2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) | 979-8-3503-2745-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACIIW59127.2023.10388147

Mutual Cross-Attention in Dyadic Fusion Networks for Audio-Video Emotion Recognition

Jiachen Luo
Centre for Digital Music
Queen Mary U. of London
London, UK
jiachen.luo@qmul.ac.uk

Huy Phan
Amazon Alexa
Cambridge, MA, USA
huypq@amazon.co.uk

Lin Wang
Centre for Digital Music
Queen Mary U. of London
London, UK
lin.wang@qmul.ac.uk

Joshua Reiss
Centre for Digital Music
Queen Mary U. of London
London, UK
joshua.reiss@qmul.ac.uk

***Abstract*—Multimodal emotion recognition is a challenging problem in the research fields of human-computer interaction and pattern recognition. How to efficiently find a common subspace among the heterogeneous multimodal data is still an open problem for audio-video emotion recognition. In this work, we propose an attentive audio-video fusion network in an emotional dialogue system to learn attentive contextual dependency, speaker information, and the interaction of audio-video modalities. We employ pre-trained models, *wav2vec*, and *Distract your Attention Network*, to extract high-level audio and video representations, respectively. By using weighted fusion based on a cross-attention module, the cross-modality encoder focuses on the inter-modality relations and selectively captures effective information among the audio-video modality. Specifically, bidirectional gated recurrent unit models capture long-term contextual information, explore speaker influence, and learn intra- and inter-modal interactions of the audio and video modalities in a dynamic manner. We evaluate the approach on the MELD dataset, and the experimental results show that the proposed approach achieves state-of-the-art performance on the dataset.**

***Index Terms*—affective computing, modality fusion, attention mechanism, deep learning**

I. Related Work

Multimodal emotion recognition in conversations is a crucial research topic in human-computer interactions [11-13]. To capture more effective emotion-relevant characteristics, multimodal fusion enables learning the internal correlation among heterogenenous multiple modalities for better emotion recognition. There are two main aspects to consider in any multimodal fusion model: identifying the best modality-specific features and effectively integrating the multimodal information [17-19].

Video data is a form of temporal data that encompasses multiple modalities. Prior research has concentrated on extracting features from acoustic, textual, and visual modalities to enhance multimodal emotion recognition in conversational settings [17-19]. Generally, feature engineering methods can be categorized into two classes: low-level handcrafted features and high-level abstract deep-learning representations [14,23].

Low-level features for audio, such as prosodic and spectral features, as well as their combination, are utilized to represent emotional features for discrimination purposes [23]. On the video end, appearance-based and geometry-based representations are employed to capture face motion parameters, including eye movements, eyebrow positions, and mouth movements [14]. However, these features do not take into account the high-level associations between them, which limits improvements in model performance. To address this issue, recent studies have utilized deep learning techniques to extract high-level representations from low-level features, leading to enhanced performance in this task [14,24].

To date, multimodal emotion recognition is roughly divided into four categories: early-level fusion [18], late-level fusion [25], score-level fusion [26], and model-level fusion [27], respectively. Specifically, model-level fusion is a compromise between the former two, where the fusion happens between the intermediate representations of the multimodal features [28]. With the popularity of model-level fusion, it has recently attracted attention in the field for its ability to fuse multimodal information. For example, Pini et al. proposed a multimodal fusion network to jointly merge static and dynamic features from different modalities into one representation [28].

While shallow fusion methods have demonstrated good performance in multimodal emotion recognition, they may not fully exploit the intricate non-linear relationships and joint distributions of multiple modalities. For instance, a simple feature concatenation may not suffice. Therefore, it is crucial to develop deep fusion models that can utilize multiple fusion operations to capture complex joint audio-visual features. To address this gap, we propose a dynamic fusion network that incorporates weighted fusion and attention mechanisms to integrate heterogeneous visual expressions with audio information.

II. Methodology

A. Problem Definition and Notation

Our goal is to accurately identify the emotion of constituent utterances presented in interactive conversations. Let us define a dialogue $U = [u_m^1, u_m^2, \ldots, u_m^N]$ in conversations, where u_m^j is the j^{th} utterance in the conversation, N is the total number of utterances. Specifically, $m \in \{a, v\}$, where a represents the audio modality and v represents the visual modality. The task is to predict the emotion e for each utterance u within a

Thanks to the China Scholarship Council and Queen Mary University of London for funding.

979-8-3503-2745-8/23/$31.00 ©2023 IEEE

finite set of emotions *E* (anger, disgust, sadness, joy, neutral, surprise and fear).

