MDPI

Article

Robust MultiModal Emotion Recognition from Conversation with Transformer-based Cross-Modality Fusion

Baijun Xie, Mariia Sidulova and Chung Hyuk Park*

Department of Biomedical Engineering, School of Engineering and Applied Science, George Washington University, Washington, DC 20052, U.S.A; {bdxie, sidul001, chpark}@gwu.edu
* Correspondence: chpark@gwu.edu;

Abstract: Decades of scientific research have been conducted on developing and evaluating methods for automated emotion recognition. With exponentially growing technology, there is a wide range of emerging applications that require emotional state recognition of the user. This paper investigates a robust approach for multimodal emotion recognition during a conversation. Three separate models for audio, video and text modalities are structured and fine-tuned on MELD dataset. In this paper, a transformer-based cross-modality fusion with EmbraceNet architecture is employed to estimate the emotion. The proposed multimodal network architecture can achieve up to 65% accuracy, which significantly surpasses any of the unimodal models. We provide multiple evaluation techniques applied to our work to show that our model is robust and can even outperform the state-of-the-art models on the MELD dataset.

Keywords: multimdoal Emotion recognition; multimodal fusion; crossmodal transformer; attention mechanism

Citation: Xie, B.; Sidulova, M.; Park, C.H. Robust MultiModal Emotion Recognition from Conversation with Transformer-based Cross-Modality Fusion. *Sensors* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

In recent years there has been a growing number of studies that attempted to recognize human emotion from either speech [1–3], text [4,5], or facial expressions [6,7]. In reality, emotional communication is a temporal and multimodal process; typical human conversations consist of a variety of cues and expressions that are rarely static. Thus, multiple studies have highlighted the importance of multi-sensory integration when processing human emotions [8–11].

Emotion recognition has extensive application prospects, including but not limited to Human-Robot Interaction (HRI), Socially Assistive Robotics (SAR), Human-Computer Interaction (HCI), and medicine. Discussion regarding the effectiveness of multimodal HRI has dominated the research in recent years. For example, in the paper by Stiefelhagen et al. [12], researchers discussed a novel multimodal HRI system, which included speech recognition, multimodal dialogue processing, visual detection, tracking, and identification of users, which combined both head-pose estimation and pointing gesture recognition. The study with human participants concluded that the incorporation of all of these modalities increased participant's engagement and made the HRI scenario more natural. For socially assistive robots (SAR) to effectively communicate with human beings, robotic systems should have the ability to interpret human affective cues and to react appropriately by exhibiting their own emotional response. In the work by Hong et al. [13], the researchers presented a multimodal emotional HRI architecture to assist in natural, engaging, bidirectional emotional communication between humans and a robot. Both body language and vocal intonation were measured to recognize the user's affection state. The results of the experiment with human participants have proved that bi-direction emotion recognition instigated more positive valence and less negative arousal during the interaction. Kim et al. [14] developed an audio-based emotion recognition system that can estimate the expression levels for valence, arousal, and dominance.

The extracted features from the speech data were used for training the automatic emotion classifier. This classifier offers emotional communications in a natural manner during human-robot interaction experiences for children with autism spectrum disorder (ASD). More recently, the study of [15] also applied deep learning methods for recognizing the emotion from the music. Several deep learning networks were employed to train the extracted spectral features from the music and predict the levels of arousal and valence.

Emotion recognition has also been heavily studied in the context of Human-Computer Interaction (HCI). Creating human-computer interaction that would be as natural and efficient as human-human interaction requires not only recognizing the emotion of the user but also expressing emotions. An example of such research could be the works of Maat and Pantic [16], where authors proposed a system that is capable of learning and analyzing the user's context-dependent behavior and adapting the interaction to support the user. Another example of an HCI system with emotion recognition is the Intelligent Tutoring System developed by Kapoor et al. [17], which incorporated multisensory data to assist in detecting frustration to predict when the user needs help. Emotion-oriented HCIs aim to not only automatically recognize emotional states but also mimic or synthesize emotions in speech, or facial expressions. As an example, previous studies focused on generating voices, which would have relevant emotional sentiment [18,19]. Learning emotions from the speech provided a way of generating convincingly emotional speech [20].

Medical specialists can also benefit from using emotion recognition systems as a diagnosing tool for a wide range of medical symptoms. For example, in the study by France et al. [21], researchers compared the acoustic trends in the speech of healthy, depressed, and suicidal individuals. Another medical example could be clinical studies of schizophrenia and the flat affect, which is indicative symptom of the diseases and is characterized as diminished emotional expressions [22,23].

In general, user affect is detected using a unique combination of body language and vocal intonation, and multimodal classification is performed using computational models, e.g., a Bayesian Network [24,25]. When applied in the robotics domain, human emotion recognition in the HRI system can especially make the interaction more natural, understandable, and intuitive [26]. In the study by Barros et al. [27], a Cross-Channel Convolutional Neural Network (CNN) structure was proposed for investigating how emotions are expressed by a robotic system changed the perception of human users. The network was able to predict human emotions using features based on facial expressions and body motions. The emotion recognition system was tested in a real HRI scenario with the iCub robot. The robot was able to detect three different emotional states and gave feedback by changing its mouth and eyebrow LEDs. In another study by Javed et al. [28], the robotic system was trained to recognize emotion states from both typically developing children and children with ASD. With the goal of measuring the overall engagement of the child during the HRI session, emotion recognition was used as one of the measured features. Therefore, it is believed that advancing emotion recognition from human beings can significantly improve the role of empathy in HRI scenarios as well.

In this study, a robust transformer-based multimodal fusion network for emotion recognition is presented. The embedding vectors from each individual modality are extracted from domain-specific models and fused via our proposed cross-modality transformer. In addition to considering the joint representation across different modalities, we introduce a robust multimodal fusion network to combine all the representation vectors from each modality. The results reach the state-of-the-art performance on the evaluated dataset.

2. Related Studies

2.1. Multimodal Fusion for Emotion Recognition

The primary fusion strategies from previous studies for multimodal emotion recognition can be classified into feature-level (early) fusion, decision-level (late) fusion, and

model-level fusion [29]. Conventionally, feature-level fusion concatenates the features from different modalities to get a joint representation, and the concatenated features are fed into a single classifier for emotion recognition. Schuller et al. [30] presented the baseline models, which concatenated the audio and visual features into a single feature vector and used support vector regression to predict the continuous affective values. Recent study [31] also investigated using Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to train different modalities of features and to combine different feature vectors via concatenation. However, feature-level fusion may suffer from the problem of data sparseness [29], so the performance of combining different modalities via a simple concatenation is limited.

Unlike feature-level fusion, decision-level fusion employs and trains separate classifiers for each modality and combines the outputs from each classifier to get the final prediction. Liu et al. [32] applied different kernel-based classifiers for different modalities and boosted the fusion results at the decision-level. However, decision-level fusion does not consider the mutual relations between different modalities, which results in losing correlated information among different modalities.

