IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 11, 2021, accepted June 23, 2021, date of publication June 28, 2021, date of current version July 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3092735

Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification

SANGHYUN LEE[1], DAVID K. HAN[2], AND HANSEOK KO[1], (Senior Member, IEEE)
[1]School of Electrical Engineering, Korea University, Seoul 136-713, South Korea
[2]Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA

Corresponding author: Hanseok Ko (hsko@korea.ac.kr)

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program, Development of Robot's Natural Language Recognition and Emotional Dialogue Technology) through the Ministry of Trade, Industry and Energy (MOTIE), South Korea, under Grant -10073162.

ABSTRACT Human communication includes rich emotional content, thus the development of multimodal emotion recognition plays an important role in communication between humans and computers. Because of the complex emotional characteristics of a speaker, emotional recognition remains a challenge, particularly in capturing emotional cues across a variety of modalities, such as speech, facial expressions, and language. Audio and visual cues are particularly vital for a human observer in understanding emotions. However, most previous work on emotion recognition has been based solely on linguistic information, which can overlook various forms of nonverbal information. In this paper, we present a new multimodal emotion recognition approach that improves the BERT model for emotion recognition by combining it with heterogeneous features based on language, audio, and visual modalities. Specifically, we improve the BERT model due to the heterogeneous features of the audio and visual modalities. We introduce the Self-Multi-Attention Fusion module, Multi-Attention fusion module, and Video Fusion module, which are attention based multimodal fusion mechanisms using the recently proposed transformer architecture. We explore the optimal ways to combine fine-grained representations of audio and visual features into a common embedding while combining a pre-trained BERT model with modalities for fine-tuning. In our experiment, we evaluate the commonly used CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets for multimodal sentiment analysis. Ablation analysis indicates that the audio and visual components make a significant contribution to the recognition results, suggesting that these modalities contain highly complementary information for sentiment analysis based on video input. Our method shows that we achieve state-of-the-art performance on the CMU-MOSI, CMU-MOSEI, and IEMOCAP dataset.

INDEX TERMS Multimodal emotion recognition, heterogeneous features, transformer, attention based multimodal, BERT.

I. INTRODUCTION

An effective communication among humans requires not only intellectual exchange but of sharing contextual emotions. While most humans are natural in perceiving others' emotional states, the sensitivities of recognizing key sentiments may not be even among us. When we look at Leonardo Da Vinci's 'Mona Lisa' many of us may judge that she is smiling and may conclude that her emotional state is positive. It turns out, however, that this seemingly prevalent observation may not necessarily be universal. In fact, an analysis of the painting has been undertaken to determine if the painting is really conveying a positive emotion [1]. Perceiving other's emotional states is obviously an important factor in peer-to-peer human interactions. It is also of paramount importance when considering effective human-to-computer interactions. As such, there has been a steady stream of efforts in developing techniques to enable machines

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

to better recognize and estimate a person's emotional state. Much of the work in emotion recognition focused on unimodal handcrafted features. From speech, Han *et al.* [2] extracted 238 Low-Level Descriptors (LLDs) at speech frame level using openSMILE [3] and had them automatically aligned with emotional labels with a recurrent neural network based Connectionist Temporal Classification (CTC) model. Tzinis *et al.* [4] compared the impact of choosing time-scales and a variety of LLDs (local features) that are relevant to global features applied to an RNN. As in other fields, there has been an explosion of deep learning techniques applied for emotion recognition to extract high-level features from raw data. The following are some of these methodologies based on speech as a unimodal feature. Han *et al.* [5] proposed speech emotion recognition using deep neural network (DNN) from raw data. Trigeorgis *et al.* [6] proposed a solution based on a Convolutional Neural Network (CNN) to recognize local emotional features with contextual considerations obtained from a Long Short-Term Memory (LSTM) network. Gao *et al.* [7] proposed a rapid end-to-end emotion recognition model with a simple structure using CNN for real-time speech emotion recognition. Zhao *et al.* [8] proposed the CTC attention model by applying attention mechanisms, allowing the model to focus on emotionally salient parts of the speech signal. For some of the emotion recognition research, visual features based on images or videos have been considered as providing salient clues to a person's emotional state. These efforts considered local features such as Gabor wavelet [9] or global geometric [10] features. Sikka *et al.* [11] proposed a method that extracts facial expressions and head posture from a video sequence and aligns them as sequential features for sentiment analysis. Cole *et al.* [12] presented a method for synthesizing neutral expressions of facial features by extracting facial landmarks. For deep learning based emotion recognition, [13]–[16] utilized CNN to extract facial features salient to expressed emotions. Another important feature for classifying emotions is the textual content of speech. Wilson *et al.* [17] proposed a keyword-based method of exploring opinion clauses. Yang *et al.* [18] proposed sentiment classification of the web blog corpus using SVM (Support Vector Machine) and CRF (Conditional Random Field) techniques. Some of the textual based approaches took advantage of recently developed word embeddings, such as Glove [19] or Word2Vec [20] as follows. Zhang *et al.* [21] used CNN while Abdul-Mageed *et al.* [22] applied RNN for classifying emotional states. Ghosal *et al.* [23] tackled the problem of individual sentiment recognition from a multi-person conversation using RNN. Zhang *et al.* [24] proposed an interactive graph-based CNN for recognizing sentiments. In addition, an attention mechanism based model that applies feature fusion for context weighting and summarizing has been proposed [25]. To improve performance, there have been some efforts of fusing different features extracted from a unimodal source. Lee *et al.* [26] proposed a method that combines two features of CNN and BERT in parallel from an audio spectrogram. Xu *et al.* [27] proposed "Hierarchical Grained and Feature Model (HGFM)" by fusion of handcrafted features and Gated Recurrent Unit (GRU) network extracted features. [28]–[30] built an emotion recognition model by fusing handcrafted features and deep learning based features from facial images. However, the effectiveness of these unimodal feature based methods was found to be insufficient to infer the speaker's sentiment as much of salient emotional features are expressed simultaneously via different modalities [31]. For example, extracting emotional elements from a sentence "It's it's (stutter) a fantasy" considering only its textual content is very difficult as it is quite ambiguous in expressing the speaker's emotional state. In such cases, considering only textual contents may not be sufficiently discriminative to discern the subject's emotions. To enhance emotion recognition performance over unimodal methods, there have been efforts in simultaneously considering multiple modalities including audio, visual, and textual features [32]–[36]. [32], [36] proposed an attention-based framework that uses attention on multimodal representations to learn the contributing features among them. [32] extracts the representation of each modality using a transformer similar to our proposed method. However, the fusion method of [32], [36] contributed to the learning of multimodal features, but the influence of text-based features was insufficient. [37] shows that text features dominate over audio and visual features in multimodal emotion recognition. As a means of fusion, [34] utilized text information by adjusting word expressions with audio and visual features. However, these latest multimodal studies have only used handicraft functions in audio and visual, and are lacking in the use of textual information [32]–[36]. Even with models utilizing multimodal fusion, the emotion recognition performance of these approaches has not been shown to be sufficient for useful applications. Further improvement, therefore, is needed. Based on these previous efforts, it can be summarized that fusion of features at various levels and modalities would help. From the findings of [27]–[29] as well as our investigation, it's been shown that handcrafted features in audio and visual domains seem to provide improved saliencies associated with emotions over deep network extracted features. Combining audio, visual, and text features with an effective fusion would lead to additional performance improvements. As it's been observed from the existing studies, textual features and the associated context deliver more influential features in determining the speaker's emotions. As such, a greater emphasis should be placed on text based features when these multimodal features are combined. While word embeddings and employment of RNNs have been shown to be effective, based on our reviews, pre-trained models based on large-scale data seem to deliver additional improvements in capturing contextual information. As such, we propose to integrate BERT in textual feature extraction in our architecture. In summary, we propose to address these challenges of recognizing emotions from analyzing utterance-level multimodal input as follows. First, we combine information from various unimodal features with

relevant saliency, then efficaciously fuse them with appropriate placement of relative weights among the modalities for accurately recognizing emotions. Our proposed model, Heterogeneous Features Unification(HFU-BERT), integrates BERT into our architecture to effectively combine heterogeneous features extracted from both handcrafted and deep learning based methods. The main contributions of this paper can be summarized as follows:

- We present an effective way of complementing and fusing both handcrafted and deep learning bottleneck features in audio, visual, and text for accurately predicting emotions therein.
- We develop an architecture to effectively combine BERT extracted textual features with other modality based features by structuring the network with appropriate emphasis placed on each feature modality.

We demonstrate experimentally that our model outperforms state-of-the-art emotion recognition models. To prove the effectiveness of our method, we test it using public multimodal sentiment analysis datasets CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) [38] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [39], and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [40]. The paper is organized as follows: Section II summarizes previous works related to this research field. Section III and IV describes video processing and modality feature extraction with details on the proposed HFU-BERT framework. Session V presents an experimental analysis to evaluate the performance of the proposed model in the MOSI, MOSEI, and IEMOCAP corpus. Section VI discusses the results. Section VII describes the ablation studies. Lastly, Section VIII presents a summary and future work.