B. Acoustic Features

We employed the pre-trained *wav2vec large* model (Uncased: layer-24) acted as a suitable choice of a pre-trained model and feature extractor for emotion classification task [29]. This model comprises of a feature encoder and a context network. The feature encoder takes a raw waveform as input to encoder local speech's information, and then these are inputted to the Transformer-based context network to produce a contextualized representation. To convert frame-level representations produced by *wav2vec* into utterance-level representation, we used an average operation across the time dimension on the *wav2vec* embedding. In total, 1024 utterance-level acoustic features were extracted ($a^{wav2vec}$).

C. Visual Features

On the video end, we utilized the pre-trained *DAN* model as the video feature extractor for emotion recognition (v^{DAN}) [16]. *DAN* consists of three key components: Feature Clustering Network (*FCN*), Multi-head cross Attention Network (*MAN*), and Attention Fusion Network (*AFN*) [16]. The *FCN* captures robust features by employing a large-margin learning objective to maximize class separability. Additionally, the *MAN* employs multiple attention heads to simultaneously focus on multiple facial areas and generate attention maps for these regions. Furthermore, the *AFN* penalizes overlapping attentions and fuses the learned features. These frame-level features are then averaged across the time dimension to obtain utterance-level video features. In total, we extracted 1024 utterance-level video features (v^{DAN}).

D. Multimodal Attentive Fusion Network

By leveraging the potential of deep learning, we introduce a framework called *AVAFN* that incorporates an audio-visual fusion model (see Figure 2). Our proposed model is mainly comprised of two stages: 1) a pre-trained model is employed to extract high-level audio and video representations, 2) a fusion module is trained to jointly learn audio-visual learning features in a new common subspace, and model the interactions in the dialogue to make the emotional state prediction.

We assume that the emotion of an utterance in a conversation is strongly dependent on four major factors: 1) the context given by the preceding utterances, 2) the speaker, 3) the listener, and 4) the emotion behind the preceding utterances. In our fusion module, we employ bidirectional *GRU* cells to capture long-distance contextual information within conversations. These *GRU* cells take the input y_t and encode the hidden state from h_{t-1} to h_t as follows: $h_t = GRU(ht-1, yt)$. The updated hidden state h_t also serves as the output for the current step. Additionally, we utilize the other three branches of bidirectional *GRU* cells to model the speaker state, listener state, and emotion state. These states are essential for capturing contextual dependencies, speaker influence, and the emotional state of the participants.

1) Contextual State: Apparently, emotions are mainly dependent on the surrounding utterances in conversational emotion recognition. The preceding utterances are contextually related to the interaction of audio and video modalities. We design the cross-attention module (*CAM*) to maintain modality-specific patterns and capture interactions of different modalities, resulting in the interaction of audio-video modality utterance (u^{CA}) (see Figure 2).

The *CAM* takes audio (*a*) and video (*v*) as input to produce joint dynamic features in one contextual utterance. These play a crucial role in learning intra- and inter-modal perspectives according to the scaled dot-product attention mechanism. Technically, we estimate the associations between audio ($a^{wav2vec}$) and video (v^{DAN}) in a crossed way via the scaled dot-product attention function, whose query (Q_m), key (K_m), and value (V_m) are the representations of modality m, H_m, under different projection spaces, where $m \in a, v$. H_m is projected to query matrix (Q_m), key matrix (K_m), and value matrix (V_m) by linear projections without bias. The specific formula are as follows:

$$\zeta H_{v-a} = softmax(Q_a K_v^{\mathsf{T}} / \sqrt{d}) V_v, \tag{1}$$

$$\zeta H_{a-v} = softmax(Q_v K_a^{\mathsf{T}} / \sqrt{d}) V_a, \tag{2}$$

Where ζH_{a-v}, ζH_{v-a} denote the propagated information from audio to video and video to audio, respectively. Finally, we update the features of one modality with the propagate information from the other modality.

$$H_a^{LN} = LayerNorm(H_a \oplus \zeta H_{v-a}), \tag{3}$$

$$H_v^{LN} = LayerNorm(H_v \oplus \zeta H_{a-v}). \tag{4}$$

To further enhance the representation capacity, a feed-forward layer and layer normalization are employed behind the cross-attention layer:

$$H_a^c = LayerNorm(H_a^{LN} \oplus FeedForward(H_a^{LN})), \tag{5}$$

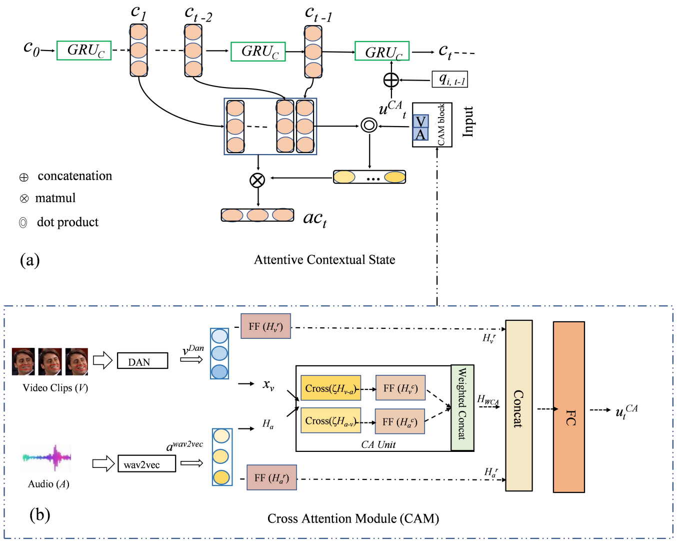
$$H_v^c = LayerNorm(H_v^{LN} \oplus FeedForward(H_v^{LN})). \tag{6}$$

To adjust the weight of the interaction between the audio and video modalities, we perform a weighted sum of the audio attention matrix (α) and the video attention matrix (β). The resulting weighted fusion attention (H_{WCA}) is computed as follows:

$$H_{WCA} = \alpha \times \text{ H}_a^c \oplus \beta \times \text{ H}_v^c \tag{7}$$

We use a residual layer on audio and video representations to preserve modality-specific information (H_a^r, H_v^r), and then pass them through a linear layer and a normalization layer. Finally, we combine them (H_a^r, H_v^r) with H_{WCA} to obtain the jointly contextual utterance representation (u_t^{CA}).

The contextual state captures the complex joint features from audio and video modalities in a single representation and propagates the overall attentive utterance-level information throughout the conversation. We model the attentive contextual state of the participants using the GRU_C cell. The state is



"Cross": cross-attention layer, "FF": feed-forward layer, "FC": fully connected layer

Fig. 1. (a) Update scheme for attentive contextual state for t utterance in a dialogue. (b) the interaction of audio-video modality utterance

updated at each time step t using the contextual state C_{t-1}, the speaker state $q_{i,t-1}$ from the previous utterance, and the bi-modal utterance representation (u_t^{CA}) at timestamp t.

$$C_t = GRU_C(C_{t-1}, (u^{CA}_t \oplus q_{i,t-1})) \tag{8}$$

2) Speaker State: Speaker usually frames their response based on the context, which includes the preceding utterances in the conversation. Therefore, we employ soft-attention on the history of interactive context to capture attentive long-context speaker interaction influences, learn conversational dependencies, and model the contribution of the attentive context-rich information. We pool the attention vector from the surrounding context history $[C_1, C_2, \ldots, C_{t-1}]$ using soft-attention (see Figure 3). This contextual attention vector ac_t can be computed as follows:

$$\begin{aligned} u^{CA}_i &= \tanh(WC_i + b), \quad 1 \leq i \leq t-1, \\ \alpha_i &= \frac{\exp(u_i^{CA})}{\sum_{i=1}^{t-1} \exp(u_i^{CA})}, \\ ac_t &= \sum\nolimits_{i=1}^{t-1} \alpha_i C_i. \end{aligned} \tag{9}$$

In Eq. (9), attention scores are calculated over the previous contextual states that represent the previous utterances. This assigns higher attention scores to emotionally relevant utterances (u^{CA}_t). Finally, the context vector C_i is calculated by globally pooling the states with α_i.

Next, we use a *GRU* cell to update the current speaker state $q_{i,t-1}$ to the new state $q_{i,t}$ based on the incoming utterance u^{CA}_t and the attentive context ac_t. The update equations for GRU_S are as follows:

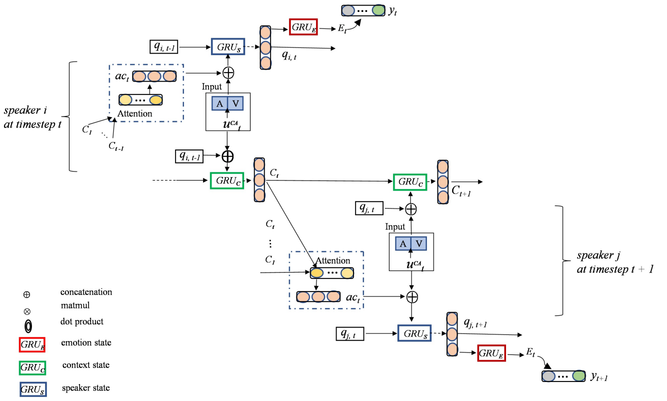


Fig. 2. Update schemes for speaker and listener state in a dialogue. Here Person i is the speaker and Persons $j \in \{1, M\}$ and $j \neq i$ are the listeners

$$q_{i,t} = GRU_S(q_{i,t-1}, (ac_t \oplus u^{CA}_t)) \tag{10}$$

3) Listener State: The listener state captures how the listener's state changes in response to the speaker's utterance, which can be observed by the participants through acoustic features, visual expressions, and other related aspects. The internal state of the participants is influenced by their emotions and the perceived effects from other participants, which may not always be explicitly expressed. In addition to emotions, this state can also encompass aspects that the participant actively tries not to express or features that are considered common knowledge and do not require explicit communication. Therefore, considering the effect on oneself is crucial for representing the internal state of the participants.

