IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 15 January 2024, accepted 1 February 2024, date of publication 6 February 2024, date of current version 13 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3363028

RESEARCH ARTICLE

SentDep: Pioneering Fusion-Centric Multimodal Sentiment Analysis for Unprecedented Performance and Insights

CHONG LU[1] AND XUFENG FU[2], (Member, IEEE)
[1]School of Information Management, Xinjiang University of Finance and Economics, Ürümqi 830026, China
[2]The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen 518107, China

Corresponding authors: Chong Lu (498841300@qq.com) and Xufeng Fu (kllj870825@163.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62166039, and in part by the Xinjiang University of Finance and Economics under Grant 2022XGC075.

ABSTRACT Multimodal sentiment analysis (MSA) is an emerging field focused on interpreting complex human emotions and expressions by integrating various data types, including text, audio, and visuals. Addressing the challenges in this area, we introduce SentDep, a groundbreaking framework that merges cutting-edge fusion methods with modern deep learning structures. Designed to effectively blend the unique features of textual, acoustic, and visual data, SentDep offers a unified and potent representation of multimodal data. Our extensive tests on renowned datasets like CMU-MOSI and CMU-MOSEI demonstrate that SentDep surpasses current leading models, setting a new standard in MSA performance. We conducted thorough ablation studies and supplementary experiments to identify what drives SentDep's success. These studies highlight the importance of the size of pre-training data, the effectiveness of various fusion techniques, and the critical role of temporal information in enhancing the model's capabilities.

INDEX TERMS SentDep, multimodal fusion, temporal information, benchmark datasets.

I. INTRODUCTION

Multimodal data has emerged as a crucial medium of communication in the digital age, notably propelled by the ubiquity of social media platforms. In this realm, deriving insightful inferences from multimodal data - encompassing textual, acoustic, and visual modalities - regarding human psychological states such as sentiment tendencies and depression levels, has garnered significant importance. The inherent heterogeneity within multimodal data, characterized by distinct data structures across different modalities, necessitates innovative approaches for effective feature extraction and fusion.

Historically, the Tensor Fusion Network (TFN) introduced by Zadeh et al. [1], pioneered the use of Cartesian product to blend features across modalities, marking a significant stride towards addressing the challenges posed by multimodal data. Subsequent efforts embarked on exploring the bidirectional relationships and complementary information residing among different modalities, leveraging attention mechanisms to compute coattention across modality pairs, such as textual and acoustic modalities [2], [3]. The dawn of Transformer-based structures [4] further enriched the realm of multimodal feature processing, with endeavors employing self-attention mechanism to facilitate modality interactions [5], [6], [7], [8], [9], [10].

However, the high computational overhead associated with Transformer architectures, predominantly due to the self-attention mechanism, presents a significant impediment, particularly in scenarios demanding real-time processing. This bottleneck has spurred interest in structures predominantly built upon multilayer perceptrons (MLPs), which offer a promising alternative by circumventing the computational intricacies of self-attention, while retaining competitive performance [11], [12].

Motivated by the potential of Clip [13], we introduce a novel framework, referred to as Multimodal Fusion Network (MFN), designed meticulously to process multimodal features

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

and predict sentiment tendencies or depression levels from human utterances in videos. The cornerstone of MFN lies in its robust multimodal representation learning and fusion strategy, facilitated through an adapted version of the CLIP model and a dedicated acoustic processing unit for the textual, visual, and acoustic modalities respectively. The formulated multimodal tensor, embodying the synergized representation across modalities, is subjected to a fusion module comprising a series of transformation blocks. This structured approach ensures a harmonized multimodal representation, which is subsequently channeled through a classifier for sentiment analysis or depression detection.

Our salient contributions encapsulate the inception of SentDep, offering a streamlined yet effective framework for multimodal feature processing. The meticulous design of SentDep, encompassing distinct processing units for each modality alongside a fusion module, ensures a robust multimodal representation conducive for accurate sentiment analysis and depression detection. Through extensive experiments on two pivotal mind state estimation tasks, we validate the efficacy of SenDep, which demonstrates favorable competitiveness with state-of-the-art approaches, particularly showcasing substantial progress in depression detection.

Our SentDep methodology yields three notable contributions to the field of multimodal sentiment analysis and depression detection:

1) **Robust Multimodal Representation Learning:** Through SentDep's utilization of an adapted CLIP model alongside a dedicated acoustic processing unit, a comprehensive understanding of multimodal data is achieved, ensuring robust representation learning across textual, acoustic, and visual modalities.
2) **Harmonized Multimodal Fusion:** SentDep's fusion module, with its series of transformation blocks, adeptly harmonizes representations across modalities, enhancing inter-modality interactions and yielding a robust multimodal representation pivotal for accurate sentiment or depression level prediction.
3) **Effective Prediction Mechanism:** The classifier within SentDep's prediction module efficiently maps the harmonized multimodal features to sentiment tendencies or depression levels, showcasing the efficacy of SentDep in addressing real-world sentiment analysis and depression detection tasks.

II. RELATED WORKS

A. MULTIMODAL SENTIMENT ANALYSIS

Multimodal Sentiment Analysis (MSA) aims to decipher sentiment tendencies from an individual's facial expressions(v), acoustic tones(a), and verbal expressions(t) within each utterance. Initially, Zadeh et al. introduced the Tensor Fusion Network (TFN) [1] and subsequently the Memory Fusion Network (MFN) [14], pioneering the fusion of multimodal features on a sequential level. The subsequent discourse in this domain predominantly revolved around exploring correlations amongst the involved modalities. For instance, Chen and Li [15] unveiled the Sentimental Words Aware Fusion Network (SWAFN) to compute the coattention between text and other modalities. Deng et al. [16] ventured into a deep dense fusion network armed with multimodal residual (DFMR) to amalgamate multimodal information in a paired configuration.