II. RELATED WORK

Multimodal sentiment comprehension is a popular research area in recent years [41]–[46]. Previous work used early fusion approach to concatenate input features from different modalities and then immediately conduct multimodal fusion [47]–[49]. Decision-level fusion method trains different models for each modality and then integrate the inference of each modality to make a final decision [50]–[52]. Gajšek *et al.* [53] employed weighted product rule for audio and video decision-level fusion. Decision-level fusion, however, cannot explore inter-modality dynamics by design. Therefore, the following efforts opted to concatenate multimodal features as a means of fusion and train them in an integrated architecture to embed inter-modal correlations in the learning process. Harwath *et al.* [54] fused dataset of images and audio in this fashion to associate spoken words and their visual representation. Similarly, Zhou *et al.* [55] combined the features of text and audio modality and proposed a semi-supervised multi-path generative neural network to better infer emotion. Duong *et al.* [56] used pooling to classify emotional states by fusion of features in image and text modalities. While these efforts focused on using bimodal fusion, others explored combining audio, visual, and language features together. Xu *et al.* [57] explored aspect-level multimodal sentiment analysis by proposing a multi-hop memory network to model the cross-modality and single-modality interactions among the three feature domains. Zadeh *et al.* [58] proposed a tensor fusion network that expresses multimodal fusion information using the product of image, audio, and visual features. Pu Liang *et al.* [59] proposed a Recurrent Multistage Fusion Network (RMFN) that decomposes a multimodal fusion problem into multiple stages using LSTM to capture synchronous and asynchronous multimodal interactions. To alleviate the added computational cost due to considering all three modalities, Liu *et al.* [60] reduced the computational complexity of the parameters by applying a low-rank multimodal fusion method that uses a low-rank tensor. Poria *et al.* [61] applied LSTM separately to text, visual, and audio first, and their extracted features are combined in a multi-level fusion learning architecture. Due to its effectiveness, the attention mechanism has recently attracted some in the field for its ability to combine multimodal features. Ghosal *et al.* [36] proposed a multi-attention recurrent network framework that learns features using attention for multimodal representation. Tsai *et al.* [41] proposed learning interactions between the modalities by designing an attention based cross-modal architecture using multimodal transformers. Also, recently, transfer learning techniques that use pre-trained networks to extract features [26], [62]–[64] have advanced significantly. BERT, a Transformer based model, has shown performance improvement by fine-tuning from pre-trained weights for a specific downstream task [65]. As demonstrated here, employing BERT with its wide availability of pre-trained weights, can save both time and cost in a variety of tasks. We present an effective fusion framework for fine-tuning by fusing heterogeneous nonverbal features that complement the linguistic expressions of BERT. The following sections describe our proposed architecture and the training process. To better understand the multimodal fusion method and the importance of BERT, we conducted ablation studies to understood the impact of our proposed model.

III. PREPROCESSING

This section presents a method for extracting heterogeneous features for the multimodal emotion recognition tasks. We introduce video preprocessing and feature extraction steps for audio signals, visual and textual information from videos. Figure 1 shows the architecture of our proposed model.

A. VIDEO PREPROCESSING

In this paper, we focus on analyzing human sentiment from utterance-level video. The multiple utterance [66] means a unit of speech bounded by sentences, and process video as utterance units. Let us assume a video to be considered as $V_i = [utt_{i,1}, utt_{i,2}, \ldots, utt_{i,L}]$ is the i^{th} utterance belongs

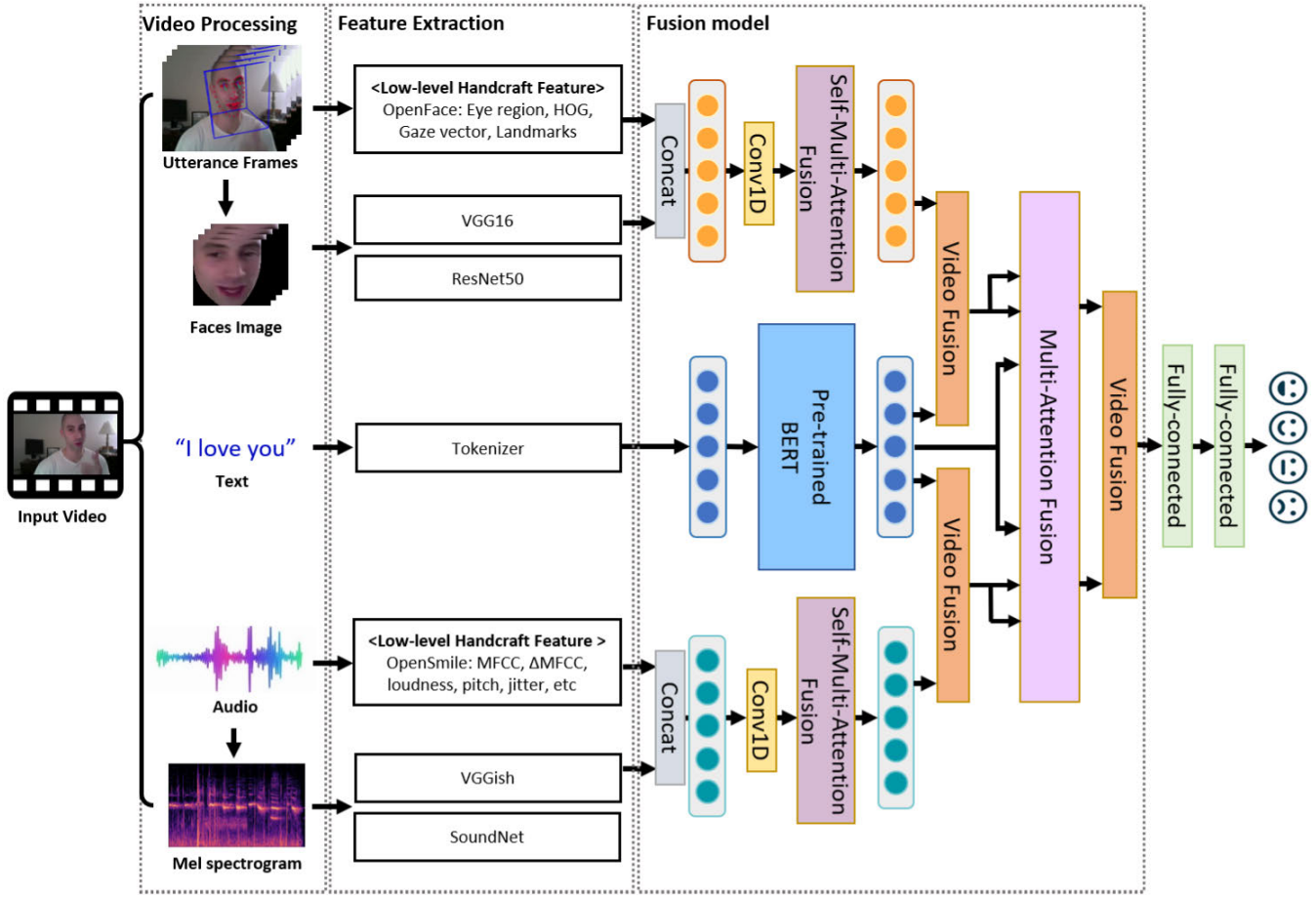

FIGURE 1. The figure above represents the architecture of the model proposed in this paper. We process raw video in utterance units to obtain audio, visual, and text information. Audio and visual modalities extract each handcraft features and bottleneck features through the toolkit and pre-trained models. Each heterogeneous feature of audio and visual is connected and the high-level representation is extracted the fusion process through the self-multi-attention fusion module. Text extracts expressions through pre-trained BERT. Then, through our BERT with heterogeneous function unification (HFU-BERT) architecture, we predict the final emotional state by simultaneously fine-tuning the BERT and fusion applying the multi-attention fusion module.

to video v_i and L is the number utterances in the video. We denote the set of modalities for each utterance as $M \in \{a, v, l\}$ as audio, visual, text and extract the feature set for the input sequence. The input features x_t and emotion label y_t of the i-*th* utterance are expressed as follows:

$$X = \left[x_t^M : 1 \leq t \leq T, x_t^M \in \mathbb{R}^{dim}\right]. \quad (1)$$

$$Y = \left[y_t : 1 \leq t \leq T, y_t \in \mathbb{R}^1\right]. \quad (2)$$

where *dim* is the input features of the modality for each sequence t. Following prior work, To obtain time series data of the same length, each feature is zero padded based on the word boundary. In addition to the handicraft features in audio and visual, we also consider the bottleneck features extracted from the pre-trained models.

B. AUDIO FEATURES

We extract three representative handcraft features and bottleneck features that are frequently utilized in the field of speech emotion recognition. As handcraft features, we extracted 1582 dimensional features using the openSMILE toolkit [3]. We follow the same procedure as Schuller *et al.*, [67], extracted with robust emotional features. The toolkit encapsulates several features including Mel Frequency Cepstral Coefficient (MFCC), $\triangle$MFCC, loudness, pitch, jitter, etc. with acoustic low-level descriptors (LLDs). LLDs capture the signal of affective states by using prosodic information from different speakers. All LLDs are extracted with window shift using a Hamming window of 25 ms step size. We specify the audio features extracted with the openSMILE toolkit as LLDs. In addition, we extract deep learning bottleneck features using SoundNet [68] and VGGish [69] models. The SoundNet network extracts rich, natural sound representations utilizing pre-trained models on large amounts of unlabeled video sound datasets. The SoundNet is a one-dimensional convolutional network and is composed of convolutional layers and pooling layers. We process audio data as network input and extract high-level 1024D bottleneck features from Conv 7-layer. In our previous work, we used VGGish Bottleneck Features to show that it is effective in

speech emotion recognition [70], [71]. The VGGish has been pre-trained for AudioSet [72], which includes an event class with more than 600 audio clips and 10 YouTube video soundtracks with over 2 million human labels. We extract the log spectrogram from the audio and process it as a VGGish network input. The VGGish input log spectrogram is 96×64. In VGGish, we extract semantically meaningful and high-level 128D embedding features from the last fully connected layer. We extract LLDs (a^{LLDs}), SoundNet (a^{sound}), and VGGish (a^{vggish}) features from raw audio, and select and fuse the features $x^a = (a^{LLDs}, a^{sound}, a^{vggish})$ suitable for the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets. through experiments. The formula is as follows.