Researchers have also taken advantage of deep learning models for model-level fusion, which also consider the interrelations between different modalities. Chen et al. [33] achieved model-level fusion by concatenating the high-level features from different modalities. More recently, Choi et al. [34] proposed a novel multimodal fusion architecture based on the neural network for classification tasks. The model architecture assured both the effectiveness of the cross-modal relationship and the robustness of fusing different modalities.

2.2. Transformer Method

Multimodal emotion recognition (MMER) with the fusion by transformer has drawn much attention recently. The transformer is a network architecture that purely depends on the attention mechanism without any recurrent structure [35]. The latest studies focus on using attention mechanisms to fuse different modalities of features for MMER [36–39]. Ho et al. [36] proposed a multimodal approach based on Multi-level Multi-Head Fusion Attention mechanism and RNN to combine audio and text modalities for emotion estimation. The paper also stated that the method of attention mechanism fusion for multiple modalities improves emotion recognition performance comparing to the unimodal approach. Another study by Huang et al. also compared decision-level and feature-level fusions by using the attention mechanism [37]. This study utilized the transformer to fuse audio and visual modalities at the multi-head attention level. The experiments showed that the proposed method could have better performance via feature-level fusion.

The previous study has investigated the use of the crossmodal transformer to reinforce a target modality by introducing the features from another modality, which also learn the attention across these two modalities features [40]. One recent study [39] proposed a multimodal learning framework based on the crossmodal transformer target for conversational emotion recognition, combining word-level features and segment-level acoustic features as the inputs. The results demonstrate the effectiveness of the proposed transformer fusion method. Another recent study [38] combined three different modalities, text, audio, and vision, with features extracted from the pre-trained Self-Supervised-Learning model. This study designed a transformer-based multimodal fusion mechanism that also considered the inter-modality connections and achieved state-of-the-art results for the task of MMER.

Even though the effectiveness of combining two different modalities by using the attention mechanism has been widely studied, the challenge emerges when there arise needs to combine three or more modalities due to the structure of multi-head attention. For this reason, most previous studies for MMER based on attention mechanism proposed and tested the network architecture for only two modalities [36,37,39]. The

study [38] deployed models for combining three modalities, but a simple feature concatenation was added at the end to combine different modalities of features. However, in practice, concatenating different high-dimensional features may result in data sparseness and degrade the performance [29].

3. Dataset

Multimodal Emotion Lines Dataset (MELD) [41] is an extended version of the EmotionLines dataset [42]. MELD dataset includes a total of 13708 samples segmented out from the TV Series *Friend's*, with the samples grouped as 9989 for training, 1109 for validation, and 2610 for testing. Each segmented sample has the following data attributes that have been used in this study: video clip, utterance, text from the utterance, and emotion labels. There are seven emotion labels available for the dataset: *Anger*, *Disgust*, *Sadness*, *Joy*, *Neutral*, *Surprise*, and *Fear*. These are typically considered to be the basic emotions, and other emotions are seen either as combinations of these basic emotions as studied by Bower [43]. Label distributions in train, validation and test datasets can be seen in Figure 1. There is an inherent imbalance in the MELD dataset,*Neutral* samples have the dominating number of samples in each of the datasets. Utterance audio clips are varied in time length but are averaged to around 5 seconds long recording. On average, there are around 7 words per one utterance. The average duration of an utterance is 3.95s, and the average number of utterances per dialogue is 9.6. The number of emotions per dialogue is 3.3 on average.

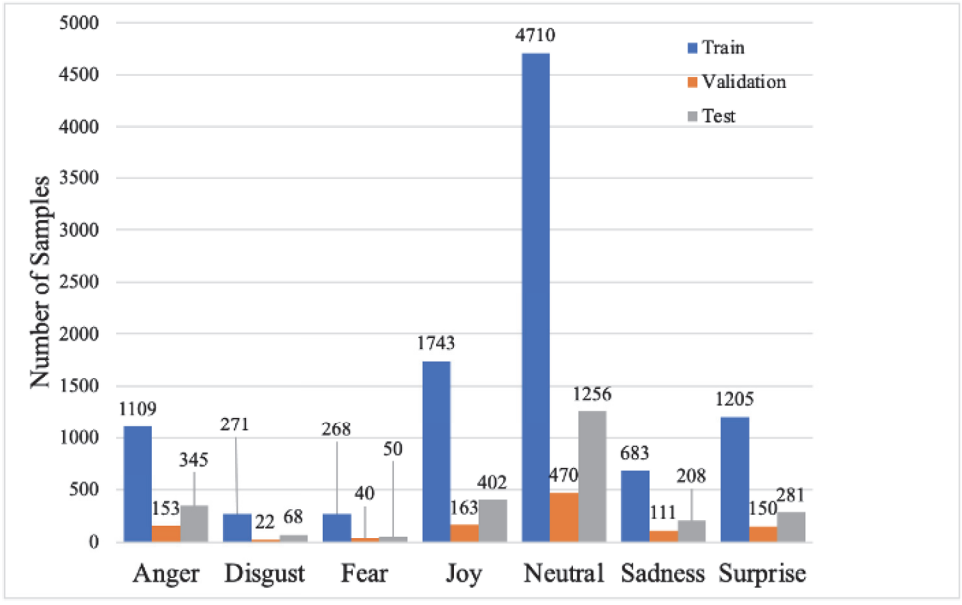

Figure 1. Number of samples per one emotion label for training, validation and testing dataset.

4. Methods

In this study, three different deep learning models are introduced for feature extraction from three different modalities. The cross-modality fusion transformer approach and the EmbraceNet for fusion are discussed for multimodal classification.

4.1. Feature Extraction

Three different deep neural networks are used for feature extraction: Generative Pre-trained Transformer (GPT), WaveRNN, and FaceNet+RNN. Transfer Learning is a training scheme that provides help for training small dataset [44]. Transfer Learning proposed first pre-training the model in a large scale dataset, and then for another smaller target dataset, pre-trained model can be fine-tuned and achieve desired performance

efficiently. Therefore, Transfer Learning paradigm is adopted in this study. In the stage of feature extraction, we first fine-tune the domain-specific models for the emotion classification task with a single modality. Then, we extract the features before the fully connected layer from the fine-tuned models, which prepare for the fusion and training in the multimodal fusion stage. Figure 2 shows a dialogue sample from the conversation of two actors, and different modalities of features from the dialogue are fed into the corresponding models. The WaveRNN will accept the audio features from the sentence's audio clip, and the sequential face images from the video clip will be fed into the FaceNet+RNN model. For the text modality, the input consists of the dialogue of the history, current sentence, and reply.