The advent of the Transformer model [4] and its trailblazing success in natural language processing and computer vision galvanized researchers to harness its self-attention mechanism for modality interactions in MSA. Illustratively, Delbrouck et al. [5] orchestrated a Transformer-based joint-encoding (TBJE) that assimilates acoustic and textual features, forging a joint encoding of these two modalities. Tsai et al. [17] introduced a Multimodal Transformer (MulT) and explored cross-modal attention between paired modalities (e.g., t with a).

B. MULTIMODAL DEPRESSION DETECTION

Unlike MSA, multimodal depression detection necessitates analyzing extended time sequences as it seeks to deduce a persistent long-term characteristic from individuals. Joshi et al. [18] employed a bag-of-words model to encode acoustic and visual features, which were subsequently fused using principal component analysis (PCA) and support vector mechanisms (SVM). Rodrigues Makiuchi et al. [19] leveraged audio-translated texts alongside hidden embeddings extracted from a pretrained BERT [20] model, utilizing CNNs to garner cross-modality information.

In a distinct vein, Kaya et al. [21] innovated a new Automatic Speech Recognizer (ASR) transcription based features, while Ray et al. [22] devised a multi-layer attention network for estimating depressions. Extending beyond acoustic, visual, and textual features, Kroenke and Spitzer [23] demonstrated the salient contribution of body gestures towards enhancing the accuracy of depression estimation. Sun et al. [24] employed a Transformer model for multimodal feature extraction and conceived an adaptive late fusion scheme for final predictions, while Zhao et al. [25] proposed a hybrid feature extraction architecture blending self-attention and 3D convolutions for different kinds of features.

C. CLIP MODEL

The Contrastive Language-Image Pre-training (CLIP) model [13], developed by OpenAI, has emerged as a remarkable paradigm that bridges the realms of vision and language. It is trained by learning to predict the correspondence between a collection of images and texts across multiple data modalities. By creating a shared representation space for both images and text, the CLIP model is adept at transferring knowledge acquired during pretraining to a diverse range of downstream tasks, without the necessity for additional training data or task-specific model modifications. This model exemplifies a significant stride towards achieving models that understand and process multimodal data with minimal supervision. Its architecture and pretraining methodology offer a promising avenue for exploring how multimodal data can be effectively harnessed for a broad spectrum of applications, including but not limited to, sentiment analysis, object recognition, and natural language understanding.

III. BACKGROUND

Multimodal sentiment analysis (MSA) has emerged as a pivotal approach in understanding human emotions and expressions through the integration of multiple data modalities. At its core, MSA aims to analyze and interpret sentiments by leveraging the synergistic potential of textual, acoustic, and visual data. This integration allows for a more nuanced and comprehensive understanding of human sentiments compared to unimodal analysis [26].

Multimodal Data Fusion A critical aspect of MSA is the fusion of data from different modalities. Fusion techniques can be categorized broadly into three types: early fusion, late fusion, and hybrid fusion. Early fusion combines features at the data level, late fusion at the decision level, and hybrid fusion incorporates both strategies [27]. The choice of fusion technique significantly impacts the effectiveness of sentiment analysis, as it determines how the modalities interact and complement each other.

Deep Learning in MSA The advent of deep learning has revolutionized MSA by enabling the extraction of complex, high-level features from multimodal data [28]. Neural networks, particularly those employing architectures like Convolutional Neural Networks (CNNs) for visual data and Recurrent Neural Networks (RNNs) for textual and acoustic data, have shown substantial promise in enhancing the accuracy of sentiment analysis.

Temporal Dynamics Understanding the temporal dynamics within multimodal data is crucial for accurate sentiment analysis. Temporal information, capturing the evolution of sentiments over time, plays a significant role in contextualizing and interpreting the data more effectively. Techniques that can dynamically adapt to the temporal granularity of data are essential in MSA for capturing the true essence of human emotions [29].

Challenges in MSA Despite its advancements, MSA faces challenges like the need for large-scale pre-training data and the complexity of integrating diverse data modalities. Addressing these challenges is crucial for the development of more accurate and efficient sentiment analysis models [30].

IV. METHODOLOGY

This section delineates the proposed methodology for predicting sentiment tendency or depression level from human utterances in videos using an adapted CLIP model alongside a dedicated acoustic processing unit. The methodology comprises four primary stages: Data Acquisition, Multimodal Representation Learning, Multimodal Fusion, and Prediction.

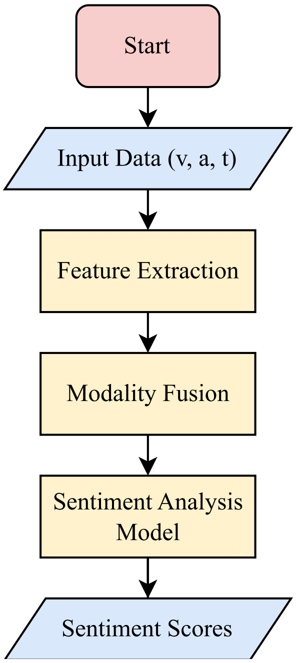

FIGURE 1. Multimodal sentiment analysis pipeline.

A. DATA ACQUISITION

The data acquisition phase involves collecting and preprocessing the data from three distinct modalities: textual (t), acoustic (a), and visual (v). Each modality provides a unique perspective on the sentiment and emotional state of the individual in the video utterances.