$$a^{LLDs} \in \mathbb{R}^{1582}. \tag{3}$$

$$a^{sound} = \left\{ a_t^{sound} | a_t^{sound} \in \mathbb{R}^{1024}, t = 0, \ldots, T \right\}. \tag{4}$$

$$a^{vggish} = \left\{ a_t^{vggisht} | a_t^{vggish} \in \mathbb{R}^{128}, t = 0, \ldots, T \right\}. \tag{5}$$

C. VISUAL FEATURES

Visual features consist of the OpenFace [73] estimators for the whole frame and the representation of the VGG [74] and ResNet [75] on the face regions. For the OpenFace features, the OpenFace toolkit is used to extract facial landmarks estimated from the eye region HOG, gaze vector, head posture, hard head shape, and facial action units [76] intensity representing facial muscle movements. Face images extracted with OpenFace zero the background according to the face contour indicated by the facial landmarks. We extract 709D with OpenFace features. Also, in OpenFace toolkit, the parts that could not detect landmarks of the face recognition module were removed and used. As a deep learning feature, we utilize the VGG16 and ResNet50, which are pre-trained facial recognition models on a large facial dataset [77]. The VGG16 extracts 4096D facial features from the fully connected 7-layer and ResNet50 extracts 2048D features from the average pool. The overall features $x^v = (v^{face}, v^{vgg}, v^{resnet})$ of each openface (v^{face}), vgg16 (v^{vgg}), and resnet50 (v^{resnet}) extracted from the visual are as follows.

$$v^{face} = \left\{ v_t^{face} | v_t^{face} \in \mathbb{R}^{709}, t = 0, \ldots, T \right\}. \tag{6}$$

$$v^{vgg} = \left\{ v_t^{vgg} | v_t^{vgg} \in \mathbb{R}^{4096}, t = 0, \ldots, T \right\}. \tag{7}$$

$$v^{resnet} = \left\{ v_t^{resnet} | v_t^{resnet} \in \mathbb{R}^{2048}, t = 0, \ldots, T \right\}. \tag{8}$$

D. TEXT PREPROCESSING

We tokenize the text on subword units by WordPiece [78] in the same as in BERT [65]. As an example, the sentences are divided into words as [''anyhow'',''it'', ''was'', ''really'', ''good''], where the word ''anyhow'' is ''any'' and ''##how'' Separated by, ''##'' indicates that the pieces belong to one word. Given a text sequence of word-piece tokens $x^l = [t_1, t_2, \ldots, t_n]$, where n is the number of sequence length. Since the embedding layer of the BERT model has a special token [CLS] added at the beginning of the sequence to obtain the representation of the whole input, the output of the last encoder layer is a $n + 1$ length sequence which is denoted as $x^l = [CLS, t_1, t_2, \ldots, t_n]$. The vector representations x^l of input tokens are computed via summing the corresponding token embedding, position embedding, and segment embedding.

IV. HFU-BERT ARCHITECTURE

This section describes the workflow of the HFU-BERT method for feature fusion shown in detail in Figure 2. We search for heterogeneous features with handcraft features and deep learning bottleneck features for both audio and visual and fuse the features suitable for the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets using the Multi-headed Attention encoder. Then we review the standard BERT model and present our multimodal extension of BERT. We describe the process of a transformer that can effectively fuse audio and image heterogeneous feature information. Finally, we report in detail the proposed multimodal architecture HFU-BERT.

A. PRE-TRAINED BERT

To utilize information from text data, we use BERT [65], a transformer-based language model [79] that achieves state-of-the-art performance on various NLP tasks. BERT is commonly trained with two steps: pre-training and fine-tuning. It is pre-trained on a large corpus of unlabeled text which includes the entire Wikipedia (about 2.5 billion words) and a book corpus (800 million words). As opposed to directional models, which read the text input sequentially, BERT is considered a bidirectional path. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). In this paper, we use the baseline BERT model which includes 12 transformer blocks, 12 attention heads, and 110 million parameters. After embedding, each word in an utterance is represented by a 768D vector. Specifically, we first fine-tune the BERT-base model using the masked language model and adjust the impact of BERT and effective fusion sentiment prediction objectives for audio and visual components. The encoder architecture of the BERT model utilizes the Self-Multi-Attention fusion module, which is described in the next section.

B. SELF-MULTI-ATTENTION FUSION

Transformer [79] is inspired by a non-repetitive neural architecture designed for sequential data, building a basic encoding block that assumes a latent adaptation that fuses various functions. Attention within Transformer is motivated by how we pay visual attention to different regions of an image or correlate the most salient features in audio. Also, in sentences, using word attention allows us to focus on contextual words. First, we extracted the effective heterogeneous features (handcraft and bottleneck features) for each audio and visual from the data and then used the concatenation method for fusion. It applies a relative positional encoding mechanism to enable state reuse without causing temporal confusion between frames. Because the dimensions of the

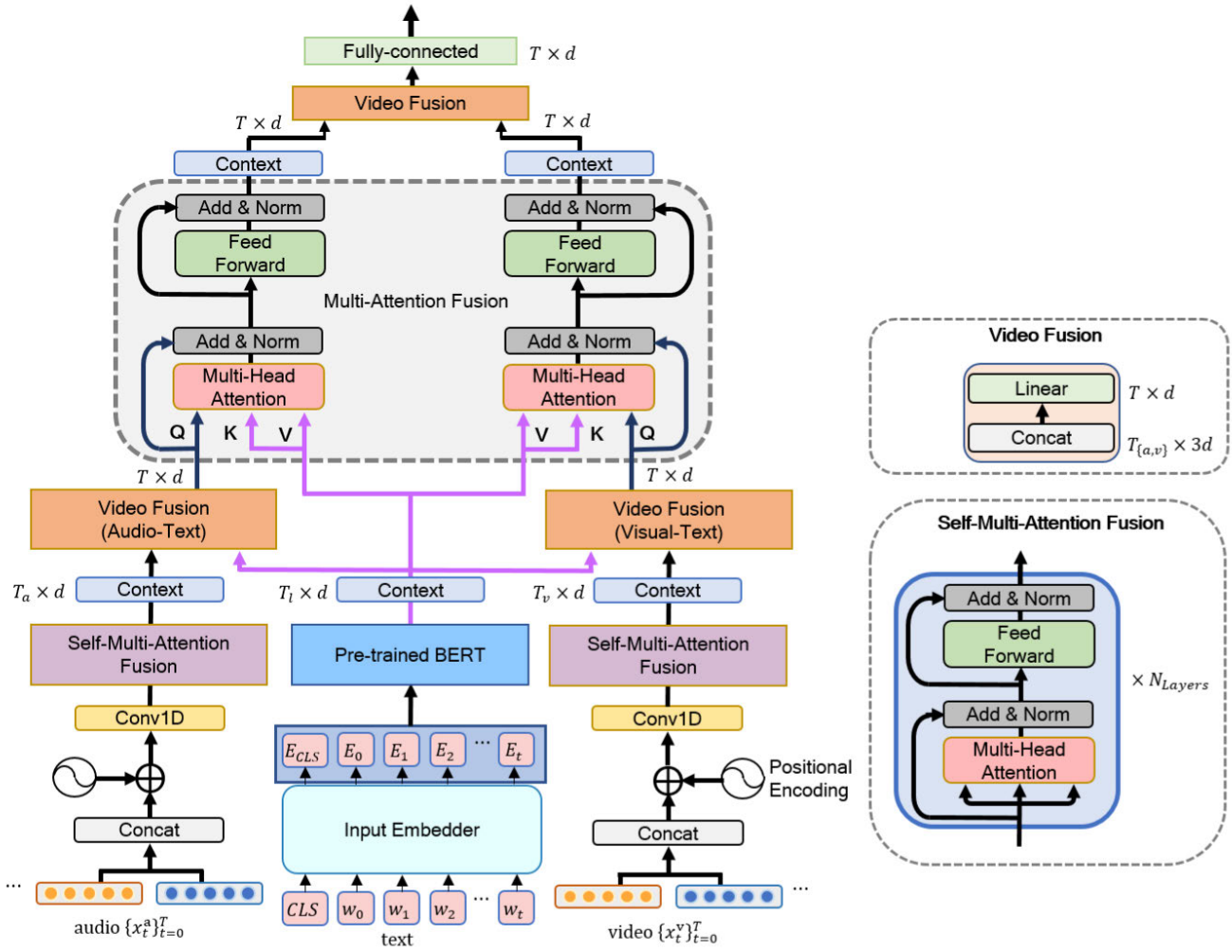

FIGURE 2. The proposed BERT with heterogeneous features unification (HFU-BERT) framework. Given an input consisting of heterogeneous features, namely, audio and visual, internal representations are produced by a corresponding feature transformer (self-multi-attention fusion module). Then, the vocabulary representations output from BERT is fused to nonverbal representations with video fusion and multi-attention fusion modules.

audio and visual features are obviously different than the text features, we use a 1D temporal convolution layer to control them in the same dimension [32].

$$\left\{\hat{x}^a, \hat{x}^v\right\} = Conv1D\left(\left\{x^a, x^v\right\}, k^{\{a,v\}}\right), \tag{9}$$

where $k^{\{a,v\}}$ represents the size of the convolution kernel for audio and visual modalities. We use the multi-head attention in [79] to fuse the attention features from each modality. The multi-head mechanism executes scaled dot-product attention multiple times in parallel. We propose not only the contextual information between each modality feature, but also the textual representation of BERT and two attention mechanisms (Self-Multi-Attention Fusion and Multi-Attention Fusion) that combine these features. As shown in the bottom-right corner of Figure 2, Self-Multi-Attention Fusion and Multi-Attention Fusion contain residual connection [80] and layer normalization [75]. Self-Multi-Attention Fusion module estimates how salient correlations exist between elements of a single modality and extracts sympathetic representation. The Self-Multi-Attention fusion can be described as follows:

$$\begin{aligned} \varepsilon(X) &= W_m\left[head_1, \ldots, head_m\right] + b_m, \\ head_i &= softmax\left(\frac{QK^T}{\sqrt{m}}\right)V, \\ &= softmax\left(\frac{\left(XW_{Q_i}\right)\left(XW_{K_i}\right)^T}{\sqrt{m}}\right)XW_{V_i}. \end{aligned} \tag{10}$$