4.1.1. Model for Text Modality

We use the GPT [45] as our language model for extracting the features from the text modality. GPT is a multi-layer transformer-based model, which was pre-trained on BookCorpus dataset [46] and fine-tuned on the MELD dialogue dataset. As shown in Figure 2, the transfer learning scheme proposed by Wolf et al. [47] was adapted. Interested readers can refer their repository[1] for more details. To estimate the emotional states of the current sentence, we combine the history of the dialogues, the current sentence, and the reply sentence to generate sentence embeddings. Then the position embeddings and segment embeddings are summed up to build a single input sequence for the transformer model to get the output sequence tokens.

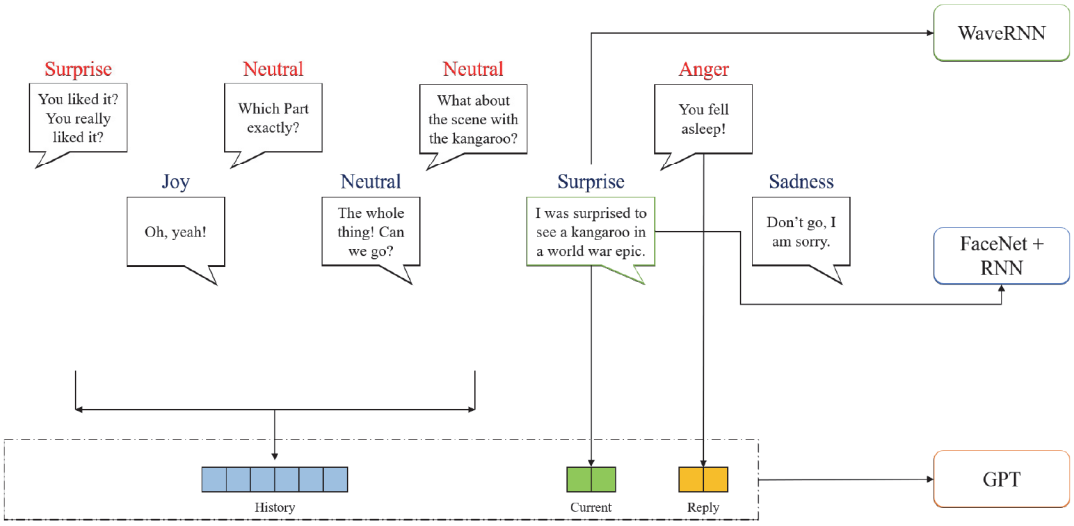

Figure 2. The cross-modality attention fusion Transformer

The optimization of the model is performed by multi-task learning. The fine-tuning is performed by introducing a combination of different loss functions: (1) next-sentence prediction loss, (2) language modeling loss, (3) emotion classification loss. For the next-sentence prediction loss, $L(s)$, a special token [CLS] is added at the end of the input sequence. The optimization goal is to train a classifier to discriminate the gold reply (ground-truth sentence) from the randomly sampled distractor (the sentence other than ground-truth). The corresponding last hidden state for token [CLS] is passed to the linear classifier, and the cross-entropy loss is used to optimize the classifier. For the language modeling loss, $L(l)$, the final hidden states of the outputs are extracted to predict the next reply tokens, and the cross-entropy loss is used to optimize the language model. For the emotion classification loss, $L(e)$, a multi-layers classifier is introduced

[1] https://github.com/huggingface/transfer-learning-conv-ai

for classifying seven emotion labels from the MELD dataset. During training, the last hidden state from the final hidden layer is fed into the classifier to predict the emotions, and the cross-entropy loss is computed to optimize the emotion model. Finally, the total loss function for optimization is the sum of these losses:

$$L(total) = L(s) + L(l) + L(e). \quad (1)$$

4.1.2. Model for Audio Modality

WaveRNN [48] model is used as an audio model to extract features from the original audio clips. The original waveform of the audio and generated spectrogram of the signal are both inputted to the model. The pre-trained model used in this study was trained on LJSpeech dataset [49]. During the pre-processing, waveforms are sampled with the 16000 Hz sample rate, and spectrograms are generated. Moreover, zero-padding for the batch inputs is applied on-the-fly during training. The input channels of waveform and spectrogram have to be 1, so we average the values from both channels of the audio inputs. All the pre-processing is implemented via the torchaudio library[2]. As shown in Figure 2, the acoustic features from the time interval of one person's speeches are extracted as the inputs for the model.

The input shapes for the WaveRNN can be defined as follows:

$$[batch\_size, n\_channel = 1, feature\_size = (n\_time - kernel\_size + 1) \times hop\_length]$$

for the input waveform, and

$$[batch\_size, n\_channel = 1, n\_freq, n\_time]$$

for the input spectrogram, where $hop\_length = 200$, $kernel\_size = 5$, and $n\_time = 240$ in this study. As can be seen in Figure 3, the feature vector with the shape of $[(n\_time - kernel\_size + 1) \times hop\_length = 47200, 512]$ before the fully connected layer of WaveRNN model is extracted, where 512 is the default feature size of WaveRNN output. A classifier consists of two consecutive linear layers at the end will give us the predictions of emotional states.

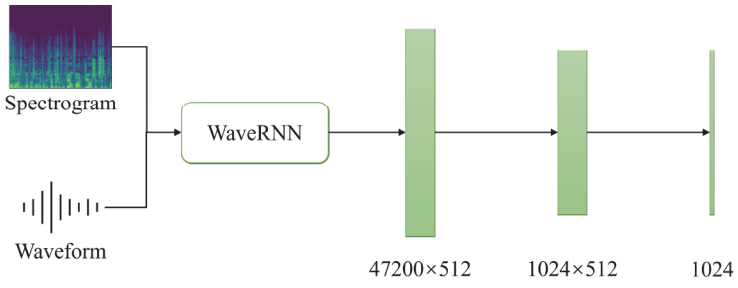

Figure 3. The classifier added to WaveRNN.

4.1.3. Model for Face Modality

Deep CNN has been widely studied for facial emotion recognition from videos by extracting the sequence of face embeddings [50–52]. Studies show that combining deep CNN and temporal models can effectively recognize facial emotion from the videos [50,52]. FaceNet is a deep CNN model that has been utilized to identify facial features from the image inputs [53]. FaceNet has an open-source library[3] implemented by Pytorch designed for face verification, recognition, and feature extraction. Multi-task CNN (MTCNN) [54] recognizes a face in the video frames and returns the sequence

[2] https://pytorch.org/audio/stable/index.html
[3] https://github.com/timesler/facenet-pytorch

of the cropped images with the detected face from the video input. Afterward, we use the Inception ResNet (V1) model, which was pre-trained on VGGFace2 [55] and CASIA-Webface [56] datasets, to extract the sequence of face embeddings from the sequence of images. The dimension of the returned embeddings, by default, equals 512. As illustrated in Figure 2, we can get a sequence of face embeddings during the time interval of one person's speech.

