be dominated by some modalities during training process, so we propose unimodal sentiment analysis auxiliary tasks for multi-task learning which forces the model to focus on all modalities. We conduct experiments on CMU-MOSI and CMU-MOSEI datasets for multimodal sentiment analysis. The results show the superior performance of SenBERT on all the metrics over previous baselines.**

***Index Terms*—multimodal sentiment analysis, pre-trained language model, multi-task learning**

I. Introduction

With the development of user-generated online content, a large amount of multimodal data has been generated, which is rewarding to exploit. Among them, sentiment analysis is an increasingly important area and has been widely applied in dialogue systems, video understanding, risk management and other fields. In particular, language modality contains extensive information, and many texture sentiment analysis models have been proposed in recent years and have achieved excellent results. Recently, Bidirectional Encoder Representations from Transformers (BERT) [1] has gained significant attention, which is designed to pre-train deep bidirectional representations. By fine-tuning on specific tasks, BERT achieves excellent performance in many downstream natural language processing tasks, including sentiment analysis task.

Despite the success of pre-trained language models, one of the limitation of existing language models is that only simple contextual features are used in both the representation and training objectives, with few explicit sense modality cues considered. During face-to-face communication, human convey information not only through verbal modalities, but also through sense such as visual and auditory modalities, and this abundance of information provides us with the benefit of understanding human behavior and intentions. A text-only approach cannot accurately determine the speaker's emotions, while the interaction between textual information and sense modalities can provide more comprehensive emotional information. Multimodal sentiment analysis (MSA) is now gaining widespread attention and is defined as a multimodal fusion problem [2], referring to the fusion of signals from different modalities into multimodal representations for downstream tasks, which improves the overall performance by providing inter-modal interactions. Therefore, it is best to perform sentiment analysis by introducing information from sense modalities, which offers more accurate emotional information.

Some previous work has only used BERT for word representation in multimodal learning [3], which is far from exploiting the performance of BERT. The output of pre-trained BERT model is high-level features containing rich information, while the features of sense modality are often obtained by preprocessing and are low-level, which make it difficult to design an effective interaction mechanism for language and sense modality. Some researchers have tried to enhance BERT with aligned sense modality data [4], [5], but it is not always feasible to align the multimodal data in realistic scenarios, and their methods are difficult to fully exploit long term dependencies across modalities.

Due to the variation in information richness of different modalities, the multimodal network branches of more information-rich modalities will converge quickly and the others will converge more slowly, which leads to the final network ignoring the knowledge of the less information-rich modalities. Especially in multimodal sentiment analysis, the language modality performs much better than the sense modality. Thus the performance can be improved by introducing unimodal sentiment prediction auxiliary tasks. Some researches introduce additional unimodal human annotations [6], which requires high labor costs. Unimodal labels can also be generated by self-supervised learning [7], but it brings instability to the model.

In this paper, we propose Sense-aware BERT (SenBERT) that is a fine-tuned BERT with explicit sense modality clues. The method first captures visual-audio interaction to generate sense modality representation vector via Low-level Cross-Attention Layer and Self-Attention Layer. After that High-level Cross-Attention Layer is used to capture the interaction between output of BERT and sense modality to generate a multimodal representation. Unimodal sentiment analysis tasks are introduced for multi-task learning to improve performance. We

*Corresponding author

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

use multimodal labels as unimodal labels, and weigh different loss functions by uncertainty of each task, which requires no additional annotation and does not bring much additional complexity to the model. To demonstrate the validity of the method, we evaluated it on the public multimodal sentiment analysis datasets CMU-MOSI [8] and CMU-MOSEI [9]. The experiments show that SenBERT improves the performance on all the metrics over previous baselines.

To sum up, the main contributions of our proposed work are three-fold:

- We propose Sense-aware BERT (SenBERT) that integrate sense modality information into BERT via multimodal multi-head attention and distinguish between the interaction of lower-level features and higher-level features
- We introduce unimodal sentiment analysis tasks for multi-task learning which forces the model to focus on all modalities during training and further improve the effectiveness of the model.
- Our methods obtain better results than the previous state-of-the-art works on public multimodal sentiment benchmark datasets CMU-MOSI and CMU-MOSEI.