We define the query as $Q = XW_Q$, the key as $K = XW_{K_i}$, and the value as $V = XW_{V_i}$, where, $W_{Q_i} \in \mathbb{R}^{f \times f}$, $W_{K_i} \in \mathbb{R}^{f \times f}$, $W_{V_i} \in \mathbb{R}^{f \times f}$, $W_m \in \mathbb{R}^{f \times f}$ and b_m are weights. Specifically, the *softmax* function measures the attention given for the time step. Hence, the time step of $head_i$ is a weighted summary of V. As shown in Figure 2, the outputs of the m attention heads are concatenated together and followed by

a linear layer.

$$Z = LayerNorm(X + \varepsilon(X)). \tag{11}$$

$$H = LayerNorm(X + Feedforward(Z)). \tag{12}$$

Finally, the entire model stacks N_{Layers} in a similar way to the BERT layers and the final hidden state of the first token (i.e., [*CLS*]) is fed to a linear transformation function for classification.

C. MULTI-ATTENTION FUSION

We introduce a multimodal fusion approach using BERT and audio and visual influenced representations. The Multi-Attention Fusion module automatically models the rich interactions between audio-visual representations (Video Fusion module: Audio-Text and Visual-Text) and BERT models. The Video Fusion module is described in the following subsection. The i-*th* head attention takes the form of Equation 13.

$$\varepsilon'(Q, K_{BERT}, V_{BERT}) = W'_m\left[head'_1, \ldots, head'_m\right] + b'_m,$$
$$head'_i = softmax\left(\frac{\left(QW'_{Q_i}\right)\left(K_{BERT}W'_{K_i}\right)^T}{\sqrt{m}}\right) \times V_{BERT}W'_{V_i}. \tag{13}$$

where W'_{Q_i}, W'_{K_i}, W'_{V_i}, $W'_m \in \mathbb{R}^{f\times f}$ and b'_m are parameters. Similarly to the BERT model, the N_{layers} are stacked to obtain the final representation, by connecting the feed-forward layer and residual connections. Finally, we use the final hidden state of the [*CLS*] token, to supply target-oriented emotions to the linear function for classification. All modules (i.e., the Video Fusion and fine-tuned BERT model) are trained simultaneously, to ensure that the model can learn the emotional content of each utterance. Furthermore, the Multi-Attention fusion module computes greater weightings for the BERT representation.

D. VIDEO FUSION

As shown in the upper-right corner of Figure 2, we describe the Video Fusion module. We added a Hadamard product operation to the concatenation, unlike the usual concatenation method. The two representations are then fused together as:

$$rep_t = [\alpha_t; \beta_t; \alpha_t \odot \beta_t]\, W_{rep} + b_{rep}. \tag{14}$$

where $\odot$ denotes Hadamard product, $W_{rep} \in \mathbb{R}^{3d\times d}$ and $b_{rep} \in \mathbb{R}^d$ are trainable weights and bias. After collecting rep_t from all time steps, we get $rep_t \in \mathbb{R}^{T\times d}$. The Video Fusion module shows its effectiveness in our ablation studies in the next section.

V. EXPERIMENTS

In this section we outline the experiments in this paper. We first start by describing the datasets, followed by presenting the experimental details.

A. DATASETS

We perform the experiments on three state-of-the-art benchmarking video sentiment analysis datasets: CMU-MOSI [38], CMU-MOSEI [39] and IEMOCAP [40]. All of the datasets include audio,text, and video modalities.

1) CMU-MOSI

CMU-MOSI consists of 2,199 short monologue video clips, examples of YouTube movie reviews specifically for multimodal emotions and emotion recognition. Unlike common emotion labels, such as happiness, anger, sadness, etc., the emotions of each sentence are annotated with the scale of emotion within the range of [−3,3], and marked from extremely negative sentiment −3 to means extremely positive +3. After dividing the video into utterances, we use 1,284 utterances as training set, 229 utterances as validation set, and 686 utterances as test set, keeping it consistent with prior works.

2) CMU-MOSEI

Similar to CMU-MOSI, this dataset is from YouTube but it is larger. CMU-MOSEI is the largest multimodal analysis data made up of 22,777 movie review videos. Additionally, that data consists of 22,856 annotated utterances. CMU-MOSEI is annotated with various emotion scores ranging from −3 to +3, identical to the CMU-MOSI dataset. To compare the models with the two datasets CMU-MOSI and CMU-MOSEI, we follow the latest prior work [32], [33], [35], using binary accuracy (i.e., Acc-2: positive or negative sentiments), seven class accuracy (i.e., Acc-7: sentiment score classification), F1-score, Mean Absolute Error (MAE), and Correlation Coefficient (Corr) as the main evaluation metrics. Specifically, when the algorithm is trained, the prediction score is the nearest integer from the set of integers −3 to +3, which classifies the data into 7 classes. In all metrics, the lower score for the MAE is the better performance and the higher the value excluding the MAE, the higher the performance. In the following experimental table, we indicate that (h) the higher the notation, the better the (l) the lower the better. Each utterance is treated as a separate multimodal example, and the training and test sets contain 16,326, 1,871, and 4,659, respectively, as in the previous work.

3) IEMOCAP

IEMOCAP is an acted multimodal emotion dataset has five sessions and each session consists of 2 actors (one male and one female) which contains 12 h data with the entire 10 actors to record the different emotions like anger, disgust, fear, sadness, neutral, happiness and excited. We selected and evaluated four commonly used emotion categories: happy, sad, anger, and neutral. We followed the experimental procedure and evaluation indicators of previous studies [32], [33], [35] and provide a comparison of model performance with other State-of-the-art models for binary accuracy (Acc-2) and F1-score. We use 4,290 utterances as training

set, 2,124 utterances as validation set, and 1,208 utterances as testing set.

B. EXPERIMENTAL DETAILS

We use a pre-trained language model from BookCorpus [81] and English Wikipedia, based on the BERTbase model published by Devlin *et al.*, [65]. The BERTbase model is specifically a model with 12 layers of transformer blocks with each block having a hidden state size of 768 and 12 attention heads and we use the same hyperparameters. Our proposed Self-Multi-Attention fusion and Multi-Attention fusion modules consist of 3 attention blocks and 4 attention heads. We use dropouts of 0.3 for training each module. Our proposed framework easily suffers from overfitting because the size of the dataset is limited. To prevent the overfitting problem, in this paper we endeavor to seek help from other auxiliary tasks. We introduce data normalization, dropout, and layer normalization as regularization methods to avoid overfitting. The dropout value is set to 0.3. We trained using the Adam optimizer [82] with an initial learning rate of 1e-5 and used a linear decay learning rate schedule with warm-up. The model was trained with a batch size of 48 for 1000 epochs and saved when the validation loss did not decrease during training. IEMOCAP dataset is trained for 100 epochs. For parameter optimization in CMU-MOSI and CMU-MOSEI, the loss function is set as the Mean Absolute Error (MAE). Additionally, IEMOCAP dataset the predicted loss function as categorical cross-entropy. The model is trained with batch-size 48 for 1000 epochs. All data is normalized to 50 equal to the length of the longest text sentence. We run our model using Pytorch, and it is trained and evaluated using two NVIDIA GeForce GTX 2080TI (11 GB memory) GPU systems.

VI. RESULTS AND DISCUSSION

In this section we outline the experiments for performance evaluation. In our approach, audio, visual and text features are taken into account to improve the recognition performance. First, we describe the analysis of features extracted from audio and visual. We experiment with the database described in Section V with the CMU-MOSI, CMU-MOSEI and IMOCAP database. Audio and Visual features are passed through using the Self-Multi-Attention Fusion module to analyze the performance. Selected heterogeneous features are entered into the module and fused by simply ''concatenation'' all features. Then, the baseline and multimodal results are described.

A. AUDIO FEATURE-LEVEL FUSION ANALYSIS

The performance of the individual features introduced in Section III is compared with that when the features are combined at the feature-level. In the experiment, select audio features including LLDs, VGGish, and SoundNet. Tables 1 to 3 show experimental results for audio features. In each table, experiment numbers 1 to 3 are single audio features and 4 to 7 are the result of combining single features.