Furthermore, we incorporate an RNN-based model to learn the temporal relations in the sequence of images. Gated recurrent unit (GRU) [57] is similar to long short-term memory (LSTM) [58] network with a forget gate but more efficient for training. The formulations of the GRU network can be stated as follows:

- Update gate z_t: defines how much of the previous memory to keep.

$$z_t = \sigma(W^z x_t + U^z h_{t-1}), \tag{2}$$

- Reset gate r_t: determines how to combine the new input with the previous memory.

$$r_t = \sigma(W^r x_t + U^r h_{t-1}), \tag{3}$$

- Cell value $\tilde{h}_t$:

$$\tilde{h}_t = \tanh(W^h x_t + U^h(h_{t-1} \odot r_t)), \tag{4}$$

- Hidden value h_t:

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}, \tag{5}$$

where $\odot$ denotes the Hadamard product.

The extracted sequence of face images has variable lengths; therefore, some samples are down-sampled to the fixed length. Due to the limitation of the GPU's memory and the consideration of training efficiency, the fixed length for the extracted images sequence in this study is set to 50 images. Oppositely, zero-padding is applied during data pre-processing to the sequences with a shorter length. Finally, during training, the last hidden state value, h_t, of the GRU outputs are extracted, and a classifier is introduced to learn the emotion classification task.

4.2. Robust Cross-Modality Fusion Transformer with EmbraceNet

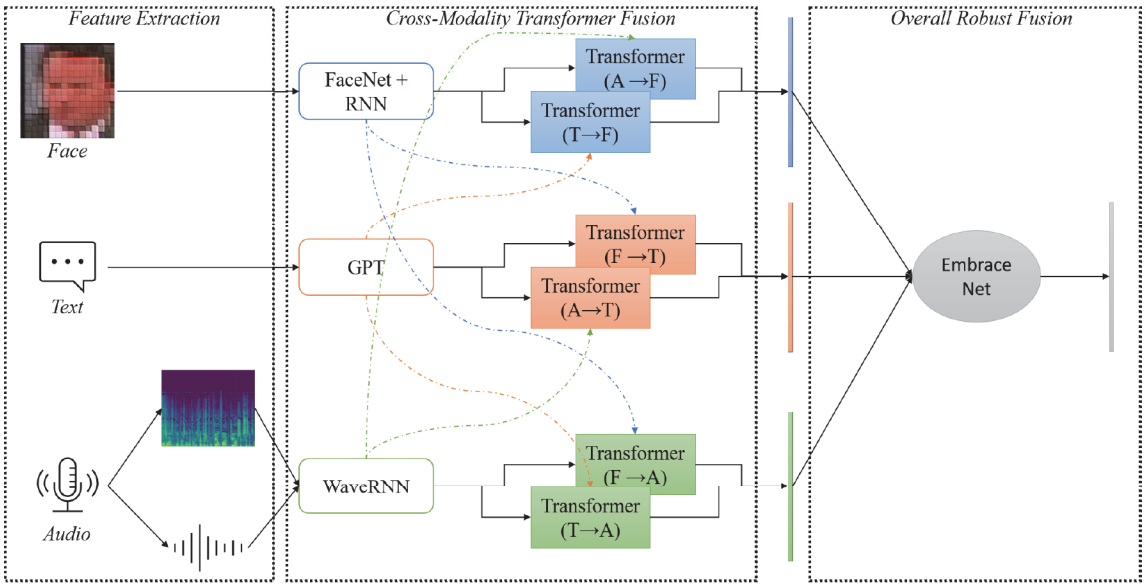

Figure 4. The crossmodal fusion transformer [40] with EmbraceNet [34]

In this section, we demonstrate the network architecture for multimodal fusion and classification, which is depicted in Figure 4. Our model consists of two main parts, the *cross-modality transformer fusion* and *overall robust fusion*.

4.2.1. Cross-Modality Transformer Fusion

The idea of the crossmodal transformer was initially proposed by Tsai et al. [40]. The crossmodal transformer can enrich the information for one modality from another modality. In this study, we adapt the network architecture to fuse different input modalities for emotion recognition. Following the formulation of [40], for example, we denote the passing of modality A information to another modality B by using "$A \rightarrow B$". The multi-head attention was proposed in works by [35], where the attention function is mapping Query, Keys, and Values to the output.

Figure 5 shows the architecture of the crossmodal transfromer. We denote the the input features for modalities A and B are $X_A \in \mathbb{R}^{T_A \times d_A}, X_B \in \mathbb{R}^{T_A \times d_B}$, where T and d are sequence length and feature size. As shown in Figure 5, We can define the Query as $Q_A = X_A W_{Q_A}$, Keys as $K_B = X_B W_{K_B}$, Values as $V_B = X_B W_{V_B}$, where $W_{Q_A} \in \mathbb{R}^{d_A \times d_k}, W_{K_B} \in \mathbb{R}^{d_B \times d_k}, W_{V_B} \in \mathbb{R}^{d_B \times d_V}$ are the weights. Then, the fused attention output vector Y from modality A to B can be represented as follows:

$$\begin{aligned} Y_{attention} &= Attention(X_A, X_B) \\ &= softmax(\frac{X_A K_B^\intercal}{\sqrt{d_k}}) V_B \\ &= softmax(\frac{X_A W_{Q_A} W_{K_B}^\intercal X_B^\intercal}{\sqrt{d_k}}) X_B W_{V_B}. \end{aligned} \tag{6}$$

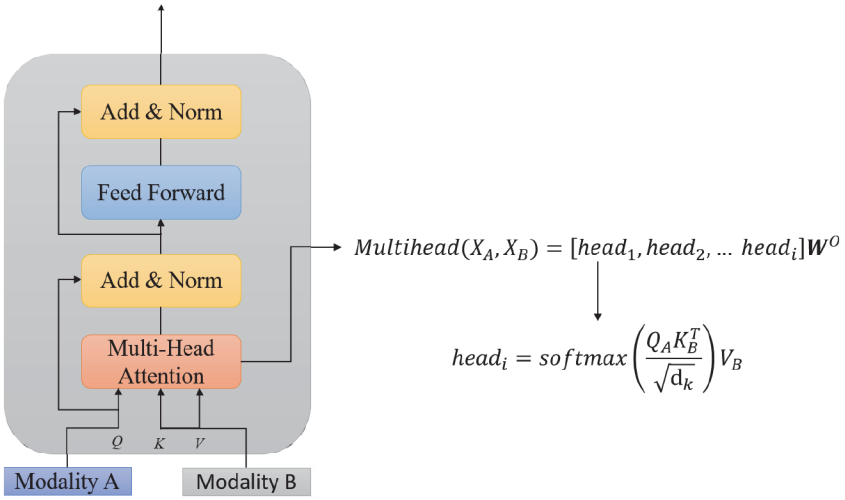

Figure 5. The architecture of the crossmodal transformer.