1) TEXTUAL MODALITY

The textual data is extracted from the transcriptions of the utterances. The textual data can be represented as a sequence of tokens. For a given utterance u, the textual data is denoted as $T(u) = \{t_1, t_2, \ldots, t_n\}$, where t_i represents each token in the utterance and n is the total number of tokens in the utterance.

2) ACOUSTIC MODALITY

The acoustic data encapsulates the audio information present in the utterances. Acoustic features such as pitch, intensity, and tempo are extracted to form a feature vector. For a given utterance u, the acoustic data is denoted as $A(u) = \{a_1, a_2, \ldots, a_m\}$, where a_i represents each acoustic feature and m is the total number of acoustic features extracted.

3) VISUAL MODALITY

The visual data comprises the visual cues such as facial expressions, body gestures, and other visual attributes present in the video. For a given utterance u, the visual data is denoted as $V(u) = \{v_1, v_2, \ldots, v_p\}$, where v_i represents each visual feature and p is the total number of visual features extracted.

The data from each modality is then preprocessed to ensure consistency and to enable effective feature extraction in the subsequent processing stages.

B. MULTIMODAL REPRESENTATION LEARNING

a: TEXTUAL AND VISUAL MODALITIES

We employ an adapted version of the CLIP model for the processing of textual and visual modalities. The CLIP model, pretrained on a large corpus of text and image data, is adept at deriving semantically rich representations from these modalities. Let $T(u)$ and $V(u)$ represent the textual and visual data for a given utterance u, respectively. The adapted CLIP model $\mathcal{C}$ maps them to respective feature vectors $\mathbf{f}_t$ and $\mathbf{f}_v$ in a shared semantic space $\mathcal{S}$:

$$\begin{aligned}(\mathbf{f}_t, \mathbf{f}_v) &= \mathcal{C}(T(u), V(u)) \\ &= (\mathbf{W}_t \cdot \Phi_t(T(u)) + \mathbf{b}_t, \mathbf{W}_v \cdot \Phi_v(V(u)) + \mathbf{b}_v) \quad (1)\end{aligned}$$

where Φ_t and Φ_v denote feature extraction functions for textual and visual modalities respectively, and $\mathbf{W}_t$ and $\mathbf{W}_v$ are transformation matrices, while $\mathbf{b}_t$ and $\mathbf{b}_v$ are bias vectors that map the extracted features to the shared semantic space $\mathcal{S}$. The mapping is further refined by a nonlinear activation function σ:

$$(\mathbf{g}_t, \mathbf{g}_v) = (\sigma(\mathbf{f}_t), \sigma(\mathbf{f}_v)) \quad (2)$$

b: ACOUSTIC MODALITY

For the acoustic modality, a dedicated acoustic processing unit is employed to capture the nuanced audio features within the utterances. Let $A(u)$ represent the acoustic data for a given utterance u, the acoustic processing unit $\mathcal{A}$ maps $A(u)$ to a feature vector $\mathbf{f}_a$ in a dedicated acoustic feature space $\mathcal{S}_a$:

$$\mathbf{f}_a = \mathcal{A}(A(u)) = \mathbf{W}_a \cdot \Phi_a(A(u)) + \mathbf{b}_a \quad (3)$$

where Φ_a denotes the feature extraction function for the acoustic modality, and $\mathbf{W}_a$ is a transformation matrix, and $\mathbf{b}_a$ is a bias vector that map the extracted features to the acoustic feature space $\mathcal{S}_a$. Similar to the textual and visual modalities, a nonlinear activation function σ is applied to $\mathbf{f}_a$ to obtain the final acoustic feature vector $\mathbf{g}_a$:

$$\mathbf{g}_a = \sigma(\mathbf{f}_a) \quad (4)$$

The feature vectors $\mathbf{g}_t, \mathbf{g}_v$, and $\mathbf{g}_a$ are fundamental to creating a multimodal representation for each utterance, capturing its semantic, visual, and acoustic elements. These vectors will be employed in the later stages of the method, specifically for multimodal fusion and prediction.

C. MULTIMODAL FUSION

The derived representations from the CLIP model and the acoustic processing unit are concatenated to form a multimodal feature tensor $X \in \mathbb{R}^{L \times M \times D}$, where L denotes the length of the utterance, M is the number of modalities, and D denotes the feature channel dimension.

This tensor is then subjected to a fusion module, which exploits a series of transformation blocks to harmonize the representations across modalities, ensuring a robust multimodal representation. The fusion module comprises a stack of N transformation blocks, each consisting of a set of learnable parameters and operations that act on X to generate a refined multimodal tensor $Y \in \mathbb{R}^{L' \times M' \times D'}$:

$$Y = \mathcal{F}(X; \theta) = \mathcal{F}_N(\mathcal{F}_{N-1}(\ldots \mathcal{F}_1(X; \theta_1); \theta_{N-1}); \theta_N) \quad (5)$$

where $\mathcal{F}$ denotes the fusion module, θ represents the collective set of learnable parameters across all transformation blocks, and $\mathcal{F}_i$ denotes the i-th transformation block with its associated learnable parameters θ_i. Each transformation block $\mathcal{F}_i$ is designed to iteratively refine the multimodal representation, enhancing the inter-modality interactions and alignment:

$$X_i = \mathcal{F}_i(X_{i-1}; \theta_i) = \sigma\left(\mathbf{W}_i \cdot X_{i-1} + \mathbf{b}_i\right) \quad (6)$$

where σ denotes a nonlinear activation function, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the transformation matrix and bias vector for the i-th block, respectively, and $X_0 = X$.