TABLE 1. Comparison of the proposed audio features on CMU-MOSI dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Num	Features	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
1	LLDs	0.571	0.1786	0.4791	1.464	0.0293
2	SoundNet	0.5755	0.2031	0.5116	1.494	0.1202
3	VGGish	0.567	0.2295	0.5557	1.475	0.16
4	SoundNet + VGGish	0.5755	0.2168	0.5468	1.477	**0.162**
5	**LDDs + VGGish**	**0.5786**	**0.2305**	**0.5563**	**1.426**	0.1512
6	LLDs + SoundNet	0.57	0.1160	0.4311	1.458	0.1148
7	LLDs + VGGish + SoundNet	0.570	0.1145	0.4772	1.438	0.0834

TABLE 2. Comparison of the proposed audio features on CMU-MOSEI dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Num	Features	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
1	LLDs	0.7061	0.3491	0.6927	0.9127	0.4091
2	SoundNet	0.6627	0.3216	0.6522	0.9664	0.3014
3	VGGish	0.7290	0.359	**0.7186**	0.8866	0.4464
4	SoundNet + VGGish	0.7231	0.3289	0.7136	0.8879	0.4516
5	**LDDs + VGGish**	**0.7301**	**0.3684**	0.7166	**0.8272**	**0.4712**
6	LLDs + SoundNet	0.7198	0.3383	0.7088	0.8894	0.4476
7	LLDs + VGGish + SoundNet	0.7293	0.3287	0.7176	0.8883	0.4566

Table 1 compares the performance of the proposed features in the CMU-MOSI dataset. LLDs+VGGish, a combination of handcraft features and deep learning features, outperforms other features except for Corr. Corr shows better results with the combination of SoundNet+VGGish features. It shows that in an audio sentiment recognition experiment, handcraft features achieve competitive performance compared to pre-trained deep learning features. For example, in Table 1, the handcraft features show competitive performance compared to the pre-trained deep learning features, and LLDs achieve higher performance at Acc-2 than VGGish. However, the method of combining all features, LLDs+VGGish+SoundNet shows lower performance than LDDs+VGGish, and it is confirmed that there exist effective features according to data. Table 2 shows the results of the CMU-MOSEI dataset. From the experimental results of the CMU-MOSEI dataset, we can see that Num.5 LLDs+VGGish performs better in most cases. CMU-MOSEI has similar data characteristics to CMU-MOSI, and has higher performance because the data size is larger than CMU-MOSI. We show the efficiency of heterogeneous features combining handcraft features LLDs and bottleneck features VGGish in CMU-MOSI and CMU-MOSEI dataset. Table 3 shows the performance results for the IEMOCAP dataset. Unlike the experimental results in Tables 1 and 2, IEMOCAP data shows that SoundNet+ VGGish has better performance in most cases (Happy, Angry and Neutral). SoundNet achieves better performance of the results of Acc-2 and F1-score of 0.8566 and 0.8453 respectively in Sad sentiment. In IEMOCAP dataset, it is shown in performance that the functions of SoundNet and VGGish play an important role in emotion recognition. We confirm that it is necessary to select and fuse effective features according to the dataset rather than increasing the performance by fusion of a lot of multiple features.

TABLE 3. Comparison of the proposed audio features on IEMOCAP dataset. We use the binary accuracy(Acc-2) and F1-score metric for each of the four emotions.

	Task	Happy		Angry		Sad		Neutral	
Num	Features	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)
1	LLDs	0.7933	0.7164	0.9333	0.9011	0.8	0.74	0.5566	0.5388
2	SoundNet	0.7959	0.7055	0.9291	0.9099	**0.8566**	**0.8453**	0.6340	0.6180
3	VGGish	0.8094	0.7774	0.9443	0.9439	0.8128	0.8157	0.6762	0.6725
4	**SoundNet + VGGish**	**0.8145**	**0.7939**	**0.9460**	**0.9479**	0.8499	0.8459	**0.6863**	**0.6830**
5	LDDs + VGGish	0.7959	0.7783	0.9376	0.9333	0.8414	0.8422	0.6694	0.6635
6	LLDs + SoundNet	0.79	0.6973	0.9333	0.9011	0.7966	0.7065	0.48	0.3171
7	LLDs + VGGish + SoundNet	0.7933	0.7051	0.9266	0.8978	0.83	0.8191	0.5566	0.5566

TABLE 4. Comparison of the proposed visual features on CMU-MOSI dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Num	Features	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
1	OpenFace	0.6122	0.226	0.6067	**1.457**	**0.2279**
2	VGG16	0.5984	0.2137	0.5899	1.5202	0.1879
3	ResNet50	0.5587	0.2015	0.5583	1.517	0.2012
4	**OpenFace + VGG16**	**0.6153**	**0.2321**	**0.6073**	1.514	0.1876
5	OpenFace + ResNet50	0.5404	0.1938	0.5431	1.601	0.1514
6	VGG16 + ResNet50	0.5832	0.2244	0.5431	1.498	0.1989
7	OpenFace + VGG16 + ResNet50	0.6076	0.2152	0.5399	1.5444	0.1898

TABLE 5. Comparison of the proposed visual features on CMU-MOSEI dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Num	Features	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
1	OpenFace	0.6984	0.3531	0.6858	0.9171	0.2039
2	VGG16	0.7024	0.3499	0.6855	0.928	0.4296
3	ResNet50	0.6925	0.3454	0.6706	0.9706	0.3969
4	**OpenFace + VGG16**	**0.7168**	**0.3641**	**0.71**	**0.8857**	**0.4703**
5	OpenFace + ResNet50	0.7106	0.3128	0.7013	0.9034	0.4158
6	VGG16 + ResNet50	0.6953	0.3428	0.4018	0.9556	0.4018
7	OpenFace + VGG16 + ResNet50	0.7131	0.3233	0.6982	0.9029	0.4094

TABLE 6. Comparison of the proposed visual features on IEMOCAP dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

	Task	Happy		Angry		Sad		Neutral	
Num	Features	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)
1	OpenFace	0.7959	0.7055	**0.9392**	**0.9098**	0.7689	0.6685	0.5025	0.346
2	VGG16	0.7959	0.7055	**0.9392**	**0.9098**	0.7689	0.6685	0.4957	0.3286
3	ResNet50	0.801	0.7173	0.9325	0.9093	0.7487	0.6667	0.4789	0.4103
4	OpenFace + VGG16	0.7959	0.7955	**0.9392**	**0.9098**	0.7706	0.6785	0.602	0.5547
5	OpenFace + ResNet50	0.7942	0.7046	**0.9392**	**0.9098**	0.7689	0.6685	0.6188	0.6187
6	VGG16 + ResNet50	0.7959	0.7055	**0.9392**	**0.9098**	0.7689	0.6685	0.5261	0.4835
7	**OpenFace + VGG16 + ResNet50**	**0.8347**	**0.812**	**0.9392**	**0.9098**	**0.8113**	**0.8104**	**0.6509**	**0.6361**

B. VISUAL FEATURE-LEVEL FUSION ANALYSIS

In this experiment, we employ visual features including OpenFace, VGG16, and ResNet50. Table 4 compares the performance of visual features on the CMU-MOSI dataset. The experimental results of the CMU-MOSI dataset in Table 4 show that OpenFace+VGG16 has better performance in Acc-2, Acc-7, and F1-score. MAE and Corr are shown to perform better with OpenFace handcraft features of 1.457 and 0.2279, respectively. Table 5 shows the experimental results for visual features by applying the CMU-MOSEI dataset. In the experimental results of the CMU-MOSEI dataset, we showed that OpenFace+VGG16 performed better in all metrics. Table 6 presents the experimental results from the IEMCOAP dataset for visual features. Unlike the experimental results in the previous table, the visual features of the IEMOCAP dataset has better performance in OpenFace+VGG16+ResNet50. In subsequent experiments, we compare how the multimodal mechanism works when keeping the high-performing features in a frozen state for each dataset.

C. BASELINE METHODS

We consider a variety of state-of-the-art models and comparisons for multimodal comparison.

- **MFN** (Zadeh *et al.*, 2018) [83] synchronizes multimodal sequences using a multi-view gated memory that stores intra-view and cross-view interactions over time.

- **RAVEN** (Wang *et al.*, 2019) [34] did language modeling by shifting word expressions with non-verbal behaviors (audio and visual).
- **MCTN** (Pham *et al.*, 2019) [37] proposed to learn robust multimodal representation using the Seq2Seq model by translating between modalities.
- **MulT** (Tsai *et al.*, 2019) [32] modeled a multimodal sequence by using an attention based cross modal transformers and combines their output in a late fusion manner to model a multimodal sequence.
- **ICCN** (Sun *et al.*, 2020) [35] used deep canonical correlation analysis (DCCA) to fuse audio and video into an outer product centered on text to determine the correlation, and then tested against multimodal embedding algorithms.
- **MFRM** (Mai *et al.*, 2020) [33] proposed a residual memory network (RMN), and time-step level fusion was introduced to model time-restricted interactions among modalities.
- **Human** [84] was asked to predict the sentiment score of each opinion utterance from CMU-MOSI dataset and human performance was recorded.

D. MULTIMODAL RESULTS

To evaluate the effectiveness of our proposed method, we compare the proposed approach with the state-of-the-art multimodal algorithm mentioned above. Table 7 to 9 shows the performance of our model compared to the state-of-the-art models. Our model for CMU-MOSI dataset in Table 7 outperforms the current state-of-the-art methods across most evaluation metrics. In the experimental results of CMU-MOSI dataset, the proposed HFU-BERT model shows 0.08 and 0.1153 performance improvements in Acc-2 and Acc-7, respectively, compared to that of the lowest state-of-the-art performance. However, a gap is still observed between HFU-BERT and human, showing space for further improvement. Table 8 shows the performance of our model compared to the previous model in the CMU-MOSEI dataset. Compared to the baseline model, our model outperforms each in Acc-2, F1-score, and Corr. However, in Acc-7 and MAE, ICCN shows better performance results at 0.5158 and 0.565, respectively, but the difference is marginal by about 0.02. Table 9 shows the superior accuracy and F1-score for each sentiment compared to previous work on the IEMCOAP dataset. MFRM performs better in happy emotion than our model, but the difference is subtle.

VII. ABLATION STUDIES

In order to analyze the usefulness of the various fusions of the HFU-BERT model, we consider three questions:

Q1:What is the effect of changing the number of modalities while training an HFU-BERT model?

Q2: When fusing with each modality in the HFU-BERT model, what is the effect of changing the position of representations fusion?

Q3: What is the effect of changing the pre-trained BERT?