Following the setting from the previous study [35], we also add a residual connection from Query to the attention output and come with the layer normalization.

$$x = LayerNorm(Y_{attention} + Q_A). \tag{7}$$

Then, a feed-forward layer is applied, which consists of two fully connected layers with a ReLU activation function:

$$f_x = Linear(x) = xA_x^\intercal + b_x. \tag{8}$$

$$x_a = ReLU(f_x). \tag{9}$$

$$f_{x_a} = Linear(x_a) = x_a A'^{\mathsf{T}}_{x_a} + b_{xa}, \tag{10}$$

where $A_x \in \mathbb{R}^{d_x \times 2d_x}$, $A'_{x_a} \in \mathbb{R}^{2d_x \times d_x}$, and $b_x \in \mathbb{R}^{2d_x}$, $b_x \in \mathbb{R}^{d_x}$.

Finally, another residual connection with normalization is used to get the final embedding representation vector from Modality A to B.

$$out_{A \to B} = LayerNorm(f_{x_a} + x). \tag{11}$$

Taking the example of the crossmodal fusion for text modality, the final attention representation of the feature is the Hadamard product of two cross-modality features as suggested in previous study [38], which is given as follows:

$$Attention_T - out_{F \to T} \odot out_{A \to T}. \tag{12}$$

where A denotes audio, F denotes face and T denotes text, and $\odot$ denotes the Hadamard product.

During training, the crossmodal transformer module transforms the source modality into the Key/Value pair to interact with Query, the target modality. The previous study shows that the crossmodal transformer can learn correlated information across modalities [40].

4.2.2. EmbraceNet for Robust Multimodal Fusion

For the multimodal emotion recognition task, we are not only considering cross-modal fusion by using the transformer but also want to assure the robustness of combining the multimodal outputs. We employ the EmbraceNet [34] into our network architecture, which focuses on dealing with cross-modal information carefully and avoids performance degradation by the cause of the partial absence of data.

As can be seen in Figure 6, the EmbraceNet consists of two main components, docking layers and an embracement layer.

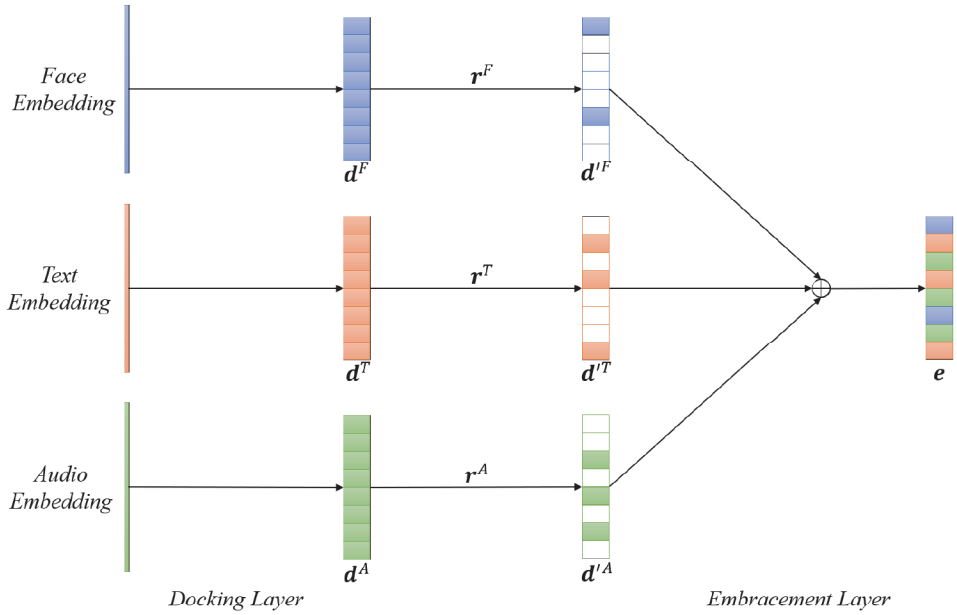

Figure 6. The EmbraceNet for multimodal fusion

4.2.3. Docking Layers

Three different deep learning models are used to get the feature vectors from three different modalities. Each modality has different characteristics and different sizes of extracted feature vectors, so the docking layers act as a preprocessing module before being fed into the embracement layer by converting different modalities features into the same size. The docking layers consist of a fully connected layer followed by a ReLU activation function. Each feature vector from different modalities is converted to the same embracement size, which in this study is equal to 256. Finally, the docking layers output m feature vectors, $\mathbf{d}^k \in \{\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, ..., \mathbf{d}^{(m)}\}$, where $m = 3$ denotes the number of modalities in this study, and $\mathbf{d}^k = [d_1^{(k)}, d_2^{(k)}, ..., d_c^{(k)}]^\intercal$,where c is the dimensionality of the vector.

4.2.4. Embracement Layer

The embracement layer is formalized as follows. Let $\mathbf{r}_i = [r_i^{(1)}, r_i^{(2)}, ...r_i^{(m)}]$, where $i \in \{1, 2, ..., c\}$ and $m \in \{1, 2, 3\}$, be a vector that can be drawn from a multinomial distribution, i.e.,

$$\mathbf{r}_i \sim \text{Multinomial}(1, \mathbf{p}), \tag{13}$$

where $\mathbf{p} - [p_1, p_2, ..., p_m]^\intercal$ are probabilities values and $\sum_k p_k - 1$. It should be noted that only one value of $\mathbf{r}_i$ is equal to 1 and the rest values are 0. Then the vector $\mathbf{r}_i$ is applied to $\mathbf{d}^k$ as

$$\mathbf{d}'^{(k)} = \mathbf{r}^{(k)} \odot \mathbf{d}^{(k)}, \tag{14}$$

where $\odot$ denotes the Hadamard product (i.e., $d_i'^{(k)} = r_i^{(k)} \cdot d_i^{(k)}$). Finally, the model combined all the features to generate a fused embedding vector $\mathbf{e} = [e_1, e_2, ..., e_c]^\intercal$, and

$$e_i = \sum_k d_i'^{(k)}, \tag{15}$$

where $\mathbf{e}$ is the final output vector for final the multimodal fusion emotion classification task. In this study, $k \in \{F, T, A\}$ and the final representation of the output vector is:

$$e_i = \sum_{k \in \{F,T,A\}} d_i'^{(k)}. \tag{16}$$

As stated in [34], the docking layers will consider the correlations between different modalities during training cause only part of the feature embedding vector for each modality will be further processed in the embracement layer. The selected features' indices are randomly changed so that each docking layer will learn to generate similar embedding vectors, and the embracement layer can generate the same output. The multinomial distribution for the selection process also acts as a regularization step, preventing the model from excessively learning from specific modalities.

5. Experiments

5.1. Computational Environment

Pytorch (version 1.8) with the CUDA version of 10.2 is utilized to develop the model and evaluate the performance of the MELD dataset. The training of the model is run on two Nvidia GeForce GTX 2080 Ti graphic cards with 11GB memory.

5.2. Training Details

The training process consists of two parts. As stated above, we first fine-tune the single modality data via three different domain-specific models. For the text modality, the inputs are the word embeddings of the sentence, including the history of a dialogue and the reply. Secondly, the input for the video modality is a sequence of face images with a fixed sequence length. Finally, the input for the audio modality is a combination of an audio waveform and spectrogram data for the audio modality.