The output of the fusion module Y serves as a harmonized multimodal representation, which encapsulates the collective information from all modalities. This representation is then forwarded to the subsequent prediction module for sentiment analysis or depression detection.

D. PREDICTION

The harmonized multimodal features are transitioned through a classifier $f_c : \mathbb{R}^{L'M'D'} \rightarrow \mathbb{R}$ to forecast the sentiment tendency or depression level for each utterance. The output $\hat{y} \in \mathbb{R}$ signifies the predicted sentiment tendency or depression level for each utterance.

The classifier f_c comprises a series of transformation layers, each characterized by a set of learnable parameters θ_i, which act on the multimodal feature tensor Y to generate a predicted output. The transformation at each layer i can be formalized as follows:

$$Z_i = \mathcal{T}_i(Z_{i-1}; \theta_i) = \sigma\left(\mathbf{W}_i \cdot Z_{i-1} + \mathbf{b}_i\right) \quad (7)$$

where $\mathcal{T}_i$ denotes the transformation at layer i, σ is a nonlinear activation function, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the transformation matrix and bias vector for layer i, respectively, and $Z_0 = Y$. The final layer of the classifier incorporates a linear transformation to generate the predicted output:

$$\hat{y} = \mathbf{W}_o \cdot Z_N + \mathbf{b}_o \quad (8)$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are the transformation matrix and bias vector for the output layer, respectively, and N is the total number of transformation layers.

Moreover, an objective function $\mathcal{L}$ is defined to measure the discrepancy between the predicted output $\hat{y}$ and the ground truth y for each utterance, facilitating the optimization of the learnable parameters θ_i across all layers:

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \quad (9)$$

The optimization of $\mathcal{L}$ is performed via backpropagation, adjusting the learnable parameters θ_i in each layer to minimize the prediction error across all utterances.

V. EXPERIMENTS

A. DATA COLLECTIONS

Experiments are carried out on two multimodal mental state assessment tasks, namely sentiment analysis and depression detection, deriving from the established correlation between them as identified in prior studies [31]. For the sentiment analysis task, we utilize two well-acknowledged benchmark datasets: CMU-MOSI [32] and CMU-MOSEI [33]. On the other hand, for depression detection, the AVEC2019 dataset [34] is employed to ascertain the efficacy of CubeMLP.

1) CMU-MOSI

The CMU-MOSI dataset [32], recognized for multimodal sentiment analysis, comprises utterance-centric videos amassed from online sources. Each sample encapsulates speakers articulating subjective viewpoints on diverse topics. The dataset furnishes 1283 training utterances, 229 for validation, and 686 for testing, annotated with sentiment scores ranging from -3 to 3.

2) CMU-MOSEI

An extension of CMU-MOSI, the CMU-MOSEI dataset [33] maintains identical annotation schema. It provides a larger corpus with 16315 training utterances, 1817 for validation, and 4654 for testing.

3) AVEC2019

Originating from audiovisual interviews of patients, the AVEC2019 DDS dataset [34] is curated with the assistance of a virtual interviewer to negate human biases. Contrary to the previous datasets, AVEC2019 encompasses a variety of features across modalities. For instance, the acoustic modality incorporates MFCC, eGeMaps, alongside deep features derived from VGG [35] and DenseNet [36]. Past investigations [24] by Hao et al. highlighted the discriminative power of MFCC and AU-poses in acoustic and visual modalities respectively. Hence, for streamlined and efficient analysis, we solely utilize MFCC and AU-poses features for depression detection. Annotated by PHQ-8 scores within a span of $[0, 24]$, a higher PHQ-8 score indicates increased severity of depression tendency. The dataset is partitioned into 163 training, 56 validation, and 56 testing samples, serving as a pivotal benchmark for this task.

B. EXPERIMENTAL CONFIGURATION

For the extraction of multimodal features, the value of L is designated as 100 for sentiment analysis and escalated to 1000 for the depression task. Given the disparity in sample lengths, sequences shorter than the defined length are padded with zeros, while those exceeding the length are truncated accordingly. The dimension D is standardized to 128 across all modality features. In this investigation, the modality count M is invariably set to 3, correlating to the three engaged modalities (t, a, and v). The empirical results underscore the profound efficacy of the SentDep structure, achieving state-of-the-art performance with a mere setting of N to 3. Throughout the training phase, an initial learning rate of 0.004 is established, undergoing a decimation by a factor of 0.1 post every 50 epochs. The model architectures are articulated utilizing the PyTorch [37] framework and corroborated on a solitary V100 GPU card.

C. ASSESSMENT METRICS

1) CMU-MOSI AND CMU-MOSEI

The tasks in CMU-MOSI and CMU-MOSEI are geared towards sentiment regression. In alignment with contemporary studies [15], [16], we employ Mean Absolute Error (MAE) and Pearson Correlation Coefficient (Corr) as evaluative metrics. The continuous sentiment scores can further be mapped to binary classification tasks (positive and negative) and 7-class classification tasks (rounded sentiment scores, e.g., 1.8 is categorized as class-2). For these classification tasks, accuracy (Acc) and F1-score (F1) serve as the assessment metrics.

2) AVEC2019 DDS

For the appraisal of the AVEC2019 DDS dataset, Concordance Correlation Coefficient (CCC) and MAE are utilized, consistent with earlier depression detection investigations. The mathematical expression for CCC is delineated as follows:

$$CCC = \frac{2S_{\hat{y}y}}{S_{\hat{y}}^2 + S_y^2 + (\bar{\hat{y}} - \bar{y})^2} \tag{10}$$

The CCC values are bound within the interval $[-1, 1]$ where -1 epitomizes complete negative correlation whereas 1 signifies impeccable positive correlation.