TABLE 7. Comparison between HFU-BERT and other state-of-the-art algorithms on CMU-MOSI dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Model Descrition	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
MFN	0.774	0.341	0.773	0.965	0.632
RAVEN	0.78	0.332	0.766	0.915	0.691
MCTN	0.793	0.356	0.791	0.909	0.676
MulT	0.811	0.391	0.810	0.889	0.686
ICCN	0.8307	0.3901	0.8302	0.862	0.714
MFRM	0.823	0.394	0.825	0.896	0.697
HFU-BERT(ours)	**0.8549**	**0.4473**	**0.8545**	**0.7223**	**0.7926**
Human	0.857	-	0.875	0.71	0.82

TABLE 8. Comparison between HFU-BERT and other state-of-the-art algorithms on CMU-MOSEI dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Model Descrition	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
MFN	0.798	0.49	0.8	0.612	0.667
RAVEN	0.798	0.496	0.806	0.607	0.67
MCTN	0.791	0.5	0.795	0.614	0.662
MulT	0.816	0.507	0.816	0.591	0.694
ICCN	0.844	**0.5158**	0.8415	**0.565**	0.713
MFRM	0.824	0.509	0.826	0.598	0.69
HFU-BERT(ours)	**0.8629**	0.4941	**0.8623**	0.5801	**0.7977**

Q4: Is it better to use HFU-BERT's fusion method than to use other fusion methods?

To highlight the importance of the number of modalities, a series of ablation studies on Q1 is conducted using the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets. However, Q2, Q3, and Q4 performed corresponding ablation studies using only the CMU-MOSEI dataset. CMU-MOSEI dataset was chosen because it has the highest number of training examples compared to other datasets. In addition, we compared and evaluated using Acc-2 and F1-score in ablation studies from Q2 to Q4.

A. COMPARISON OF THE EFFECTS OF EACH MODALITY FUSION

To explore the underlying information of each modality, we carry out an experiment to compare the performance among unimodal, bimodal and trimodal(HFU-BERT) systems. Table 10 shows the ablation studies for CMU-MOSI dataset. We perform audio and visual features by applying a combination of the aforementioned high-performance features. In Table 10, A is the audio modality, an audio feature that combines LLDs and VGGish features. V is the visual modality, and it is a video feature that combines OpenFace+VGG16 features. T is the text modality extracted by the pre-trained BERT model. Unimodal showed that text dominates over audio and visual. In bimodal, it can be seen that A+V without text has much lower performance than those with text, and T+V has the high performance. Our HFU-BERT showed the highest performance in Acc-2, Acc-7, F1-score, and MAE by fusion of A+V+T. Table 11 shows the ablation studies for CMU-MOSEI dataset. The features of A, V and T are the same as those of

TABLE 9. Comparison between HFU-BERT and other state-of-the-art algorithms on IEMOCAP dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Task	Happy		Angry		Sad		Neutral	
Model Description	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)
MFN	0.865	0.84	0.85	0.837	0.835	0.821	0.696	0.692
RAVEN	0.873	0.858	0.873	0.867	0.834	0.831	0.697	0.693
MulT	0.848	0.819	0.739	0.702	0.777	0.741	0.625	0.597
ICCN	0.8741	0.8472	0.8862	0.8802	0.8626	0.8593	0.6973	0.6847
MFRM	**0.876**	**0.859**	0.894	0.894	0.861	0.854	0.707	0.698
HFU-BERT (ours)	0.8617	0.8552	**0.9392**	**0.9357**	**0.8819**	**0.8822**	**0.7655**	**0.7649**

TABLE 10. Unimodal, bimodal and trimodal results of HFU-BERT on CMU-MOSI dataset. Here, "L" means text modality, "A" denotes audio modality, and "V" represents visual modality. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Methods	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
Unimodal					
A	0.5786	0.2305	0.5563	1.426	0.1512
V	0.7301	0.3684	0.7166	0.8272	0.4712
T	0.8412	0.4334	0.7370	0.794	0.8410
Bimodal					
T+V	**0.8549**	0.4290	0.8542	0.7298	**0.7962**
T+A	0.8434	0.3816	0.8413	0.8434	0.7524
A+V	0.577	0.2	0.57532	1.5352	0.1497
Trimodal (HFU-BERT)					
A+V+T	**0.8549**	**0.4473**	**0.8545**	**0.7223**	0.7926

TABLE 11. Unimodal, bimodal and trimodal results of HFU-BERT on CMU-MOSEI dataset. Here, "L" means text modality, "A" denotes audio modality, and "V" represent visual modality. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Methods	Acc-2(h)	Acc-7(h)	F1-score(h)	MAE(l)	Corr(h)
Unimodal					
A	0.7301	0.3684	0.7166	0.8272	0.4712
V	0.7168	0.3641	0.71	0.8857	0.4703
T	0.849	0.4657	0.8490	0.6125	0.7650
Bimodal					
T+V	0.8587	0.4711	0.8555	0.5849	0.7934
T+A	0.8595	0.4807	0.8589	0.5779	0.7956
A+V	0.7321	0.3564	0.7209	0.8677	0.4736
Trimodal (HFU-BERT)					
A+V+T	**0.8629**	**0.4941**	**0.8623**	**0.5801**	**0.7977**

CMU-MOSI dataset. In unimodal, T showed higher performance than A and V. In the case of bimodal, in Table 11, T+A achieves the best performance and A+V shows the lowest performance. The comparison of T+V, T+A, and A+V shows that T has a dominant effect on emotion recognition. In the CMU-MOSEI dataset, our HFU-BERT fused A+V+T outperformed all performance. Table 12 shows the ablation studies for IEMOCAP dataset. A is the audio modality, which is an audio feature that combines SoundNet and VGGish features. V is the visual modality, and it is a video feature that combines OpenFace+VGG16+ResNet50 features. T is text modality and is a feature extracted by the pre-trained BERT model. In unimodal, most of the A showed high performance (angry, sand, neutral). T+A, which combines text and audio in bimodal, has high performance in most cases (happy, angry, sad). Our HFU-BERT showed the highest performance in happy, angry, and sad by fusion of A+V+T.

B. COMPARISON OF THE CENTRAL MODALITY INFLUENCE OF HFU-BERT MODEL

We perform a comparison of the influence of the central modality in the HFU-BERT model. As shown in Figure 2, the proposed HFU-BERT model is fused in the audio and visual representations and the Video Fusion module, centering on the text modality. As shown by Equation 13 in the Multi-Attention Fusion module of HFU-BERT, two alternative fusion strategies (video focused, audio focused, and BERT focused) are applied to the attention mechanism by taking different value combinations of the query, key, and value. Our experiment compares the impact of robust modalities by changing these central modalities. Figure 4 shows the impact of the fusion performance between modalities. HFU-BERT (robust-text) is robustly

TABLE 12. Unimodal, bimodal and trimodal results of HFU-BERT on IEMOCAP dataset. Metric used is Acc-2: binary accuracy, Acc-7: seven class accuracy, F1-score, MAE: mean absolute error, and corr: correlation coefficient.

Task	Happy		Angry		Sad		Neutral	
Methods	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)	Acc-2(h)	F1-score(h)
Unimodal								
A	0.8145	0.7939	0.9360	0.9179	0.8499	0.8459	0.6863	0.6830
V	0.8431	0.8326	0.9392	0.9119	0.8333	0.815	0.661	0.6573
T	0.86	**0.8587**	0.9207	0.9167	0.8448	0.8427	0.7554	0.7553
Bimodal								
T+V	0.8503	0.8484	0.9258	0.9248	0.8549	0.8534	**0.803**	**0.8884**
T+A	0.855	0.8478	0.9274	0.9238	0.8634	0.8628	0.7622	0.7618
A+V	0.7959	0.7055	0.9376	0.9090	0.8684	0.8651	0.6475	0.6448
Trimodal (HFU-BERT)								
A+V+T	**0.8617**	0.8552	**0.9392**	**0.9357**	**0.8819**	**0.8822**	0.7655	0.7649