To update the intra-listener state $q_{j,t-1}$ to $q_{j,t}$, we adapt the visual modality v_t^{DAN} and the bi-modal utterance representation u^{CA}_t at timestamp t with the listener state $q_{j,t-1}$ (see Figure 4):

$$q_{j,t} = GRU_L(q_{j,t-1}, (v_t^{DAN} \oplus u^{CA}_t)) \tag{11}$$

4) Emotion State: The emotional state E_t represents the emotional mood of the participant and the inferred emotion class of the current utterance at time step t. The emotion state is updated based on the context state C_t, speaker state $q_{i,t}$, and listener state $q_{j,t}$ (see Figure 4). This update is done using a GRU_E model, which combines all these factors as follows:

$$E_t = GRU_E(E_{t-1}, (C_t \oplus q_{i,t} \oplus q_{j,t})) \tag{12}$$

5) Classification: We feed emotion state E_t into a fully connected network to get the final emotion inferences of all utterances:

$$P_t = softmax(W_{smax}E_t + b_{smax}) \tag{13}$$

$$y_t = argmax(P_t[k]) \tag{14}$$

Categorical cross-entropy loss is used as the loss function, and L2-regularization is applied by adding a penalty to the cost function.

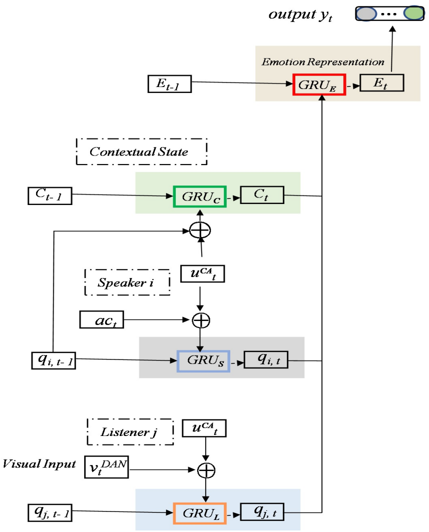


Fig. 3. Update schemes for emotion state in a dialogue. Here Person i is the speaker and Persons $j \in \{1, M\}$ and $j \neq i$ are the listeners

III. Experiments

A. Database and Metrics

We evaluated our proposed model on a multi-modal and multi-party dataset for conversational emotion recognition: Multi-modal EmotionLines Dataset (MELD). MELD is a multi-modal scenario that includes audio, visual, and text modalities for each utterance in a conversation [22].

Table 1: Dataset Distribution on Multi-modal EmotionLines Dataset

	Train	Validation	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281
Total	9989	1109	2610

MELD contains 13,708 utterances with pre-defined seven emotions (anger, disgust, sadness, joy, neutral, surprise, and fear) from 1,433 dialogues of the TV series ”Friends”. Each conversation involves two or more speakers. For a fair comparison, we conducted audio-video modality experiments using the predefined train/validation/test splits in MELD, which consist of 9,989, 1,109, and 2,610 utterances, respectively (see Table 1). Due to the unbalanced distribution of utterances across different emotion labels in the dataset, we evaluated the mean classification performance using precision, recall, and weighted F1-score for the seven emotion categories.

B. Baselines and State-of-the-Art

For a comprehensive evaluation, we compared our model with the following baselines and state-of-the-art models in multimodal emotion recognition (Table 2). We utilized the published results from references [30-33].

- M2F2 proposed a multi-head fusion attention module to extract emotion-rich latent representations of emotion-relevant features from the audio, text, and visual modalities. Specifically, they introduced a new adaptive margin triple loss function to help the extractor module effectively learn representations [30].
- MM-DFN designed a graph-based dynamic fusion approach to fuse multimodal context features for emotion recognition [31].
- CTC introduced MELD with Fixed Audiovisual Information via Realignment using recent active speaker detection and automatic speech recognition models [32].
- GA2MIF employed a multimodal fusion approach named Graph and Attention based Two-stage Multi-source Information Fusion for emotion detection in conversations [33].

C. Experimental Setup

In this study, the dataset is divided into a training set, validation set, and test set at a ratio of 8:1:1 [22]. We implemented our proposed model using the PyTorch 1.11.0 framework. The model was trained with the Adam optimizer, using an initial learning rate of 1e-4 and a batch size of 32. Cross-entropy loss was employed as the loss function. To prevent overfitting, the network was regularized using the L2 norm of the model’s parameters with a weight of 3e-4. Additionally, a dropout rate of 0.3 was applied during training.