II. Related Work

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis has become an important topic that integrates information from heterogeneous data such as language, visual and acoustic modalities in order to understand human emotions. Some previous works aligned different modality sequences based on word boundaries and then fused them based on the aligned sequences. Gu et al. [10] designed a hierarchical multimodal architecture with attention to perform word-level fusion. Wang et al. [11] proposed a recurrent attended variation embedding network to generate the multimodal-shifted word representation. Pham et al. [12] introduced a method of learning joint representations based on translation from a source to a target modality. However, manual word-alignment process requires additional labor costs and time costs, and neglect long term dependencies across modalities. Therefore, recent studies have focused on the fusion of unaligned sequence data. Tsai et al. [13] constructed Multimodal Transformer (MulT) that focuses on interactions between multimodal sequences spanning different time steps. Lv et al. [14] designed progressive reinforcement strategy and message hub to encourage a more efficient multimodal fusion. Han et al. [15] proposed bi-bimodal fusion network that pairwise fusion process proceeds progressively through stacked complementation layers.

Our work focuses on multimodal fusion with unaligned sequences. Although previous works have proposed reasonable multimodal fusion mechanisms, they have bottlenecks in unimodality, especially in language modality where most studies either use traditional word representation model such as Glove [16] or only regard BERT [1] as a feature extractor [3], which leads to some experiments [4] showing that using BERT based unimodal model performs better than multimodal models. Our methods are based on BERT and then integrate other modality information into it.

B. Pre-trained Language Model and Fine-tuning

Pre-trained language models (PLMs) have been widely applied in natural language processing, and have significantly improved the state of the art across various natural language processing tasks [1], [17], [18]. Peters et al. [19] proposed Embeddings from Language Models (ELMo) which learns deep context-dependent representations on a large text corpus. Bidirectional Encoder Representations from Transformers (BERT) [1] performs better than ELMo, because it employs a bidirectional Transformer encoder to fuse both the left and right context, and is pre-trained on Masked Language Model Task and Next Sentence Prediction Task via a large cross-domain corpus. There are some PLMs that have been proposed to improve the BERT model, such as RoBERTa [20] and ALBERT [21].

Fine-tuning the pre-trained BERT model has been a key factor in improving the performance of downstream tasks [22]. A new trend in recent years is fine-tuning the pre-trained BERT model with external resources, such as knowledge graphs [23], [24], semantic role label [25] and characters [26]. There is currently some work on fine-tuning BERT for multimodal sentiment analysis. Yang et al. [5] utilized masked multimodal attention which capture the interaction between text and audio modality to fine-tune BERT. Rahman et al. [4] employed multimodal adaptation gate which enables BERT to accept multimodal nonverbal data during fine-tuning. However, their models requires the multimodal sequence data to be aligned, which is not always be feasible in practice and ignore the long-range dependencies between elements from different modalities. In this paper, We explore fine-tuning BERT based on unaligned multimodal sequence data.

C. Multi-task Learning

Multi-task learning aims to utilize the useful information contained in multiple related tasks to help improve the generalization performance of all tasks [27]. Some previous work applied multi-task learning to multimodal sentiment analysis. Akhtar et al. [28] proposed a multi-task framework that performs sentiment and emotion analysis both. Yu et al. [7] conducted Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) that perform unimodal sentiment analysis with a label generation module that acquires independent unimodal supervisions. Different from Self-MM, Different from Self-MM, our method performs unimodal sentiment analysis based on multimodal labels.

III. Methods

As illustrated in Fig. 1, our SenBERT model comprises of three parts: (1) a **Sense Block** that capture audio-visual interaction to generate sense modality representation vector and perform unimodal sentiment analysis auxiliary task of sense modality; (2) a **Language Block** that use BERT to get contextual text feature and perform unimodal sentiment