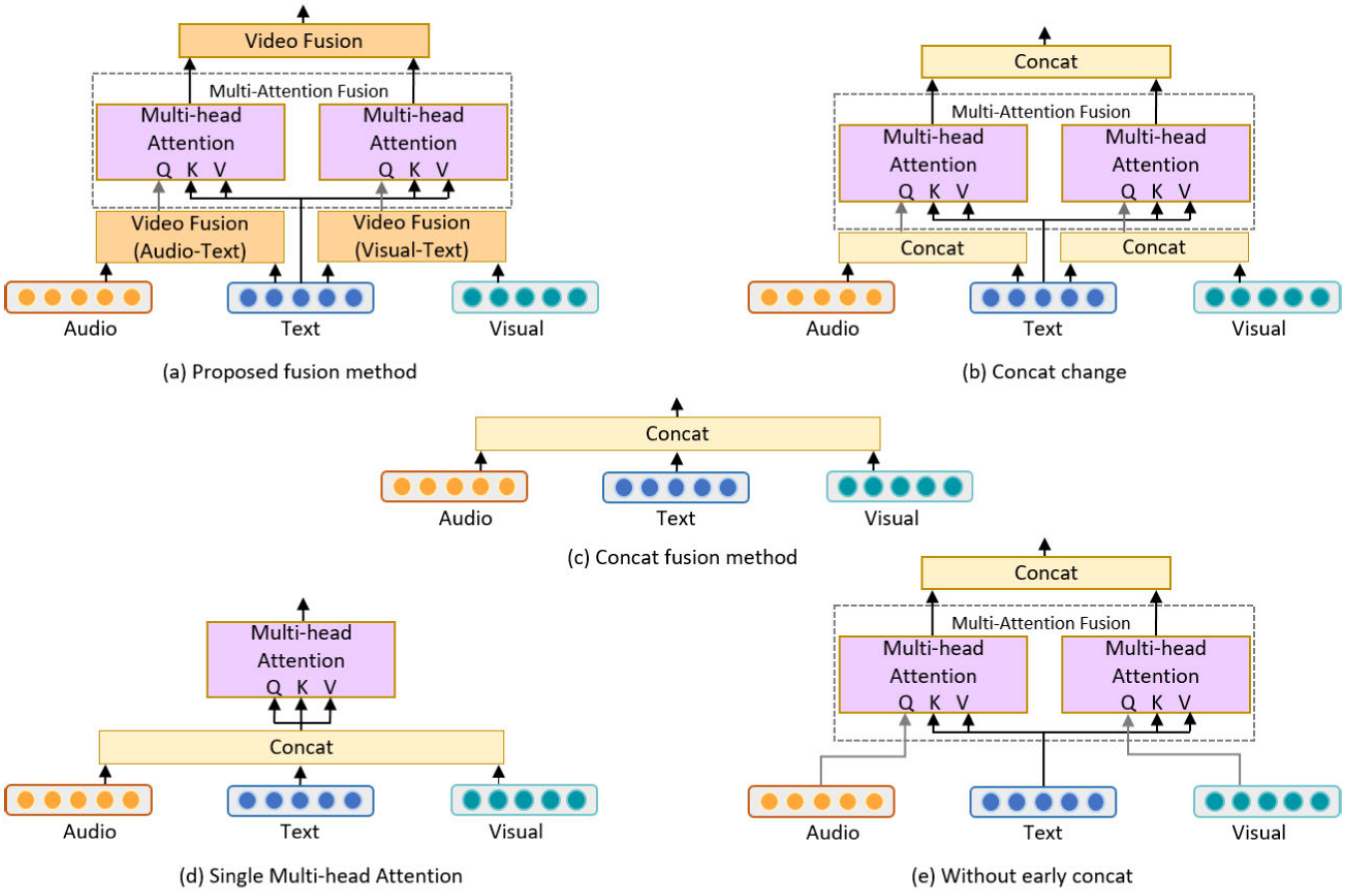

FIGURE 3. Comparison of ablation studies for various fusion methods. (a) is our proposed HFU-BERT fusion method. (b) is a fusion method that replaced video fusion module with concatenate in HFU-BERT. (c) fuses three modalities in a concatenate method. (d) is the method using single multi-head attention. (e) is the fusion method excluding early concatenate.

designed with BERT representation as a central position in the Video Fusion module and Multi-Attention Fusion module. Similarly, HFU-BERT (robust-audio) and HFU-BERT (robust-visual) are robustly designed with each representation as a central position for the Video Fusion module and the Multi-Attention Fusion module. We show that the robust-text (Acc-2: 0.863, F1-score: 0.862) is more robust than the robust-visual (Acc-2: 0.859, F1-score: 0.858) and robust-audio (Acc-2: 0.854, F1-score: 0.853) modalities centered fusion. It is essential to highlight that the modality with only BERT representation performs significantly better than other modalities.

C. COMPARISON OF THE EFFECTS OF PRE-TRAINED BERT IN THE HFU-BERT MODEL

We compare the impact of pre-trained BERT on text representation in the HFU-BERT model. To compare with the BERT model, we use publicly available 300-dimensional word2vec [20] vector trained on 100 billion words from Google News. As shown in Figure 6, the HFU-BERT model we proposed proceeds by changing BERT in text representation. Compare experiments using three methods:

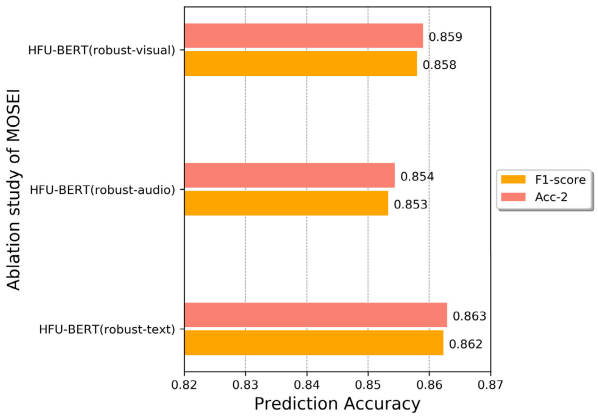

FIGURE 4. Comparison of modality effects in MOSEI according to three fusion strategies in HFU-BERT.

We use the fine-tuning model HFU-BERT (Fine-tuning BERT) and the model HFU-BERT (Non-pretrained BERT) using only the BERT structure, not the pre-trained BERT, and the HFU-BERT (Word2vec) applying word2vec. In Figure 6, HFU-BERT (Word2vec) has better performance

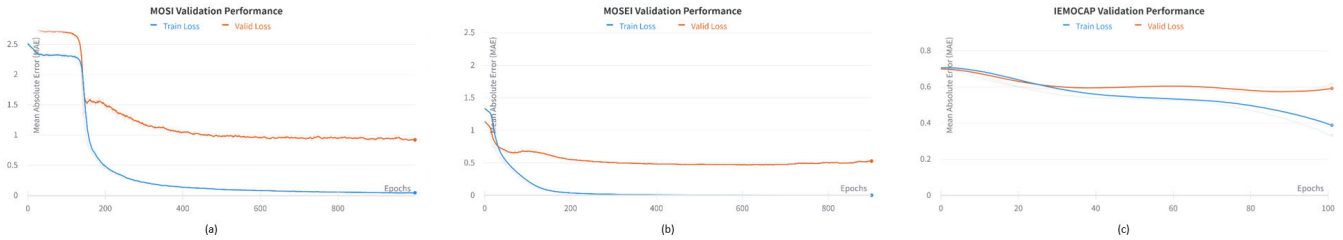

FIGURE 5. Validation set convergence of HFU-BERT model. (a), (b) and (c) show the calculated losses for both training and validation processes in CMU-MOSI, CMU-MOSEI and IEMOCAP datasets.

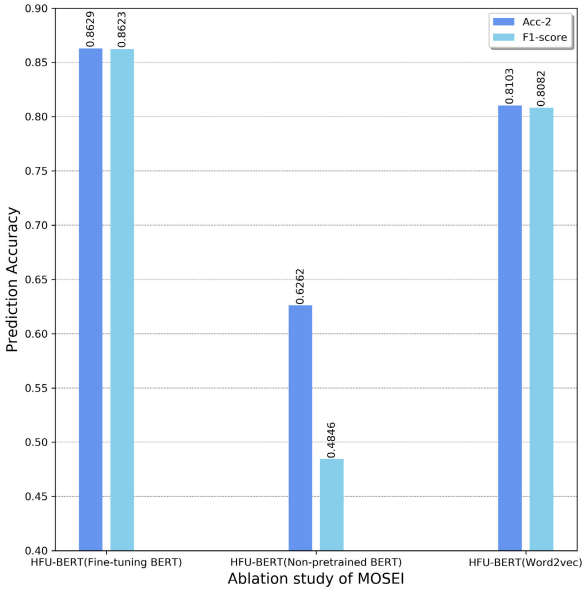

FIGURE 6. Comparison of the effect of fine-tuning in the HFU-BERT model using CMU-MOSEI dataset.

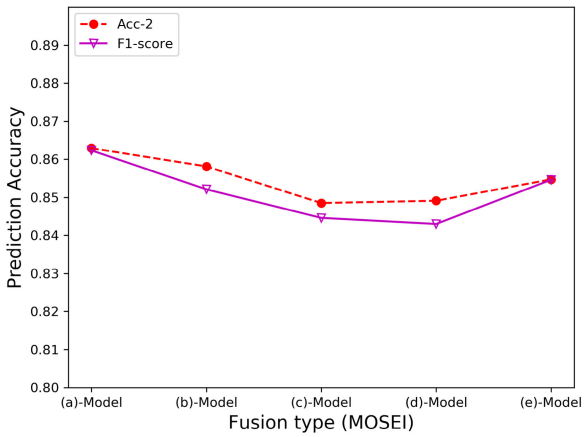

FIGURE 7. Comparison of various fusion methods on MOSEI dataset.

than HFU-BERT (Non-pretrained BERT) and shows the effect of pre-training model. However, when compared to the same two pre-trained models, HFU-BERT (Fine-tuning BERT) shows higher performance than HFU-BERT (Word2vec). In our experiment, HFU-BERT (Fine-tuning BERT) showed the highest performance, and BERT showed the effect of fine-tuning.

D. COMPARISON WITH THE OTHER FUSION METHODS

We use several models to test our design decisions using CMU-MOSEI dataset. Consists of five major fusions to perform multimodal fusion, as shown in Figure 3: (a) Proposed fusion method, (b) Concat change, (c) Concat fusion method, (d) Single Multi-head Attention, (e) without early concat. (a) is the fusion method proposed by us, and (b) is the fusion method in which the Video Fusion module is changed to concatenate. Here we can compare the effects of the Video Fusion module. (c) is a fusion method that simply applies concatenate using three basic modalities. (d) is a model when only one Multi-head Attention is used, and can be compared with the two existing Multi-head Attentions of (b). (e) is the part directly applied to Multi-head Attention without early concatenate, and the effect of concatenate can be confirmed compared to (b). The results of the fusion design are shown in Figure 7. (a) is Acc-2(0.8629) and F1-score(0.8623), and (b) is Acc-2(0.8581) and F1-score(0.8521) to show the effect of Video Fusion (a) in the two comparisons. The simple fusion methods of (c) are Acc-2 (0.8485) and F1-score (0.8446), and Acc-2 showed the lowest performance. (d) Acc-2 (0.8491), F1-score (0.843) is compared with (b), we can see that the proposed Multi-Attention Fusion two Multi-head Attentions is effective. (e) is Acc-2 (0.8547) and F1-score (0.8546). Compared to (b), early connection was effective, but the effect was marginal. The proposed HFU-BERT fusion method showed superior performance compared to other fusion methods.