IV. Results and Discussion

In this section, we compare the performance of *AVAFN* with state-of-the-art methods on the MELD dataset for audio, video, and bi-modal emotion recognition. Additionally, we provide a discussion and analysis to demonstrate the effectiveness of the proposed method.

A. Comparison with Baselines and State-of-the-Art Models

To demonstrate the effectiveness of our proposed method, we compare it with state-of-the-art approaches in emotion recognition tasks on the benchmark MELD dataset (refer to Table 2 and Figure 5). Based on these results, we draw the following observations. Firstly, among the single-modal methods, audio features demonstrate relatively better performance compared to visual features on the MELD dataset. Particularly, the utilization of the pre-trained *wav2vec* audio embedding proves to be effective in representing emotion-rich features. Secondly, in the case of bi-modal methods, the audio-video emotion recognition models outperform most of the single-modal emotion recognition analyses on the MELD dataset. Overall, our proposed *AVAFN* framework exhibits superior performance compared to the existing best model, achieving a 3.31% higher F1-score on the MELD dataset.

Our proposed approach demonstrates a strong ability to accurately infer emotions such as neutral, anger, sadness, surprise, joy, and disgust, which are explicitly expressed through audio and video modalities (refer to Figure 5). However,

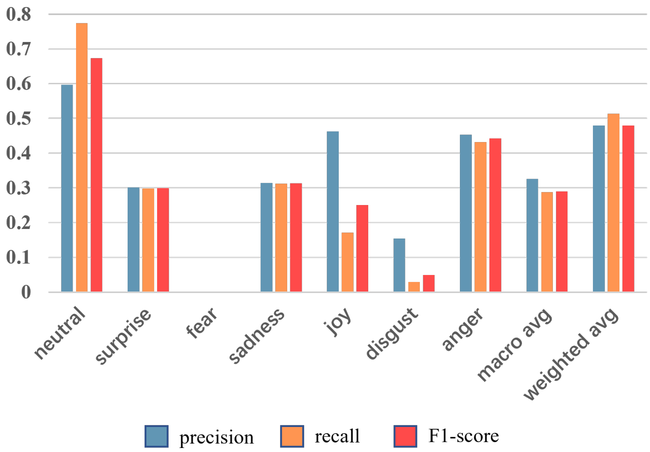


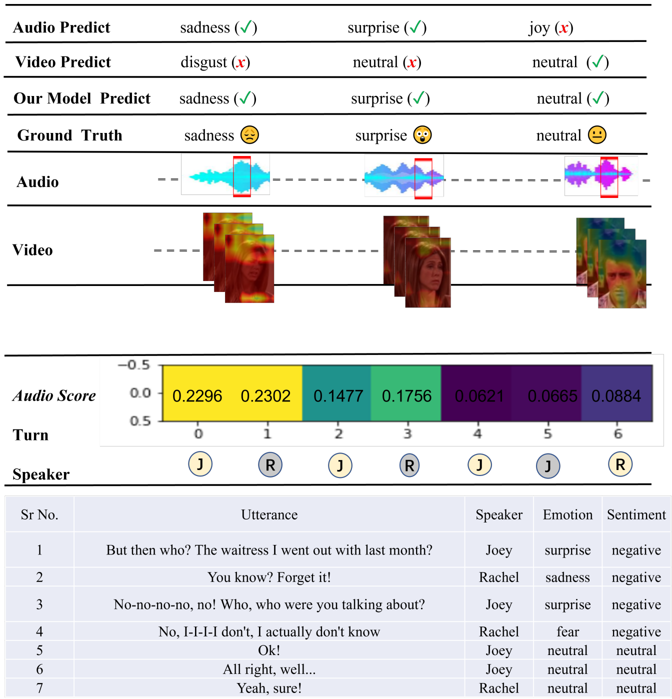
Fig. 4. Precision, recall, F1 score performance of the proposed on the MELD dataset.

when it comes to the emotion of fear, most existing models struggle to recognize it, potentially due to the implicit ways in which speakers express fear and the limited number of samples available for this emotion in the dataset. While the M2F2 model can capture long-term contextual dependencies, it fails to consider the correlations between different modalities. In contrast, our AVAFN framework effectively leverages the structural similarities within each modality, maintains consistency across modalities, and learns speaker-sensitive dependencies, resulting in improved emotion recognition.