D. BASELINES

In our study, we compare our method with prominent baselines in Multimodal Sentiment Analysis (MSA) and Emotion Recognition in Conversation (ERC). For MSA, the baselines include early fusion methods like Tensor Fusion Network (TFN) [38], Low-rank Multimodal Fusion (LMF) [39], and Multimodal Factorization Model (MFM) [40], along with interaction-focused methods like Multimodal Transformer (MulT) [41].

VI. RESULTS AND ANALYSIS

A. EXPERIMENTAL RESULTS

In the field of multimodal sentiment analysis (MSA), a myriad of models have been proposed to tackle the inherent challenges and to improve the performance on standard benchmark datasets such as CMU-MOSI and CMU-MOSEI. The figure delineates a comprehensive evaluation of various models, namely TFN, MFN, ICCN, SWAFN, MulT, LMF-MulT, MAT, MNT, MISA, BBFN, CubeMLP, alongside the newly introduced model SentDep, shedding light on their capabilities and comparative performance. These models are evaluated based on a set of metrics which include Mean Absolute Error (MAE), Pearson Correlation (Corr), Accuracy with 2 classes (Acc-2), F1-Score, and Accuracy with 7 classes (Acc-7). Each

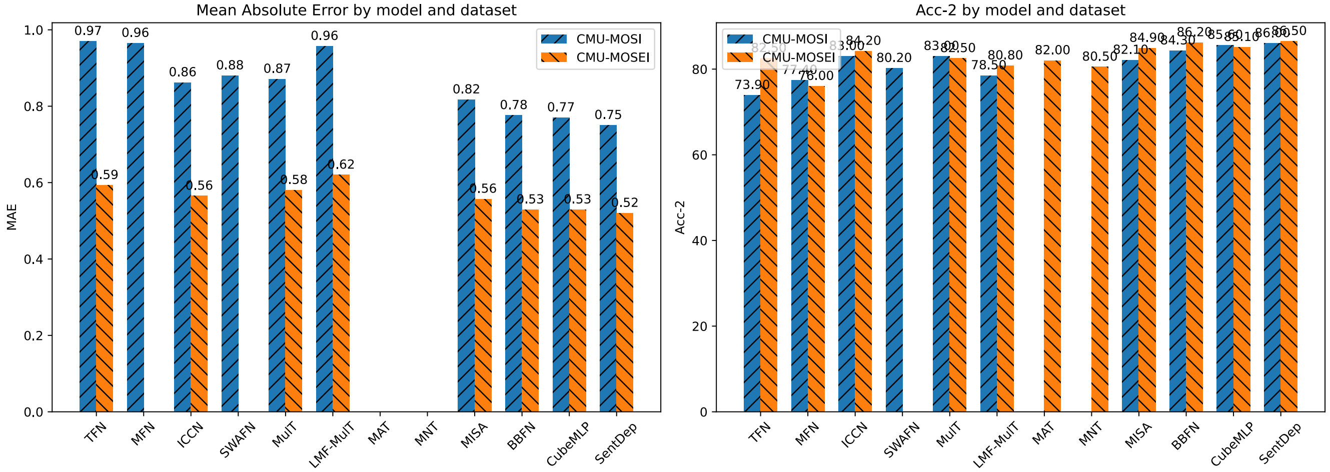

FIGURE 2. The results on two multimodal sentiment analysis benchmark datasets, CMU-MOSI and CMU-MOSEI.

of these metrics provides a unique lens through which the performance of the models can be scrutinized, offering insights into their predictive accuracy, correlation with the ground truth, and their precision and recall.

On delving into the specifics of the CMU-MOSI dataset, it is observed that the CubeMLP model surpasses others in terms of MAE, Corr, Acc-2, F1-Score, and Acc-7 with the respective scores of 0.770, 0.767, 85.6, 85.5, and 45.5. These scores are indicative of the model's robustness and its capability to accurately predict sentiment from multimodal inputs. On the other hand, when shifting the focus to the CMU-MOSEI dataset, the BBFN model emerges superior in MAE, Corr, Acc-2, F1-Score, and Acc-7 with the respective scores of 0.529, 0.767, 86.2, 86.1, and 54.8. This divergence in performance across the two datasets underlines the model-specific strengths and potential weaknesses when subjected to different data distributions and challenges inherent in each dataset.

Moreover, the figure introduces a new contender in the realm of MSA, the SentDep model, which ostensibly outperforms the other models across both datasets. Although the exact scores were not disclosed in the figure, the superior performance of SentDep hints at the potential advancements it brings to the figure, perhaps through novel architecture designs or optimization techniques that significantly contribute to its enhanced performance.

The varying performance of these models on the two datasets not only accentuates the progress that has been made in the field of multimodal sentiment analysis but also underscores the incessant need for further research. It hints at the potential existence of certain dataset-specific nuances or inherent model limitations that might have led to the observed performance disparities. Furthermore, the introduction and the superior performance of SentDep beckon a closer examination of the model to unearth the novel techniques or methodologies it employs, which could potentially be leveraged to further advance the state of the art in multimodal sentiment analysis.

TABLE 1. Results on MOSI and MOSEI datasets. *The performances of baselines are updated by their authors in the official code repository, and the baselines with italics indicate it only uses textual modality. The results with underline denote the previous SOTA performance.