E. COMPARISON OF VALIDATION SET CONVERGENCE IN HFU-BERT MODEL

Figure 5 shows the loss values for the training and validation process of the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets of the HFU-BERT model. (a) and (b) select the best predictive model at 554 and 196 epochs, respectively, when repeated for 1000 epochs. (c) is a model trained on IEMOCAP dataset, reaching the best performance at 87 epochs when repeated for 100 epochs. In the experiments, several effective methods have been used to reduce

overfitting, but overfitting could not be completely avoided. The CMU-MOSEI dataset (b) with the most training data converges with better performance. Therefore, recording a large number of multimodal emotion databases is very important to facilitate development in this research topic.

VIII. CONCLUSION

In this paper, we proposed a HFU-BERT model that effectively fuses multimodal emotion recognition using pre-trained BERT model for multimodal languages and heterogeneous features unification for audio and visual. The proposed HFU-BERT integrated visual and acoustic modalities into heterogeneous features and was successfully fine-tuned using BERT. Our method was shown to exceed the state-of-the-art in three challenging benchmarks: CMU-MOSI, CMU-MOSEI, and IEMOCAP. In order to analyze the effect of each modal, an ablation study was performed on HFU-BERT. A potential limitation of our proposed model was increased computations due to the generation of more trainable weights and hyperparameters. Moreover, further research needs to be conducted to confirm the robustness of our proposed model. In future work, we will explore how to learn a better representation audio, visuals and text using models similar to BERT.

REFERENCES

[1] L. Marsili, L. Ricciardi, and M. Bologna, "Unraveling the asymmetry of Mona Lisa smile," *Cortex*, vol. 120, pp. 607–610, Nov. 2019.
[2] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proc. Interspeech*, Sep. 2018, pp. 932–936.
[3] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 835–838.
[4] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 190–195.
[5] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–5.
[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.
[7] M. Gao, J. Dong, D. Zhou, Q. Zhang, and D. Yang, "End-to-end speech emotion recognition based on one-dimensional convolutional neural network," in *Proc. 3rd Int. Conf. Innov. Artif. Intell. (ICIAI)*, 2019, pp. 78–82.
[8] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. W. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 1–5.
[9] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, p. 149.
[10] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–6.
[11] K. Sikka, G. Sharma, and M. Bartlett, "LOMo: Latent ordinal model for facial analysis in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5580–5589.
[12] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3703–3712.
[13] Y. Tang, "Deep learning using linear support vector machines," 2013, *arXiv:1306.0239*. [Online]. Available: http://arxiv.org/abs/1306.0239
[14] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, and M. Mirza, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.
[15] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 302–309.
[16] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.
[17] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? finding strong and weak opinion clauses," in *Proc. AAAI*, vol. 4, 2004, pp. 761–769.
[18] C. Yang, K. H.-Y. Lin, and H.-H. Chen, "Emotion classification using Web blog corpora," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Nov. 2007, pp. 275–278.
[19] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*. [Online]. Available: http://arxiv.org/abs/1310.4546
[21] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*. [Online]. Available: http://arxiv.org/abs/1510.03820
[22] M. Abdul-Mageed and L. Ungar, "EmoNet: Fine-grained emotion detection with gated recurrent neural networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 718–728.
[23] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*. [Online]. Available: http://arxiv.org/abs/1908.11540
[24] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5415–5421.
[25] W. Jiao, H. Yang, I. King, and M. R. Lyu, "HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition," 2019, *arXiv:1904.04446*. [Online]. Available: http://arxiv.org/abs/1904.04446
[26] S. Lee, D. K. Han, and H. Ko, "Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition," *Sensors*, vol. 20, no. 22, p. 6688, Nov. 2020.
[27] Y. Xu, H. Xu, and J. Zou, "HGFM : A hierarchical grained and feature model for acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6499–6503.
[28] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.
[29] X. Xia, J. Liu, T. Yang, D. Jiang, W. Han, and H. Sahli, "Video emotion recognition using hand-crafted and deep learning features," in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–6.
[30] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.
[31] I. R. Shaffer, "Exploring the performance of facial expression recognition technologies on deaf adults and their children," in *Proc. 20th Int. ACM SIGACCESS Conf. Comput. Accessibility*, Oct. 2018, pp. 474–476.
[32] 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.
[33] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Trans. Affect. Comput.*, early access, Jun. 8, 2020, doi: 10.1109/TAFFC.2020.3000510.
[34] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L. P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7216–7223.

[35] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 8992–8999.

[36] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3454–3466.

[37] 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 *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 6892–6899.

[38] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*. [Online]. Available: http://arxiv.org/abs/1606.06259

[39] 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 (Long Papers)*, vol. 1, 2018, pp. 2236–2246.

[40] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, p. 335, Dec. 2008.

[41] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*. [Online]. Available: http://arxiv.org/abs/1806.06176

[42] T. Baltrusaitis, 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. 2019.

[43] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[44] 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, pp. 3–14, Sep. 2017.

[45] Y. Wang, J. Wu, and K. Hoashi, "Multi-attention fusion network for video-based emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 595–601.

[46] N. Samadiani, G. Huang, B. Cai, W. Luo, C.-H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, Apr. 2019.

[47] A. Lazaridou, N. The Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," 2015, *arXiv:1501.02598*. [Online]. Available: http://arxiv.org/abs/1501.02598

[48] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 1–8.

[49] Y. Gu, X. Li, S. Chen, J. Zhang, and I. Marsic, "Speech intention classification with multimodal deep learning," in *Proc. Can. Conf. Artif. Intell.* Berlin, Germany: Springer, 2017, pp. 260–271.

[50] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 949–954.

[51] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," 2018, *arXiv:1805.00705*. [Online]. Available: http://arxiv.org/abs/1805.00705

[52] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 284–288.

[53] S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešič, and V. Štruc, "Towards efficient multi-modal emotion recognition," *Int. J. Adv. Robotic Syst.*, vol. 10, no. 1, p. 53, Jan. 2013.

[54] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 649–665.

[55] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[56] C. T. Duong, R. Lebret, and K. Aberer, "Multimodal classification for analysing social media," 2017, *arXiv:1708.02099*. [Online]. Available: http://arxiv.org/abs/1708.02099

[57] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 371–378.

[58] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*. [Online]. Available: http://arxiv.org/abs/1707.07250

[59] P. Pu Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," 2018, *arXiv:1808.03920*. [Online]. Available: http://arxiv.org/abs/1808.03920

[60] Z. Liu, Y. Shen, V. Bharadhwaj Lakshminarasimhan, P. Pu Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*. [Online]. Available: http://arxiv.org/abs/1806.00064

[61] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 873–883.

[62] Z. Han, H. Zhao, and R. Wang, "Transfer learning for speech emotion recognition," in *Proc. IEEE 5th Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput., (HPSC) IEEE Int. Conf. Intell. Data Secur. (IDS)*, May 2019, pp. 96–99.

[63] S. Parthasarathy, V. Rozgic, M. Sun, and C. Wang, "Improving emotion classification through variational inference of latent variables," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7410–7414.

[64] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," *Frontiers Comput. Sci.*, vol. 2, p. 9, Feb. 2020.

[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[66] D. Olson, "From utterance to text: The bias of language in speech and writing," *Harvard Educ. Rev.*, vol. 47, no. 3, pp. 257–281, Sep. 1977.

[67] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1–4.

[68] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," 2016, *arXiv:1610.09001*. [Online]. Available: http://arxiv.org/abs/1610.09001

[69] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[70] Z. Lian, Y. Li, J. Tao, and J. Huang, "Investigation of multimodal features, classifiers and fusion methods for emotion recognition," 2018, *arXiv:1809.06225*. [Online]. Available: http://arxiv.org/abs/1809.06225

[71] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6494–6498.

[72] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[73] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[76] P. Ekman and W. V. Friesen, *Facial Action Coding System: Investigator's Guide*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[77] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015.

[78] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762

[80] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450
[81] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 19–27.
[82] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[83] 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. 1–8.
[84] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.

SANGHYUN LEE received the B.S. degree from the School of Electrical Engineering, Hallym University, South Korea, in 2017. He is currently pursuing the dual M.S. and Ph.D. degrees in electrical engineering with Korea University, Seoul, South Korea. His research interests include multimodal-based emotion recognition, speech recognition, speech emotion recognition, deep learning, and its applications.

DAVID K. HAN received the B.S. degree from Carnegie Mellon University and the M.S.E. and Ph.D. degrees from Johns Hopkins University. After years of serving as a Scientist at NSWC and ONR, he joined the University of Maryland, in 2005, as a Faculty Member and the Deputy Director of CECD. From 2007 to 2009, he was the Distinguished IWS Chair Professor with the United States Naval Academy, Annapolis, MD, USA. In January 2009, he returned to ONR, as a Program Officer at the Ocean Engineering Team. From 2012 to 2014, he was the Deputy Director of research with ONR. From 2014 to 2016, he was an Associate Director of basic research in artificial intelligence (AI) and robotics with ASD (Research and Engineering). He is currently a Senior Scientist of AI with CISD, Army Research Laboratory.

HANSEOK KO (Senior Member, IEEE) received the B.S. degree in electrical engineering from Carnegie Mellon University, in 1982, the M.S. degree in electrical engineering from Johns Hopkins University, in 1988, and the Ph.D. degree in electrical engineering from CUA, in 1992. At the onset of his career, he was with WOL, Annapolis, MD, USA, where his work involved signal and image processing. In March 1995, he joined as a Faculty Member with the Department of Electronics and Computer Engineering, Korea University, where he is currently a Professor.