Furthermore, for the proposed multimodal fusion model, we combine all of the extracted features from the domain-specific models before the last fully connected layers. We use the stochastic gradient descent (SGD) optimizer with a learning rate of 0.001, and the cross-entropy loss for the multiclass classification problem.

5.3. Evaluation Metrics

We evaluate the performance of multimodal emotion recognition task using these evaluation metrics: *Accuracy*, *Balanced Accuracy*, *Precision*, *Recall*, *F1* score. Using the notations of the true positive (TP), true negative (TN), false positive (FP), false negative (FN), the expression of these metrics are given as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{17}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}, \tag{18}$$

$$Precision = \frac{TP}{TP + FP}, \tag{19}$$

$$Recall = \frac{TP}{TP + FN}, \tag{20}$$

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}, \tag{21}$$

where

$$Sensitivity = \frac{TP}{TP + FN}, \tag{22}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{23}$$

It should be noted that the *Balanced Accuracy* is commonly used for evaluating imbalanced datasets; thus, it is believed that it would be effective for evaluating the MELD dataset. *Balanced Accuracy* for multiclass classification is defined as the average *Recall* obtained from each of the classes. The *Precision* shows how many positive predicted samples are truthfully positive, and the *Recall* tells how many positive samples are correctly classified as positive by the model. *F1* score takes both precision and recall into account which is the harmonic mean of *Precision* and *Recall*.

5.4. Performance Evaluation

We evaluate the performance of our fusion model via the following strategies: (1) Compare the performances of the classification for the unimodal and for the multimodal models and (2) Contrast the performances of our proposed method with existing studies. The visualization of the performance includes using confusion matrix and the feature embedding visualization of the MELD dataset using t-distributed Stochastic Neighbor Embedding (t-SNE) [59].

6. Results and Discussion

6.1. Performance Comparison between Single Modality and Multi-Modalities

There are seven emotion labels from the MELD dataset (*Neutral*, *Joy*, *Sadness*, *Anger*, *Surprise*, *Fear*, and *Disgust*, that) that were trained to be recognized. Since the MELD dataset has already been divided into training, validation, and test sets, we built our cross-modality fusion model, tuned the training hyper-parameters based on the train and validation set, and evaluated the model on the test set to get final results. Table 1 shows the performances of single modality and the proposed multi-modal fusion model. The fusion evaluation is performed by evaluating the classification results of the fused representation vectors. The fusion model outperforms all single modality models from the weighted average metrics. Specifically, the proposed fusion model

achieved a *Precision* of 63.1%, a *Recall* of 65.0%, and an *F1* score of 64.0% based on the weighted average. The text-modality model contributes the most to the final fusion results, achieving a weighted average *F1* of 61.8%.

Table 1: The performance of the classification results for MELD dataset. (PRE: precision, REC: recall, F1: F1 score).

Emotion	Modality											
	Audio			Face			Text			Multimodal		
	PRE	REC	F1	PRE	REC	F1	PRE	REC	F1	PRE	REC	F1
Neutral	48.5	99.8	65.2	48.1	99.0	64.7	72.7	82.1	77.1	74.0	84.9	79.1
Joy	45.5	1.2	2.4	15.0	0.7	1.4	54.4	53.7	54.1	56.7	57.2	56.9
Sadness	0.0	0.0	0.0	0.0	0.0	0.0	40.1	33.2	36.3	49.6	29.3	36.9
Anger	60.0	0.9	1.7	0.0	0.0	0.0	53.4	39.1	45.1	51.4	44.7	47.8
Surprise	16.7	0.4	0.7	0.0	0.0	0.0	53.0	60.0	56.3	54.4	61.4	57.7
Fear	0.0	0.0	0.0	0.0	0.0	0.0	16.1	10.0	12.3	50.0	6.0	10.7
Disgust	0.0	0.0	0.0	0.0	0.0	0.0	52.4	16.2	24.7	47.1	11.8	18.8
Weighted Avg.	40.0	48.4	43.8	25.5	47.8	33.3	61.0	62.6	61.8	**63.1**	**65.0**	**64.0**

Table 2 presents the overall performance of single and multimodal models. Evaluating the performance of each individual modality, the modality of text has the highest *Accuracy* and *Balanced Accuracy* from the experiments. However, the face modality has the lowest results. All evaluation metrics show that our multimodal fusion model outperforms the unimodal results.

Table 2: The overall classification results of single modal and multimodal.

Modality	Accuracy	Balanced Accuracy	F1
Audio	48,4	14.6	43.8
Face	47.8	14.3	33.3
Text	62.6	42.0	61.8
Multimodal	**65.0**	**42.2**	**64.0**

The inherent imbalance issue from the MELD dataset causes the low performance of the *Balanced Accuracy*, which is also reflected in the confusion matrix of the multimodal results. As can be seen in Figure 7, most of the predictions of samples tend to lie in the

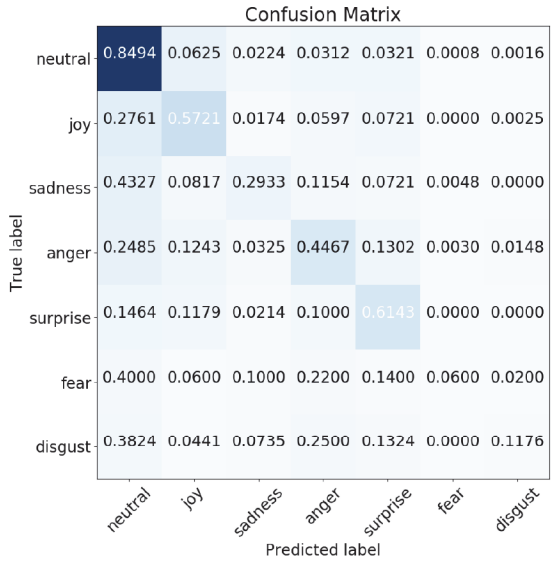

Figure 7. The confusion matrix for the multimodal results.

first column of the confusion matrix, which is the *Neutral* emotional class.The reason for this issue is that over 47% of the samples from the MELD dataset are labeled as *Neutral*. So the model is influenced by the data imbalance and learns more weights for the *Neutral* emotional class.

6.2. Comparison with Existing Studies

Table 3 compares the performance of the proposed model with the existing studies that also implemented the multimodal architecture and tested it on the MELD dataset. Most of the previous studies [36,39] only considered the audio and text modalities. However, a study by Siriwardhana et al. [38] proposed a multimodal fusion model for combining the modality of audio, face, and text and has achieved state-of-the-art results. The results from the previous studies attribute that both *Accuracy* and *F1* scores are improved by combining multiple modalities compared with a single modality. The previous results have also yielded that single modality models for both audio and face modalities cannot learn any distinct emotion features, which is also supported by the results of our experiments. Comparing with the study by [38], our proposed method achieves higher *Accuracy* and equivalent *F1* scores, which matches the state-of-the-art performance and propounds the robustness in multimodal emotion classification.