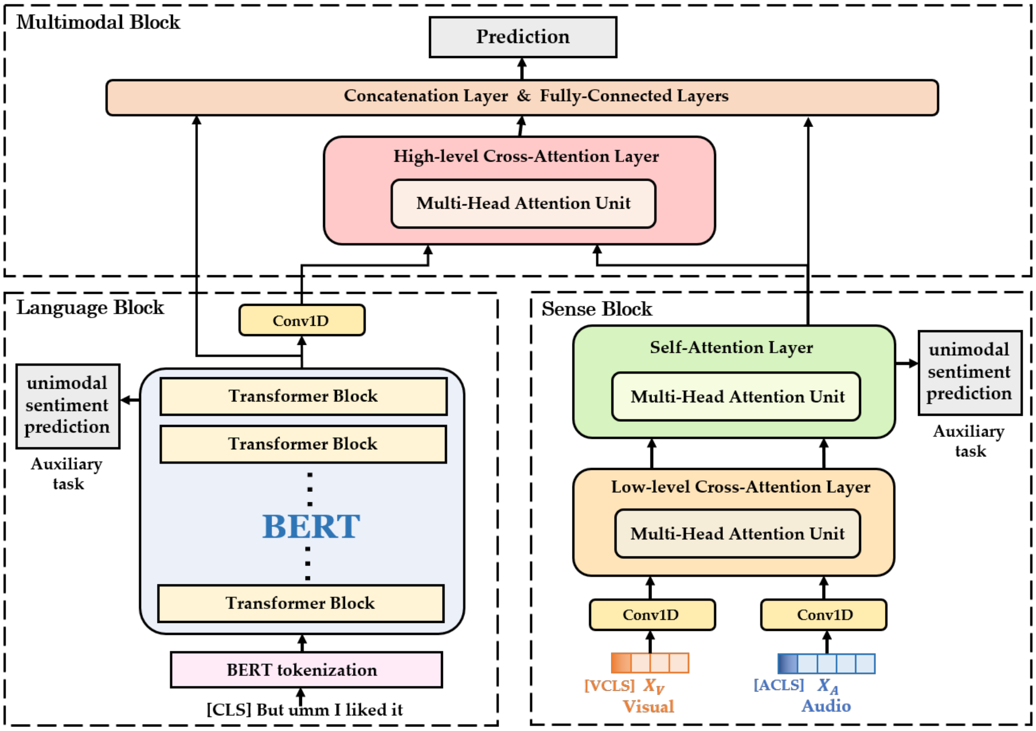


Fig. 1. Overview architecture of the Sense-aware BERT model. The details of the attention layers in SenBERT is shown in Fig. 2

analysis auxiliary task of language modality; (3) a **Multimodal Block** that integrate sense modality information into BERT and obtain the final multimodal representation to perform multimodal sentiment analysis task.

A. Problem Definition

In this work, we employ a pre-trained BERT from language (L) modality with the text sequence of word-piece tokens $T = \{T_1, T_2, \ldots, T_n\}$ as input, where n indicates the length of the sequence. At the same time, the other two modalities corresponding to this text is provided: vision (V) and acoustic (A), which are extracted from the video clip, and we uniformly refer to them as sense (S) modality denoted by $X_{\{V,A\}} \in \mathbb{R}^{T_{\{V,A\}} \times d_{\{V,A\}}}$, where $T_{(\cdot)}$ and $d_{(\cdot)}$ represent the sequence length and feature dimension. Our goal is to integrate unaligned data of sense modality in BERT and obtain representations that are effective for sentiment analysis tasks.

B. Audio-Visual Interaction

Following previous work about multimodal fusion [13], we use attention mechanism for unaligned audio-visual interaction. Since the visual and auditory features have different dimensions, we use a 1D temporal convolutional layer to convert them to the same size for the dot-products in the subsequent cross-attention layer. After that we argument position embedding (PE) to gain temporal information:

$$\hat{X}_{\{V,A\}} = \text{Conv1D}(X_{\{V,A\}}, k_{\{V,A\}}), \tag{1}$$

$$Y_{\{V,A\}} = \hat{X}_{\{V,A\}} + \text{PE}(T_{\{V,A\}}, d). \tag{2}$$

Motivated by the [CLS] inside BERT [1], we created [VCLS] for visual modality as well as [ACLS] for audio modality and initialize them by applying Mean-pooling layers on $Y_{\{V,A\}}$:

$$Z_V^{[0]} = [VCLS] \oplus Y_V, \tag{3}$$

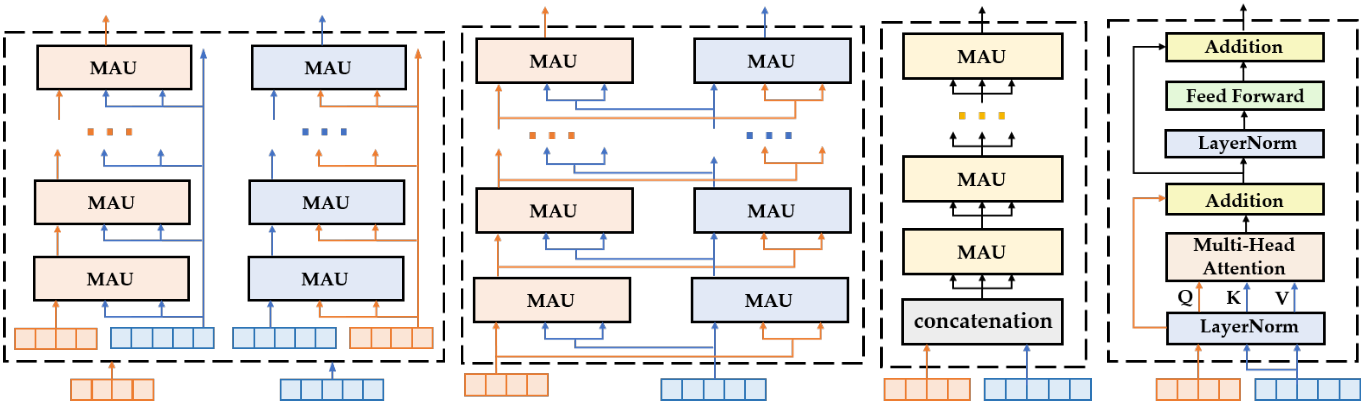
$$Z_A^{[0]} = [ACLS] \oplus Y_A, \tag{4}$$

where $\oplus$ represents concatenation operator. [ACLS] and [VCLS] can capture global information for sentiment analysis in subsequent attention interactions.

Multi-Head Attention Unit. Multi-Head Attention Unit (MAU) is the basic component of the attention layers of our model, and is designed based on the original transformer encoder layer [29]. As shown in Fig. 2(d), MAU contains multi-head cross-attention, residual connection [30] and layer normalization [31]. For the sake of simplicity, we use two vectors X_α, X_β to represent the sequence of different modalities, where α and β represent two modalities, where $\alpha, \beta \in \{L, V, A, S\}$, and denotes visual (V) and audio (A) in this section. Multi-head Attention (MA) is computed as:

$$\begin{aligned} MC(X_\alpha, X_\beta, X_\beta) &= Concat(head_1, \cdots, head_n)W_o \\ head_i &= \text{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right) V_\beta \\ &= \text{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta}. \end{aligned} \tag{5}$$

Querys, Keys and Values are defined as $Q_\alpha = X_\alpha W_{Q_\alpha}, K_\beta = X_\beta W_{K_\beta}, V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in$



(a) Low-level Cross-Attention Layer

(b) High-level Cross-Attention Layer

(c) Self-Attention Layer

(d) Multi-head Attention Unit (MAU)

Fig. 2. The details of the attention layers in SenBERT.

$\mathbb{R}^{d_\alpha \times d_k}, W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}, W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ and $W_o \in \mathbb{R}^{hd_v \times hd_v}$ are the parameters of linear projection. Modality α will update its sequence based on the information from modality β.

Low-level Cross-Attention Layer. We exploit the temporal relationship between unaligned visual and speech sequences based on attention mechanism. As shown in Fig. 2(a), Low-level Cross-Attention Layer contains some Multi-Head Attention Units. Unlike language modality, the representation of visual and auditory modality is obtained by feature extraction, which is low-level features. Each modality is reinforced with low-level features from another modality, allowing the model to preserve low-level information about the modality for higher performance, as demonstrated in previous work [13]. Z_α will continue to be reinforced by $Z_\beta^{[0]}$:

$$Z_\alpha^{[i]} = MAU(Z_\alpha^{[i-1]}, Z_\beta^{[0]}, Z_\beta^{[0]}) \tag{6}$$

where $\alpha, \beta \in \{V, A\}$, MAU represents complete Multi-Head Attention Unit. The final output of Low-level Cross Attention Layer is denoted by Z_V^{CA} and Z_A^{CA}.