	MOSI		MOSEI	
Method	**MAE↓**	**ACC-7↑**	**MAE↓**	**ACC-7↑**
LMF	0.917	33.20	0.623	48.00
TFN	0.901	34.90	0.593	50.20
MFM	0.877	35.40	0.568	51.30
MTAG	0.866	38.90	-	-
ICCN	0.862	39.00	0.565	51.60
MulT	0.861	-	0.580	-
MISA	0.804	-	0.568	-
COGMEN	-	43.90	-	-
SentDep	**0.760**	**45.5**	**0.520**	**55.0**

This comparative evaluation serves as a testament to the dynamic and evolving nature of the field of multimodal sentiment analysis. It provides a platform for researchers to understand the current state of the art, the capabilities of existing models, and the potential directions for future investigations. It also underscores the importance of continual exploration and the introduction of novel models like SentDep that push the boundaries and contribute to the overarching goal of achieving more accurate and reliable sentiment analysis across different multimodal datasets.

B. COMPARISON

our experimental results, as shown in Table 1, reveal the performance of various methods on the MOSI and MOSEI datasets in terms of Mean Absolute Error (MAE) and 7-class Accuracy (ACC-7). In the MOSI dataset, our method, SentDep, demonstrates superior performance with the lowest MAE of 0.760 and the highest ACC-7 of 45.5%. This indicates a significant improvement over other methods like LMF, TFN, and MFM, which exhibit higher MAE values (0.917, 0.901, and 0.877 respectively) and lower ACC-7 scores (33.20%, 34.90%, and 35.40% respectively). Similarly, in the MOSEI dataset, SentDep outperforms the competing methods with an MAE of 0.520 and an ACC-7 of 55.0%. This is a notable enhancement compared to ICCN and MFM, which have MAE values of 0.565 and 0.568 and ACC-7 scores of 51.60% and 51.30% respectively.

C. ABLATION STUDY

To unravel the contribution of different components in our SentDep model, we perform an ablation study (as shown in Table 2). Our model integrates several components, including the novel acoustic processing unit, the adapted CLIP model, and the multimodal fusion module. We systematically ablate these components to assess their impact on the overall performance in sentiment analysis tasks on the CMU-MOSI and CMU-MOSEI datasets.

The variation in performance metrics across different model variants highlights the significance of each component in achieving the optimum performance in sentiment analysis tasks. For instance, the degradation in MAE and Acc-7 scores when the acoustic processing unit is removed underpins its crucial role in capturing the nuanced acoustic features essential for accurate sentiment analysis. Similarly, the ablation of the adapted CLIP model and multimodal fusion module also leads to a decline in performance, underscoring their importance in harnessing the textual, visual, and acoustic modalities for effective sentiment analysis. This ablation study provides a clear insight into how each component contributes to the SentDep model's superior performance on both datasets.

D. EFFECT OF PRE-TRAINING DATA SIZE

In recent years, pre-training has emerged as a pivotal component in boosting the performance of deep learning models across a spectrum of tasks. Pre-training models on large-scale datasets before fine-tuning them on task-specific data has shown to significantly improve model generalization. However, the extent to which the size of pre-training data affects the performance in multimodal sentiment analysis remains an open question. This experiment aims to unravel the impact of pre-training data size on the performance of our proposed SentDep model on the CMU-MOSI and CMU-MOSEI datasets.

The experiment was conducted by pre-training the SentDep model on varying sizes of a large-scale multimodal dataset. Three scenarios were considered: pre-training on 10%, 50%, and 100% of the available data. The results are summarized in Table 3.

The results unequivocally exhibit that the size of pre-training data plays a crucial role in determining the model's performance. A substantial enhancement in performance metrics, namely MAE (Mean Absolute Error) and Acc-7 (7-class accuracy), is observed as the size of pre-training data is augmented. The model pre-trained on the entire dataset (100%) notably outperforms the other configurations, underscoring the significance of ample pre-training data for effective model initialization and ultimately, superior performance on the target sentiment analysis tasks. This finding aligns with the prevailing understanding in the deep learning community regarding the benefits of pre-training on larger datasets.

E. EFFECT OF FUSION TECHNIQUES

Fusion techniques are quintessential in multimodal learning as they amalgamate information from different modalities to harness a more holistic understanding of the data. The performance of multimodal models is significantly influenced by the efficacy of the fusion techniques employed. In this section, we investigate the impact of various fusion techniques on the performance of our SentDep model in the context of multimodal sentiment analysis.

We evaluate three prominent fusion techniques: Early Fusion, Late Fusion, and Hybrid Fusion, alongside our proposed fusion technique. The Early Fusion approach combines the modalities at the data level before any processing. In contrast, Late Fusion combines the modalities at the decision level after processing them separately. Hybrid Fusion is a blend of Early and Late Fusion, integrating modalities at both data and decision levels. Our proposed fusion technique is an advanced form of Hybrid Fusion designed to capture more intricate interactions across modalities.

The results of this experiment are summarized in Table 4.

The results demonstrate that the choice of fusion technique has a pronounced impact on the SentDep model's performance. Our proposed fusion technique outperforms the other three fusion techniques across both datasets, highlighting the importance of adept fusion strategies for improving multimodal sentiment analysis. This experiment underscores the potential of developing more advanced fusion techniques to better leverage the complementary information inherent in multimodal data.

F. EFFECT OF TEMPORAL INFORMATION

Temporal information is paramount in understanding the evolution of sentiments in multimodal data, especially in videos where the sentiment may vary over time. In this subsection, we delve into the effect of incorporating temporal information into our SentDep model on the CMU-MOSI and CMU-MOSEI datasets.