Model	Modality	Acc	F1
N. Ho et al.	A	48.8	45.3
	T	61.7	59.0
	Multimodal (A + T)	63.3	60.6
Z. Lian et al.	Audio	46.9	38.2
	Text	60.6	58.3
	Multimodal (A + T)	62.0	60.5
S. Siriwardhana et al.	Multimodal (A + F + T)	64.3	63.9
Cross-modality Fusion (proposed)	Audio	48.4	32.1
	Face	47.8	31.4
	Text	62.6	61.2
	Multimodal (A + F + T)	**65.0**	**64.0**

Table 3: The comparison of the proposed model with existing studies. (A: Audio, F: Face, T: Text)

6.3. t-SNE Visualization for the MELD dataset

Figure 8 shows the visualization of the embedding outputs from the last fully connected layer of our proposed model. The embedding vectors are projected into a 2D plan by using t-SNE [59]. As can be seen, the sparseness of the *Neutral* class spans over all the other classes, which makes training more challenging. However, it is not particularly surprising given the fact that emotion *Neutral* may contain characteristics from either emotion.

It can also be seen that the cluster of *Surprise* emotion is father away than *Neutral* emotion data points, meaning the model generate distinct features for this class, which also can be reflected from Table 1 that emotion *Surprise* get highest *F1* score among the other emotion classes except for *Neutral*.

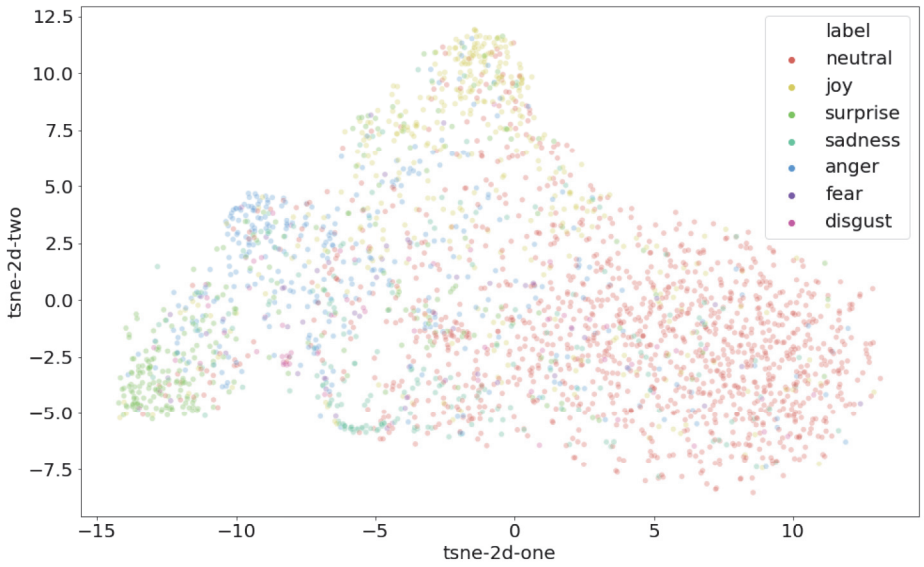

Figure 8. The t-SNE visualization for the MELD dataset.

7. Conclusions

This study demonstrates a robust multimodal emotion classification architecture, which includes crossmodal transformer fusion to combine three different modalities of input information. The architecture considers both the joint relations between the modalities and robustly fuses different sources of representation vector. Three separate prediction models are adapted to identify emotion from audio, visual, and textual inputs. Text, audio, and image inputs were trained by GPT, WaveRNN, and FaceNet+GRU, respectively. Designed transformer-based fusion mechanism with EmbraceNet has demonstrated the ability to solve the task of multimodal feature fusion from multiple pre-trained models. EmbraceNet takes the attention outputs from the crossmodal models, embraces them to build a fused representation of the emotion embedding vectors. The experiment's metrics have shown that our multimodal classification model outperforms every single modality model. However, due to innate imbalance in the dataset, *Balanced Accuracy* is generally lower than *Accuracy*. Future studies should consider introducing data augmentation techniques to handle the imbalanced data issue.

Furthermore, experimental results on the MELD datasets demonstrate the effectiveness of the proposed method. The performance of our method can reach the previous state-of-the-art strategies [38] (with 0.7% performance improvement on *Accuracy*). The performance of the *F1* score is equivalent. Nevertheless, our proposed network architecture extends the idea of multimodal emotion recognition with the crossmodal transformer, and the structure of the network can also be expanded for a higher number of input modalities. For future study, other input modalities such as different physiological measurements should also be added to this network architecture. Besides, the emotions stimulated by actors from a comedy can be exaggerated and different from real emotions, which could also lead to biased results [60]. Therefore, more multimodal datasets should be evaluated in the future work.

Author Contributions: Conceptualization, B.X. and C.H.P.; methodology, B.X.; software, B.X. and M.S.; validation, B.X. and M.S.; formal analysis, B.X.; investigation, B.X. and M.S.; resources, C.H.P.; data curation, B.X. and M.S.; writing—original draft preparation, B.X.; writing—review and editing, B.X., M.S. and C.H.P.; visualization, B.X. and M.S.; supervision, C.H.P.; project administration, C.H.P.; funding acquisition, C.H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSF 1846658: "NSF CAREER: Social Intelligence with Contextual Ambidexterity for Long-Term Human-Robot Interaction and Intervention (LT-HRI^2)."