Self-Attention Layer. In order to facilitate further cross-modal interaction to obtain a more uniform sense vector representation, we concatenate the output of the Low-level Cross-Attention Layer into $Z_S^{CA} = [Z_V^{CA}; Z_A^{CA}]$, then pass it to the self-attention transformer [29] for processing to generate Z_S^{SA}. The calculation process of MAU is the same as in Low-level Cross-Attention Layer, with the difference that $\alpha = \beta = S$. The architecture of Self-Attention Layer is shown in Fig. 2(c). This layer will update each element of sense modality based on the information of all other elements. Note that self-attention transformer does not change the shape of the input, and Z_S^{SA} can still be split into two parts: $Z_S^{SA} = [Z_V^{SA}; Z_A^{SA}]$.

C. Sense-aware BERT

The text sequence is first passed through the pre-trained BERT model, and the output of the last encoder layer is treated as the language feature, denoted by $Z_L = [[CLS], T_1, T_2, \cdots, T_n]$, where n represents the sequence length. $[CLS]$ incorporates global information about the sequence and will be used for subsequent unimodal sentiment analysis in multi-task learning. We similarly unify the dimension of the output BERT with the sense representation vector by a 1D temporal convolutional layer. After that, the information of sense modality is integrated into BERT through High-level Cross-Attention Layer.

High-level Cross-Attention Layer. Unlike the visual and audio modalities, the features of the language modality are informative after BERT and are considered high-level features. The interaction method in Low-level Cross-Attention Layer may cause the high-level features to not receive a clear supervisory signal for updates, resulting in poor performance. As shown in Fig. 2(b), we employ progressive reinforcement strategy for language-sense interaction, where the two modalities progressively reinforce each other:

$$\begin{aligned} Z_\alpha^{[i]} &= MAU(Z_\alpha^{[i-1]}, Z_\beta^{[i-1]}, Z_\beta^{[i-1]}) \\ Z_\beta^{[i]} &= MAU(Z_\beta^{[i-1]}, Z_\alpha^{[i-1]}, Z_\alpha^{[i-1]}) \end{aligned} \tag{7}$$

where $\alpha, \beta \in \{L, S\}$, MAU represents complete Multi-Head Attention Unit. Note that the text sequence and the sense sequence pass through an additional layer of self-attention before entering the high-level cross-Attention layer, which reduces the conflict between unimodal and multimodal tasks. In other word, the features used for the unimodal task will be adapted to the subsequent multimodal task with one more layer of transformer block, which is beneficial for multi-task learning.

D. Multi-task Fine-tuning

In addition to the multimodal sentiment analysis task, we introduce unimodal sentiment analysis task to force the training process to focus on all modalities and make the unimodal network perform better.

Multimodal Task. The final multimodal representation $[CLS_{final}]$ is obtained by concatenating

$[CLS_{HCA}], [VCLS_{HCA}], [ACLS_{HCA}]$ in the multimodal block, $[CLS]$ in the language block, and $[SCLS]$ in the sense block.

After that we pass $[CLS_{final}]$ through a linear layer for multimodal sentiment prediction, which is a regression task. We use L1 Loss as the optimization objective, and the loss funtion is denoted by $\mathcal{L}_{mul}$.

Unimodal Task. For language modalities, we use the output of BERT: $[CLS_{uni}]$. For sense modality, we use the output of Self-Attention Layer: $[SCLS_{uni}] = [VCLS_{uni}, ACLS_{uni}]$. The above unimodal representations will be fed to a fully connected layer for unimodal sentiment analysis. We use L1 Loss as the optimization objective, and the loss funciton of unimodal tasks is denoted by $\mathcal{L}_{lan}$ and $\mathcal{L}_{sen}$ respectively.

Due to the variability of sentiment labeling across modalities [6], unimodal sentiment analysis tasks may introduce noise if they also use the same labels as multimodal ones, and different tasks contain different levels of noise. To alleviate this problem, I use an uncertainty-based multi-task loss function [32], [33], which gives higher weights to tasks with lower uncertainty through the idea of probabilistic modeling.