While the performance of the visual modality is relatively lower, it does not imply that it lacks effective emotional information. Specifically, the emotion of sadness tends to be misclassified as fear or disgust. In a dialogue system, a short utterance consisting of a single word, like ”okay,” can express three different emotions: joy, neutrality, and sadness. The visual cues, such as a frown or a crying face, are crucial for recognizing sadness. In our work, we utilize a pre-trained *DAN* model to extract high-level visual representations that are relevant to emotions. This approach demonstrates the effectiveness of incorporating prior localized temporal knowledge from traditional feature-based transfer learning methods, which greatly enhances highlight detection.

Table 2: Average weight F1 score performance of the proposed and state-of-the-art methods for audio-video fusion on the MELD.

model	auio	video	audio-video	references
M2F2	39.63	32.44	35.74	[30]
MM-DFN	42.72	32.34	44.67	[31]
CTC	40.54	35.28	39.81	[32]
GA2MIF	_	_	43.54	[33]
AVAFN	**45.2**	**36.2**	**47.98**	_



Audio Predict	sadness (✓)	surprise (✓)	joy (x)
Video Predict	disgust (x)	neutral (x)	neutral (✓)
Our Model Predict	sadness (✓)	surprise (✓)	neutral (✓)
Ground Truth	sadness 😞	surprise 😲	neutral 😐

Sr No.	Utterance	Speaker	Emotion	Sentiment
1	But then who? The waitress I went out with last month?	Joey	surprise	negative
2	You know? Forget it!	Rachel	sadness	negative
3	No-no-no-no, no! Who, who were you talking about?	Joey	surprise	negative
4	No, I-I-I-I don't, I actually don't know	Rachel	fear	negative
5	Ok!	Joey	neutral	neutral
6	All right, well...	Joey	neutral	neutral
7	Yeah, sure!	Rachel	neutral	neutral

Fig. 5. Attention weight visualization of our model from the cases in MELD dataset.

The audio modality shows promise in effectively addressing the aforementioned issues by providing rich emotional features in the joint representation. The time-dependent acoustic content at the utterance level encompasses various information, including pitch, energy, tone, and loudness. This knowledge plays a crucial role in capturing the essence and progression of emotion-relevant information from speakers in emotion-aware spoken dialogue systems. Integrating audio information with visual information offers significant advantages, such as introducing additional audio features, disambiguating visual information, and bridging the gap to real-world environments for improved recognition. For instance, as depicted in Figure 6, a sequence of video frames with unchanged facial expressions can convey ambiguous emotions in different situations, such as joy, neutrality, or sadness. However, when corresponding audio information, such as the speaker’s high voice and sobs, is introduced, it becomes easier to discern the sentiment of the utterance as positive. In general, deploying joint representations from different modalities tends to yield better judgment of the emotional state by leveraging complementary information.

The *AVAFN* framework surpasses the state-of-the-art model in emotion recognition by achieving a 3.31% higher F1-score, which can be attributed to two main factors. Firstly, *AVAFN* integrates the audio and video features by jointly fusing them into a shared space using cross-attention calibration and weighted fusion with temporal awareness. This fusion module effectively combines the audio and visual modalities by assigning significant attention weight to emotion-relevant acoustic features, enhancing the details present in each individual modality, and adaptively integrating the implicit complementary content to amplify the interactions and correlations. Secondly, *AVAFN* incorporates attentive long-distance contextual information from the surrounding utterance history, allowing for better control of information flow during emotion transitions and capturing speaker-sensitive emotional dynamics in multiturn conversations. These factors enable *AVAFN* to maintain speaker-sensitive dependencies and facilitate synchronization.

B. Ablation Studies

AVAFN has four branches of bi-directional *GRU* cells to capture intra- and cross-modal interactions of audio and video, and learn attentive contextual information, and model speaker state and listener information. Emotion recognition in conversations requires understanding the temporal dynamics and dependencies among utterances. *GRU*, as a variant of recurrent neural networks, is known for its ability to model sequential data and capture long-term dependencies. It is designed to allow information to flow through the network over longer time steps. By utilizing *GRU*, our model can effectively capture the temporal evolution of emotions in conversations. Additionally, *GRU* provides flexibility in modeling various types of dependencies within the conversation. It can learn to focus on relevant contextual information while disregarding irrelevant or noisy inputs. This adaptability allows *GRU* to capture the nuanced relationships between utterances and the emotional context within the conversation, enabling it to capture the long-term dependencies necessary for accurate emotion recognition.

The results of the ablation studies presented in Table 3 demonstrate the importance of various components of our proposed *AVAFN* framework for emotion recognition in dialogue systems. The results indicate that the long-distance emotion-relevant contextual information is crucial for dialog-aware emotion recognition, and the combination of the attentive contextual module (*AVAFN w/o* ACM) improves the performance by at least 2.64% when compared to models that do not use it. Additionally, the self-speaker module (*AVAFN w/o* SSM) and intra-listener module (*AVAFN w/o* ILM) contribute to the performance of the model by capturing speaker influence and listener state, respectively.