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Acknowledgments: The authors would like to appreciate the National Science Foundation for supporting this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. Fifteenth annual conference of the international speech communication association, 2014.
2. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters* **2018**, *25*, 1440–1444.
3. Neliyanchath, A.; Paramasivam, P.; Yenigalla, P. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7179–7183.
4. Chatterjee, A.; Gupta, U.; Chinnakotla, M.K.; Srikanth, R.; Galley, M.; Agrawal, P. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior* **2019**, *93*, 309–317.
5. Batbaatar, E.; Li, M.; Ryu, K.H. Semantic-emotion neural network for emotion recognition from text. *IEEE Access* **2019**, *7*, 111866–111878.
6. Tarnowski, P.; Kołodziej, M.; Majkowski, A.; Rak, R.J. Emotion recognition using facial expressions. *Procedia Computer Science* **2017**, *108*, 1175–1184.
7. Cohen, I.; Garg, A.; Huang, T.S.; others. Emotion recognition from facial expressions using multilevel HMM. Neural information processing systems. Citeseer, 2000, Vol. 2.
8. Regenbogen, C.; Schneider, D.A.; Finkelmeyer, A.; Kohn, N.; Derntl, B.; Kellermann, T.; Gur, R.E.; Schneider, F.; Habel, U. The differential contribution of facial expressions, prosody, and speech content to empathy. *Cognition & emotion* **2012**, *26*, 995–1014.
9. Regenbogen, C.; Schneider, D.A.; Gur, R.E.; Schneider, F.; Habel, U.; Kellermann, T. Multimodal human communication—targeting facial expressions, speech content and prosody. *Neuroimage* **2012**, *60*, 2346–2356.
10. Jessen, S.; Kotz, S.A. The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage* **2011**, *58*, 665–674.
11. Müller, V.I.; Habel, U.; Derntl, B.; Schneider, F.; Zilles, K.; Turetsky, B.I.; Eickhoff, S.B. Incongruence effects in crossmodal emotional integration. *Neuroimage* **2011**, *54*, 2257–2266.
12. Stiefelhagen, R.; Ekenel, H.K.; Fugen, C.; Gieselmann, P.; Holzapfel, H.; Kraft, F.; Nickel, K.; Voit, M.; Waibel, A. Enabling multimodal human–robot interaction for the karlsruhe humanoid robot. *IEEE Transactions on Robotics* **2007**, *23*, 840–851.
13. Hong, A.; Lunscher, N.; Hu, T.; Tsuboi, Y.; Zhang, X.; dos Reis Alves, S.F.; Nejat, G.; Benhabib, B. A Multimodal Emotional Human-Robot Interaction Architecture for Social Robots Engaged in Bidirectional Communication\vspace* 7pt. *IEEE transactions on cybernetics* **2020**.
14. Kim, J.C.; Azzi, P.; Jeon, M.; Howard, A.M.; Park, C.H. Audio-based emotion estimation for interactive robotic therapy for children with autism spectrum disorder. 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI). IEEE, 2017, pp. 39–44.
15. Xie, B.; Kim, J.C.; Park, C.H. Musical emotion recognition with spectral feature extraction based on a sinusoidal model with model-based and deep-learning approaches. *Applied Sciences* **2020**, *10*, 902.
16. Maat, L.; Pantic, M. Gaze-X: Adaptive, affective, multimodal interface for single-user office scenarios. In *Artifical Intelligence for Human Computing*; Springer, 2007; pp. 251–271.
17. Kapoor, A.; Burleson, W.; Picard, R.W. Automatic prediction of frustration. *International journal of human-computer studies* **2007**, *65*, 724–736.
18. Murray, I.R.; Arnott, J.L. Synthesizing emotions in speech: Is it time to get excited? Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. IEEE, 1996, Vol. 3, pp. 1816–1819.
19. Walker, M.A.; Cahn, J.E.; Whittaker, S.J. Improvising linguistic style: Social and affective bases for agent personality. Proceedings of the first international conference on Autonomous agents, 1997, pp. 96–105.
20. Schröder, M. Emotional speech synthesis: A review. Seventh European Conference on Speech Communication and Technology, 2001.
21. France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M.; Wilkes, M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering* **2000**, *47*, 829–837.
22. Edwards, J.; Jackson, H.J.; Pattison, P.E. Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review* **2002**, *22*, 789–832.

23. Streit, M.; Wölwer, W.; Gaebel, W. Facial-affect recognition and visual scanning behaviour in the course of schizophrenia. *Schizophrenia research* **1997**, *24*, 311–317.
24. Sebe, N.; Cohen, I.; Huang, T.S. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*; World Scientific, 2005; pp. 387–409.
25. Kessous, L.; Castellano, G.; Caridakis, G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces* **2010**, *3*, 33–48.
26. Samani, H.A.; Saadatian, E. A multidisciplinary artificial intelligence model of an affective robot. *International Journal of Advanced Robotic Systems* **2012**, *9*, 6.
27. Barros, P.; Magg, S.; Weber, C.; Wermter, S. A multichannel convolutional neural network for hand posture recognition. International Conference on Artificial Neural Networks. Springer, 2014, pp. 403–410.
28. Javed, H.; Lee, W.; Park, C.H. Toward an Automated Measure of Social Engagement for Children With Autism Spectrum Disorder—A Personalized Computational Modeling Approach. *Frontiers in Robotics and AI* **2020**, *7*, 43.
29. Wu, C.H.; Lin, J.C.; Wei, W.L. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing* **2014**, *3*.
30. Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; Pantic, M. Avec 2012: the continuous audio/visual emotion challenge. Proceedings of the 14th ACM international conference on Multimodal interaction, 2012, pp. 449–456.
31. Huang, J.; Li, Y.; Tao, J.; Lian, Z.; Wen, Z.; Yang, M.; Yi, J. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, 2017, pp. 11–18.
32. Liu, M.; Wang, R.; Li, S.; Shan, S.; Huang, Z.; Chen, X. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. Proceedings of the 16th International Conference on multimodal interaction, 2014, pp. 494–501.
33. Chen, S.; Jin, Q. Multi-modal conditional attention fusion for dimensional emotion prediction. Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 571–575.
34. Choi, J.H.; Lee, J.S. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion* **2019**, *51*, 259–270.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* **2017**.
36. Ho, N.H.; Yang, H.J.; Kim, S.H.; Lee, G. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access* **2020**, *8*, 61672–61686.
37. Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Niu, M. Multimodal transformer fusion for continuous emotion recognition. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3507–3511.
38. Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S. Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion. *IEEE Access* **2020**, *8*, 176274–176285.
39. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2021**, *29*, 985–1000.
40. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, 2019, Vol. 2019, p. 6558.
41. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* **2018**.
42. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Ku, L.W.; others. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379* **2018**.
43. Bower, G.H. How might emotions affect learning. *The handbook of emotion and memory: Research and theory* **1992**, *3*, 31.
44. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **2009**, *22*, 1345–1359.
45. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training **2018**.
46. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.
47. Wolf, T.; Sanh, V.; Chaumond, J.; Delangue, C. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149* **2019**.
48. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient neural audio synthesis. International Conference on Machine Learning. PMLR, 2018, pp. 2410–2419.
49. Ito, K.; Johnson, L. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.
50. Ouyang, X.; Kawaai, S.; Goh, E.G.H.; Shen, S.; Ding, W.; Ming, H.; Huang, D.Y. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 577–582.
51. Abdulsalam, W.H.; Alhamdani, R.S.; Abdullah, M.N. Facial emotion recognition from videos using deep convolutional neural networks. *International Journal of Machine Learning and Computing* **2019**, *9*, 14–19.

52. Leong, F.H. Deep learning of facial embeddings and facial landmark points for the detection of academic emotions. Proceedings of the 5th International Conference on Information and Education Innovations, 2020, pp. 111–116.
53. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
54. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* **2016**, *23*, 1499–1503.
55. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018, pp. 67–74.
56. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* **2014**.
57. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* **2014**.
58. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation* **1997**, *9*, 1735–1780.
59. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.
60. Franzoni, V.; Vallverdù, J.; Milani, A. Errors, biases and overconfidence in artificial emotional modeling. IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume, 2019, pp. 86–90.