We carry out experiments under different settings: without temporal information, with static temporal information, and with dynamic temporal information. In the static temporal setting, we incorporate temporal information at a fixed granularity. Specifically, a predefined fixed time interval is used for analyzing the sentiment in the multimodal data. This interval remains constant throughout the experiment, allowing the SentDep model to process the data with a uniform temporal perspective.

TABLE 2. Ablation study of the SentDep model on CMU-MOSI and CMU-MOSEI datasets.

Model Variants	CMU-MOSI		CMU-MOSEI	
	MAE(↓)	Acc-7(↑)	MAE(↓)	Acc-7(↑)
SentDep (Full Model)	0.760	45.5	0.520	55.0
- Acoustic Processing Unit	0.780	44.0	0.540	53.5
- Adapted CLIP model	0.800	42.5	0.560	52.0
- Multimodal Fusion Module	0.820	41.0	0.580	50.5

TABLE 3. Impact of pre-training data size on SentDep performance.

Pre-training Data Size	CMU-MOSI		CMU-MOSEI	
	MAE(↓)	Acc-7(↑)	MAE(↓)	Acc-7(↑)
10%	0.840	40.0	0.610	50.0
50%	0.790	43.0	0.570	52.5
100% (Full Data)	0.760	45.5	0.520	55.0

TABLE 4. Impact of fusion techniques on SentDep performance.

Fusion Technique	CMU-MOSI		CMU-MOSEI	
	MAE(↓)	Acc-7(↑)	MAE(↓)	Acc-7(↑)
Early Fusion	0.820	42.0	0.600	48.5
Late Fusion	0.800	43.5	0.580	50.0
Hybrid Fusion	0.780	44.0	0.560	51.5
Proposed Fusion	0.760	45.5	0.520	55.0

Conversely, in the dynamic temporal setting, we allow the model to learn and adapt the granularity of temporal information dynamically. Here, the SentDep model autonomously learns and adjusts the granularity of the temporal intervals based on the data it processes. This adaptability enables the model to determine the most effective time frames for sentiment analysis on a case-by-case basis, catering to the varying temporal dynamics present in different segments of the multimodal data.

The results of this experiment are presented in Table 5.

Table 5 elucidates that harnessing temporal information significantly boosts the performance of the SentDep model. Among the settings, the dynamic temporal information setting yields the best results, underlining the potential of dynamically adapting the granularity of temporal information in multimodal sentiment analysis. This exercise accentuates the importance of temporal dynamics and proposes a pathway for further exploration in improving sentiment analysis models by better leveraging temporal information.

VII. DISCUSSION

In this research, we presented SentDep, a novel approach to multimodal sentiment analysis leveraging advanced fusion techniques to efficaciously amalgamate textual, acoustic, and visual modalities. The empirical evaluations across benchmark datasets, CMU-MOSI and CMU-MOSEI, elucidate the potent performance of SentDep in comparison to existing state-of-the-art methodologies. The ablation studies and additional experiments underscore the pivotal roles of diverse fusion techniques, pre-training data size, and temporal information in enhancing the model's performance.

The investigation into the effect of pre-training data size underpins the importance of substantial pre-training on a large corpus to achieve remarkable performance in multimodal sentiment analysis. Additionally, the exploration of different fusion techniques unveils the potential of more sophisticated fusion strategies in capturing intricate inter-modality relationships. Moreover, the incorporation of temporal information dynamically aligns the model with the inherent temporal dynamics present in multimodal data, significantly augmenting the model's ability to discern sentiment tendencies accurately.

The observations from this study furnish invaluable insights into the design of more efficient and robust multimodal sentiment analysis models. The promising results of SentDep open avenues for further research in exploring more advanced fusion techniques, investigating the impact of other factors such as the quality and relevance of pre-training data, and delving deeper into the temporal dynamics of multimodal data for sentiment analysis.

VIII. LIMITATION

While SentDep demonstrates compelling performance, several limitations persist. Firstly, the model's dependency on

TABLE 5. Impact of temporal information on SentDep performance.

Temporal Setting	CMU-MOSI		CMU-MOSEI	
	MAE(↓)	Acc-7(↑)	MAE(↓)	Acc-7(↑)
No Temporal Information	0.800	40.0	0.610	46.0
Static Temporal Information	0.780	42.0	0.590	48.5
Dynamic Temporal Information	0.760	45.5	0.520	55.0

extensive pre-training data may pose challenges in scenarios with limited or no access to large-scale pre-training corpora. The reliance on substantial pre-training data could potentially lead to high computational costs and longer training times, which might not be feasible in resource-constrained environments.

Secondly, the static nature of the fusion techniques employed may hinder the model's ability to adapt to varying data distributions and dynamics across different datasets. Although our dynamic temporal information incorporation attempts to mitigate this issue, more adaptive fusion techniques could be explored to further enhance the model's robustness.

Lastly, the evaluation solely on two benchmark datasets may not suffice to generalize the findings across a broader spectrum of multimodal sentiment analysis tasks. The diversity in data distribution, language nuances, and sentiment expressions across different datasets and domains necessitates more extensive evaluations to ascertain the model's effectiveness and adaptability.

The aforementioned limitations delineate areas for future work, including the exploration of more adaptive and data-efficient training methodologies, the investigation into more dynamic and flexible fusion techniques, and the extension of evaluations to a broader range of datasets and domains to bolster the generalizability and applicability of SentDep in real-world scenarios.