We combine the two tasks to obtain the final optimization objective:

$$\begin{aligned}\mathcal{L} =& \frac{1}{2\sigma_{mul}^2}\mathcal{L}_{mul} + \frac{1}{2\sigma_{lan}^2}\mathcal{L}_{lan} + \frac{1}{2\sigma_{sen}^2}\mathcal{L}_{sen} \\ &+ \ln(1+\sigma_{mul}^2) + \ln(1+\sigma_{lan}^2) + \ln(1+\sigma_{sen}^2) \qquad (8)\end{aligned}$$

where σ_{mul}, σ_{lan} and σ_{sen} are noise parameters for each task. If the σ of a task is larger, it means that the task is noisier and has higher uncertainty, so the model will give a lower weight to this task in loss function. The last three items act as a regulariser, making the noise not to increase much.

IV. Experiments

A. Datasets and Evaluation Metrics

We use CMU-MOSI [8] and CMU-MOSEI [9] to evaluate our proposed model. CMU-MOSI is a prevalent multimodal sentiment analysis dataset, consisting of 2,199 utterance-video segments sliced from 93 videos in which 89 distinct narrators are sharing opinions on some topics. Each segment is manually annotated with a real number score ranged from -3 to +3, indicating the relative strength of negative (score below zero) or positive (score above zero) emotion. CMU-MOSEI is the extension of CMU-MOSI, The dataset contains 23,453 video segments which are extracted from 5,000 videos involving 1,000 distinct speakers and 250 different topics, Its labeling style is the same as CMU-MOSI.

In our experiments, following the previous works [3], [7], [13], we employ four metrics to evaluate the performance of the baselines and proposed model. For binary sentiment classification task, we report binary classification accuracy (Acc-2) and weighted F1 score (F1-Score). For regression task, we report mean absolute error (MAE) and Pearson correlation (Corr).

B. Baselines

To evaluate the rationality and effectiveness of our methods, We compare proposed model with the following recent and competitive baselines:

- **TFN** [34]: Tensor Fusion Network explicitly represents unimodal, bimodal, and trimodal interactions between behaviors by three-fold Cartesian product and outer product.
- **LMF** [35]: Low-rank Multimodal Fusion decomposes high-order tensors into many low-rank factors to improve effificiency, then performs multimodal fusion based on these factors.
- **CIA** [36]: Context-aware Interactive Attention learns the inter-modal interaction among the participating modalities through an auto-encoder mechanism.
- **MISA** [3]: Modality-Invariant and -Specific Representations projects each modality into modality-invariant subspace and modality-specific subspace to provide a holistic view of the multimodal data.
- **ICCN** [37]: Interaction Canonical Correlation Network use deep canonical correlation analysis to learn correlations between different modalities.
- **MulT** [13]: Multimodal Transformer uses directional pairwise cross-modal attention to translate one modality to another, which captures interactions between multimodal sequences across distinct time steps.
- **PMR** [14]: Progressive Modality Reinforcement is an upgraded version of MulT. The model allows the message hub and each modality progressively reinforce each other via cross attention to obtain more effective multimodal representations.
- **CM-BERT** [5]: Cross-Modal BERT introduce masked multimodal attention which capture the interaction between text and audio modality to fine-tune the pre-trained BERT model.

C. Implementation Details

To extract low-level feature of visual modality, video frames are processed by Facet [38] to generate a set of features that are composed of 35 facial action units, which represent the facial muscle movement, including facial landmarks, head pose, etc. To extract low-level feature of audio modality, COVAREP [39] is utilized for generating features of acoustic signals, includes 12 Mel-frequency cepstral coeffificients (MFCCs), pitch tracking, speech polarity, spectral envelope, etc.

We use uncased BERT-Base as the pre-trained BERT in our proposed SenBERT model. The rest of the parameters are initialized randomly. Low-level Cross-Attention Layer and High-level Cross-Attention Layer both have 2×4 attention blocks and 8 attention heads. Self-Attention Layer has 4 attention blocks and 8 attention heads. We train each module with dropouts of 0.3. We use Adam [40] as optimizer and use a linear decay learning rate schedule with warm-up. To get better performance, the learning rate is 5e-5 for BERT and 1e-3 for other parameters. The batch size is 32 across two datasets. The hyper-parameters are determined according to the performance from the validation set.