Table 3: Albation study on the MELD dataset

Method	w-average F1
(*AVAFN w/o* ACM)	45.35
(*AVAFN w/o* SSM)	46.58
(*AVAFN w/o* ILM)	46.92
(***AVAFN***)	**47.98**

The attentive contextual module (*AVAFN w/o* ACM) contributes the most to the performance of the model by incorporating long-term emotion-relevant contextual information from surrounding history utterances. The cross-attention coupled with weighted fusion in the fusion unit introduces the underlying interaction between audio and video features, which enhances the dependencies and stability across various modalities. By introducing more attentive weight to relevant video data, we empowered the information on the video representation to help the audio content effectively adjust the weight of words by temporal-awareness. Our fusion unit effectively deploys different information into a new space to magnify the details embedded in a single modality from the contextual level and adapatively integrates implicit complementary information to strengthen the interactions and correlations between different information subspaces, reducing information gaps.

C. Case Studies

The conversational snippet presented in Figure 6 highlights the effectiveness of our proposed framework in detecting emotion shifts during a conversation. The snippet starts with Rachel being in a sadness state while Joey is the speaker. However, Joey changes his focus and questions Rachel on her state, which leads to her feeling fear. Our framework is able to accurately infer the emotion shift from sadness to fear, demonstrating its ability to capture the dynamics of emotional states during a conversation.

The challenge in this scenario is that it can be difficult to recognize the emotion of Rachel from her calm tone of fear. By giving more weight to frightened expression, our framework is able to effectively complement the audio modality with the visual modality to capture more rich emotional features. The cross-attention coupled with weighted fusion is crucial in selecting the most critical information from all modalities, which enhances the model's ability to detect emotion shifts and accurately classify the emotional state of the speaker. Overall, our framework aims to strike a balance between different modalities and improve the comprehensive capabilities of the model for emotion-aware spoken dialog systems (See Figure 6).

V. Conclusion

In this paper, we proposed the *AVAFN* framework based on high-level pre-trained models, namely *wav2vec* and *DAN*, in a dynamic manner. The cross-attention module, as the core unit of the *AVAFN* framework, played a crucial role in explicitly modeling inter- and intra-modal interactions within and between audio and video modalities. Additionally, bidirectional *GRU* components were adopted to capture long-distance contextual dependencies and model the state of the speaker and listener. Our approach achieved state-of-the-art performance in utterance-level recognition when evaluated on the standard benchmark MELD dataset.

While the achieved performance in multimodal emotion recognition is encouraging, there are still some limitations that need to be addressed to improve emotion recognition further. Most benchmark datasets used for the multimodal emotion recognition task rely on controlled databases with non-realistic recording conditions and error-less text transcriptions. Therefore, in practical applications, to develop an applicable and generalizable model, we need to construct it for the problem of multimodal emotion recognition in conversations that can handle the following scenarios: 1) exploring unexpected automatic speech recognition errors; 2) performing inference in cases of noisy or absent modalities, or unaligned temporal multimodal data.

References

[1] Y. Wang, S. Wei, W. Tao, L. Antonio, D.W. Yang, X.L. Li, S.Y. Gao, Y.X. Sun, W.F. Ge, W. Zhang, W.Q. Zhang, "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, vol. 83-84, pp. 19-52, 2022.

[2] L.C. Sun, B. Liu, J.H. Tao, L. Zheng, “Multimodal cross- and self-attention network for speech emotion recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4275-4279, 2022.

[3] P. Kumar, S. Jain, B. Raman, P.P. Roy, M. lwamura,“End-to-end triplet loss based emotion embedding system for speech emotion recognition,” *25th International Conference on Pattern Recognition (ICPR)*, pp. 1-8, 2021.

[4] JY.Y. Jiang, W. Li, M.S. Hossain, M. Chen, A. Alelaiwi, M.A. Hammadi, “A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition,” *Information Fusion*, vol. 53, pp. 209-221, 2020.

[5] K. Roemmich, N. Andalibi, “Data subjects’ conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1-34, 2021.

[6] F.Y. Liu, G.D. Zhang, X. Sheng, J. Lei, Q.F. Xiang, B. Jiang, P.K. Chen, “Predicting students’ performance in e-learning using learning process and behaviour data,” *Scientific Reports*, vol. 12, no. 453, pp. 1-15, 2022.

[7] K.N. Wang, H.R. Xie, D. Zou, K.L. Chou, “Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources,” *Information Fusion*, vol. 77, pp. 107-117, 2022.