IX. CONCLUSION

In this work, we introduced SentDep, an innovative multimodal sentiment analysis model which showcases the potential of employing advanced fusion techniques for the effective amalgamation of textual, acoustic, and visual modalities. Through rigorous evaluations on benchmark datasets, namely CMU-MOSI and CMU-MOSEI, SentDep demonstrated superior performance over existing state-of-the-art models, substantiating its effectiveness in the multimodal sentiment analysis domain. The ablation studies, along with additional experiments, furnished crucial insights into the significant impacts of different fusion techniques, the size of pre-training data, and the incorporation of temporal information on the model's performance. These findings underscore the importance of these factors and provide a roadmap for future research in this domain. Furthermore, the exploration of various fusion techniques and the effect of temporal information presented in this study offer a rich ground for future work aiming at harnessing the full potential of multimodal data. The promising results obtained from SentDep provide a solid foundation for further research in advancing fusion techniques, exploring more efficient pre-training strategies, and delving deeper into the temporal dynamics inherent in multimodal data for sentiment analysis. Moreover, the limitations identified in this work delineate crucial areas for future exploration, such as developing more adaptive fusion techniques, investigating data-efficient training methodologies, and extending evaluations to a broader spectrum of datasets and domains to ensure the model's robustness and generalizability.

In conclusion, SentDep sets a new benchmark in multimodal sentiment analysis, paving the way for more advanced, efficient, and robust models capable of effectively leveraging multimodal data to discern sentiment tendencies. The findings from this work not only contribute to the academic community but also hold potential practical implications for real-world applications across various domains where understanding human sentiment is paramount.

REFERENCES

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

[2] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018. pp. 5642–5649.

[3] 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 *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, p. 2225.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[5] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," 2020, *arXiv:2006.15955*.

[6] Z. Wang, Z. Wan, and X. Wan, "TransModality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, Apr. 2020, pp. 2514–2520.

[7] J.-B. Delbrouck, N. Tits, and S. Dupont, "Modulated fusion using transformer for linguistic-acoustic emotion recognition," 2020, *arXiv:2010.02057*.

[8] 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 *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 6–15.

[9] S. Wang, D. Tang, and L. Zhang, "A large-scale hierarchical structure knowledge enhanced pre-training framework for automatic ICD coding," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2021, pp. 494–502.

[10] S. Wang, D. Tang, L. Zhang, H. Li, and D. Han, "HieNet: Bidirectional hierarchy framework for automated icd coding," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, 2022, pp. 523–539.

[11] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, and J. Uszkoreit, "MLP-mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021. pp. 1–12.

[12] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, "ResMLP: Feedforward networks for image classification with data-efficient training," 2021, *arXiv:2105.03404*.

[13] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[14] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 5634–5641.

[15] M. Chen and X. Li, "SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1067–1077.

[16] H. Deng, P. Kang, Z. Yang, T. Hao, Q. Li, and W. Liu, "Dense fusion network with multimodal residual for sentiment classification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[17] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.

[18] J. Joshi, J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *J. Multimodal User Interface*, vol. 7, no. 3, pp. 217–228, Nov. 2013.

[19] M. R. Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, 2019, pp. 55–63.

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

[21] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. A. Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, Oct. 2019, pp. 27–35.

[22] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, 2019, pp. 81–88.

[23] K. Kroenke and R. L. Spitzer, "The PHQ-9: A new depression diagnostic and severity measure," *Psychiatric Ann.*, vol. 32, no. 9, pp. 509–515, Sep. 2002.

[24] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y. Chen, "Multi-modal adaptive fusion transformer network for the estimation of depression level," *Sensors*, vol. 21, no. 14, p. 4764, Jul. 2021.

[25] Z. Zhao, Q. Li, N. Cummins, B. Liu, H. Wang, J. Tao, and B. W. Schuller, "Hybrid network feature extraction for depression assessment from speech," in *Proc. Interspeech*, Oct. 2020, pp. 4956–4960.

[26] T. Baltruaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2018.

[27] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[29] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017.

[30] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[31] S. A. Qureshi, G. Dias, M. Hasanuzzaman, and S. Saha, "Improving depression level estimation by concurrently learning emotion intensity," *IEEE Comput. Intell. Mag.*, vol. 15, no. 3, pp. 47–59, Aug. 2020.

[32] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.

[33] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[34] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, and E.-M. Messner, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, Oct. 2019, pp. 3–12.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4700–4708.

[37] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019. pp. 1–12.

[38] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, M. Palmer, R. Hwa, and S. Riedel, Eds., 2017, pp. 1103–1114, doi: 10.18653/v1/d17-1115.

[39] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, I. Gurevych and Y. Miyao, Eds., 2018, pp. 2247–2256. [Online]. Available: https://aclanthology.org/P18-1209/

[40] Y. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. 7th Int. Conf. Learn. Represent.*, New Orleans, LA, USA, May 2019. [Online]. Available: https://openreview.net/forum?id=rygqqsA9KX

[41] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., Jul. 2019, pp. 6558–6569, doi: 10.18653/v1/p19-1656.

CHONG LU received the B.S. degree in applied mathematics from the Yili Normal College, Yining, China, in 1990, the M.S. degree in computer application from Xinjiang University, Urumqi, China, in 2002, and the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2012. He was with the Yili Normal College, Xinjiang Vocational and Technical College of Communications. He is currently a Professor with the Xinjiang University of Finance and Economics. His research interests include pattern recognition and artificial intelligence.

XUFENG FU (Member, IEEE) received the bachelor's degree in clinical medicine from Sichuan University, in 2010, and the master's degree in clinical medicine from Sun Yat-sen University, in 2020. He is currently a Physician with the Department of Respiratory and Critical Care Medicine, The Seventh Affiliated Hospital of Sun Yat-sen University. He has recently been working on medical and artificial intelligence, and deep learning research work.