TABLE I
COMPARISON WITH BASELINES ON CMU-MOSI AND CMU-MOSEI BENCHMARK

Model	MOSI				MOSEI				Data Setting
	MAE	Corr	Acc-2	F1-Score	MAE	Corr	Acc-2	F1-Score	
TFN*	0.901	0.698	-/80.8	-/80.7	0.593	0.700	-/82.5	-/82.1	Unaligned
LMF*	0.917	0.695	-/82.5	-/82.4	0.623	0.677	-/82.0	-/82.1	Unaligned
CIA*	0.914	0.689	79.8/-	79.5/-	0.680	0.590	80.4/-	78.2/-	Aligned
ICCN*	0.860	0.710	-/83.0	-/83.0	0.565	0.713	-/84.2	-/84.2	Aligned
MISA*	0.804	0.764	80.79/82.1	80.77/82.03	0.568	0.724	82.59/84.23	82.67/83.97	Aligned
MulT*	0.871	0.698	-/83.0	-/82.8	0.580	0.703	-/82.5	-/82.3	Unaligned
PMR†	-	-	-/83.6	-/83.4	-	-	-/82.4	-/82.1	Unaligned
CM-BERT‡	0.729	0.791	-/84.5	-/84.5	-	-	-/83.6	-/83.6	Aligned
SenBERT (Ours)	**0.702**	**0.805**	**83.67/85.37**	**83.66/85.40**	**0.534**	**0.768**	**84.57/85.39**	**84.59/85.15**	Unaligned

*: from [3]; †: from [14]; ‡: from [5]. For Acc-2 and F1-Score, we use the segmentation marker -/- to report results, where the the left-side score is calculated as ”negative/non-negative”, while the right-side score is calculated as ”negative/positive”

V. RESULTS AND ANALYSIS

A. Comparison with Baselines

We list the results with baselines on the two datasets in Table I. As for ”Data Setting”, we divide it into two categories: Unaligned and Aligned. Aligned setting requires an additional step of manually aligning the data of different modalities according to word boundaries, while unaligned setting directly uses unaligned sequence data for multimodal fusion. Performance is generally better in aligned settings. It can be observed that the proposed SenBERT model achieves the best performance and outperforms other models in all the evaluation metrics across the CMU-MOSI dataset and CMU-MOSEI dataset. The results demonstrate the superiority of our proposed model, showing the effectiveness of integrating sense modality into BERT during fine-tuning. It is notable that our model with unaligned setting performs superior to all the model with aligned setting, which is an encouraging result as we are able to perform better even with less labor cost and time cost.

B. Ablation Study

To further explore the contributions of SenBERT, we conduct comprehensive ablation studies using the unaligned version of CMU-MOSI. The results are shown in Table II.

Role of Modalities. We first explore the effect of different modalities on our model performance. We examine performance of the model with language-only modality and sense-only modality. For the language-only model, we directly use $[CLS]$ of BERT output for sentiment prediction. For the sense-only model, we employ the output of the Self-Attention Layer in the Sense Block for the task. From the experimental results, it can be seen that removing language modality or sense modality brings a degradation in model performance. This proves the need for a multimodal perspective on sentiment analysis and the necessity to integrate sense information in BERT during fine-tuning. It is also observed that the language modality itself can achieve excellent performance, significantly stronger than the sense modality, mainly due to the benefits of pre-training.

TABLE II
ABLATION STUDIES ON CMU-MOSI DATASET.

Ablation	MAE	Corr	Acc-2	F1-Score
Role of Modalities				
Language Only	0.741	0.762	82.65/84.16	82.64/84.19
Sense Only	0.876	0.657	76.87/77.66	76.86/77.60
Role of Auxiliary Tasks				
W/O $\mathcal{L}_{lan}$	0.732	0.771	82.94/84.60	82.90/84.61
W/O $\mathcal{L}_{sen}$	0.743	0.759	82.36/84.15	82.32/84.16
Role of Fine-tuning				
Fixed	0.764	0.729	81.92/83.08	81.95/83.16
Random	1.139	0.508	69.53/69.97	69.61/70.15
Full Model, All Modalities	**0.702**	**0.805**	**83.67/85.37**	**83.66/85.40**

Role of Auxiliary Tasks. We also perform ablation study on the design of multi-task learning. It can be observed that the performance degrades without auxiliary tasks. It shows that multi-task learning is important to improve the performance of multimodal sentiment analysis. In particular, the unimodal task of sense modality is more effective in our task. We believe that this is because sense modality has poorer performance and is less convergent than language modality, while introducing the unimodal sentiment analysis task of sense modality can encourage the model training process to take into account the sense modality instead of the language modality alone, thus making the sense modality better complementary to provide more comprehensive sentiment information.