[8] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, “Deep residual learning for image recognition,” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[9] T.L. Nguygen, S. Kavuri, M. Lee, “A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips,” *Neural Networks*, vol. 118, pp. 208-219, 2019.

[10] D. Wan, M.Y. Xu, D.Y. Huang, W.S. Lin, M.H. Dong, X.G. Yu, H.Z. Li, “Audio and face video emotion recognition in the wild using deep neural networks and small datasets,” *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 506-513, 2016.

[11] J.J. Deng, C.H.C. Leung, Y.X. Li, “Multimodal emotion recognition using transfer learning on audio and text data,” *Computational Science and Its Applications (ICCSA)*, pp. 552-563, 2021.

[12] L. Scgineveld, A. Othmani, H. Abdelkawy, “Leveraging recent advances in deep learning for audio-Visual emotion recognition,” *Pattern Recognition Letters*, vol. 146, pp. 1-7, 2021.

[13] X. Chang, W. Skardek, “Multi-modal residual perceptron network for audio–video emotion recognition,” *sensors*, vol. 21, no, 16, pp. 1-17, 2021.

[14] S.Q. Zhang, S.L. Zhang, T.J. Huang, W. Gao, Q. Tian, “Learning affective features with a hybrid deep model for audio–visual emotion recognition,” *IEEE transactions on circuits and systems for video technology*, vol. 28, no. 10, pp. 3030-3043, 2018.

[15] Z. Wang, L.Z. Wang, H. Huang, “Joint low rank embedded multiple features learning for audio–visual emotion recognition,” *Neurocomputing*, vol. 388, pp. 324-333, 2020.

[16] Z.Y. Wen, W.Z. Lin, T. Wang, G. Xu, “Distract your attention: multi-head cross attention network for facial expression recognition,” *arXiv:2109.07270*, 2021.

[17] K. Sailunaz, M. Dhaliwal, J. Rokne, R. Alhajj, “Emotion detection from text and speech: a survey,” *Social Network Analysis and Mining*, vol. 8, pp. 1-26, 2018.

[18] Y. Wang, W. Song, W. Tao, A. Liotta, D.W. Yang, X.L. Li, S.Y. Gao, Y.X. Sun, W.F. Ge, W. Zhang, W.Q. Zhang, “A systematic review on affective computing: emotion models, databases, and recent advances,” *Information Fusion*, pp. 19-52, 2022.

[19] R.A. Khalil, E. Jones, M.I. Babar, T. Jan, M.H. Zafar, T. Alhussain, “Speech emotion recognition using deep learning techniques: a review,” *IEEE Access*, vol. 7, pp. 117327-117345, 2019.

[20] S. Pini, O.B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, B. Huet, “Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild,” *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 536-543, 2017.

[21] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, “Context-dependent sentiment analysis in user generated videos,” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 873-883, 2017.

[22] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, “MELD: a multimodal multi-party dataset for emotion recognition in conversations,” *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 527-536, 2019.

[23] M.B. Akcay, K. Oguz, “Speech emotion recognition: Emotional ˘models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, pp. 56-75, 2020.

[24] Y. Fan, X.J. Lu, D. Li, Y.L. Liu, “Video-based emotion recognition using CNN-RNN and C3D hybrid networks,” *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 445-450, 2016.

[25] M. Nahar, M. E. Ali, ”A deep ensemble approach of anager detection from audio-textual conversations,”*International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1-8, 2022.

[26] K. Vishi, V. Mavroeidis. “An evaluation of score level fusion approaches for fingerprint and finger-vein biometrics,” *arXiv:1805.10666*, 2018.

[27] S.Q. Zhang, S.L. Zhang, T.J. Huang, W. Gao, Q. Tian, “Learning affective features with a hybrid deep model for audio–visual emotion recognition,” *IEEE Transaction Circuits System Video Technology*, vol. 28, no. 10, pp. 3030-3043, 2018.

[28] S. Pini, O.B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, B. Huet, “Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild,” *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 536-543, 2017.

[29] S. Schneider, A. Baevski. R. Collobert, M. Auli, “wav2vec: unsupervised pre-training for speech recognition,” *arXiv:1904.05862*, 2019.

[30] Vishal Chudasama, et al. “M2FNet: multi-modal fusion network for emotion recognition in conversation,” *arXiv:2206.02187*, 2022.

[31] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, N. Onoe, ”MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations,” 2022, :2203.02385.

[32] H. Carneiro, C. Weber, S. Wermter, ”Whose emotion matters? speaker detection without prior knowledge,” 2022, *arXiv:2211.15377*.

[33] J. Li, X.P. Wang, G.Q. Lv, Z.G. Zeng, ”GA2MIF: graph and attention based two-stage multi-source information fusion for conversational emotion detection,” *IEEE Transactions on affective computing*, pp. 1-14, 2023.