Role of Fine-tuning. Lastly, we examine the effect of fine-tuning strategy. Here, we consider two methods of applying BERT: fixed pre-trained parameters and random initialization. It can be seen that the performance drops without fine-tuning strategy. With fixed parameters, BERT is simply treated as a feature extractor, making it difficult to adapt the model to new data distributions and multimodal interaction scenarios. With random initialization, BERT of such magnitude cannot work due to the lack of long-term pre-training on large amounts of data. This observation clearly demonstrates the necessity of fine-tuning BERT in multimodal sentiment analysis.

VI. Conclusion

In this paper, we introduced the Sense-aware BERT (SenBERT) for multimodal sentiment analysis. Different from previous works, we integrate visual and audio modalities into the pre-trained BERT model rather than only used text information during fine-tuning. We employ low-level and high-level cross-attention layer to capture the interaction between different modalities. Additionally, Unimodal sentiment analysis task is used for multi-task learning to further enhance performance. Our experiments demonstrated the superior performance of SenBERT on the CMU-MOSI and CMU-MOSEI datasets over previous baselines. Ablation studies were performed to further study the influence of the individual components in SenBERT. In fact, our methods not only enable the fusion of sense modality information, but also provide a framework for the fusion of other heterogeneous information in BERT. Moreover, the pre-processed features of visual and audio modality limit the performance. In the future, we will conduct an end-to-end network for multimodal sentiment analysis.

Acknowledgment

This research work has been funded by the Joint Funds of the National Natural Science Foundation of China (Grant No. U21B2020).

References

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

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[3] D. Hazarika, R. Zimmermann, and S. Poria, “Misa: Modality-invariant and-specific representations for multimodal sentiment analysis,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[4] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, “Integrating multimodal information in large pretrained transformers,” in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020. NIH Public Access, 2020, p. 2359.

[5] K. Yang, H. Xu, and K. Gao, “Cm-bert: Cross-modal bert for text-audio sentiment analysis,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 521–528.

[6] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, “Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3718–3727.

[7] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis,” *arXiv preprint arXiv:2102.04830*, 2021.

[8] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[9] A. Zadeh and P. Pu, “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 (Long Papers)*, 2018.

[10] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, “Multimodal affective analysis using hierarchical attention strategy with word-level alignment,” in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2225.

[11] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.

[12] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, “Found in translation: Learning robust joint representations by cyclic translations between modalities,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.

[13] 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 *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[14] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, “Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2554–2562.

[15] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, and S. Poria, “Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis,” in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 6–15.

[16] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[17] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli, “Cloze-driven pretraining of self-attention networks,” *arXiv preprint arXiv:1903.07785*, 2019.

[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.

[19] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations. arxiv 2018,” *arXiv preprint arXiv:1802.05365*, vol. 12, 1802.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.

[22] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?” in *China national conference on Chinese computational linguistics*. Springer, 2019, pp. 194–206.

[23] A. Yang, Q. Wang, J. Liu, K. Liu, Y. Lyu, H. Wu, Q. She, and S. Li, “Enhancing pre-trained language representations with rich knowledge for machine reading comprehension,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2346–2357.

[24] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, “K-bert: Enabling language representation with knowledge graph,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2901–2908.

[25] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, “Semantics-aware bert for language understanding,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9628–9635.

[26] W. Ma, Y. Cui, C. Si, T. Liu, S. Wang, and G. Hu, “Charbert: Character-aware pre-trained language model,” *arXiv preprint arXiv:2011.01513*, 2020.

[27] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[28] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," *arXiv preprint arXiv:1905.05812*, 2019.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[32] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, 2018.

[34] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[35] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[36] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5647–5657.

[37] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.

[38] iMotions, "Facial expression analysis," 2017.

[39] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.