sensors

MDPI

Article

Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality 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, USA; bdxie@gwu.edu (B.X.); sidul001@gwu.edu (M.S.)
* 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 the MELD. In this paper, a transformer-based crossmodality fusion with the 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.

Keywords: multimodal emotion recognition; multimodal fusion; crossmodal transformer; attention mechanism

check for updates

Citation: Xie, B.; Sidulova, M.; Park, C.H. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion. *Sensors* **2021**, *21*, 4913. https://doi.org/10.3390/s21144913

Academic Editors: Soo-Hyung Kim and Gueesang Lee

Received: 1 July 2021
Accepted: 17 July 2021
Published: 19 July 2021

1. Introduction

In recent years, there has been a growing number of studies that have 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 multisensory 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], the 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 participants' engagement and made the HRI scenario more natural. For Socially Assistive Robots (SARs) 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 affective state. The results of the experiment with human participants proved that bidirectional 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 to train 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 to recognize 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 is the work of Maat and Pantic [16], where the 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 that would have relevant emotional sentiment [18,19]. Learning emotions from the speech provides a way of generating convincing emotional speech [20].

Medical specialists can also benefit from using emotion recognition systems as a diagnostic tool for a wide range of medical symptoms. For example, in the study by France et al. [21], the researchers compared the acoustic trends in the speech of healthy, depressed, and suicidal individuals. Another medical example is clinical studies of schizophrenia and flat affect, which is an indicative symptom of the disease and is characterized as diminished emotional expression [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 to investigate how emotions are expressed by a robotic system and 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 were extracted from domain-specific models and fused via our proposed crossmodality 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 reached 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 obtain a joint representation, and the concatenated features are fed into a single classifier for emotion recognition. Schuller et al. [30] presented 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. A recent study [31] also investigated using a 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 obtain 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 a neural network for classification tasks. The model architecture ensured both the effectiveness of the crossmodal relationship and the robustness of fusing different modalities.

2.2. Transformer Method

Multimodal Emotion Recognition (MMER) with fusion by the 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 focused on using attention mechanisms to fuse different modalities of features for MMER [36–39]. Ho et al. [36] proposed a multimodal approach based on a multilevel 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 improved the emotion recognition performance compared 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 had better performance via feature-level fusion.

The previous study investigated the use of the crossmodal transformer to reinforce a target modality by introducing the features from another modality, which also learns 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 demonstrated 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 intermodality 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 the need arises to combine three or more modalities due to the structure of multi-head attention. For this reason, most previous studies on MMER based on the attention mechanism proposed and tested the network architecture for only two modalities [36,37,39]. The study [38] deployed models for combining three modalities with transformer-based fusion, but a simple feature concatenation was added at the end to combine different modalities' features. However, in practice, concatenating different high-dimensional features may result in data sparseness and degrade the performance [29].

3. Dataset

The Multimodal Emotion Lines Dataset (MELD) [41] is an extended version of the EmotionLines dataset [42]. The MELD includes a total of 13,708 samples segmented out from the TV series *Friends*, with the samples grouped as 9989 for training, 1109 for validation, and 2610 for testing. Each segmented sample has the following data attributes, which were 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 the training, validation and test datasets can be seen in Figure 1. There is an inherent imbalance in the MELD: *Neutral* samples dominate in each of the datasets. Utterance audio clips vary in time length, but are averaged to around 5 s-long recordings. On average, there are around seven words per one utterance. The average duration of an utterance is 3.95 s, and the average number of utterances per dialogue is 9.6. The number of emotions per dialogue is 3.3 on average.

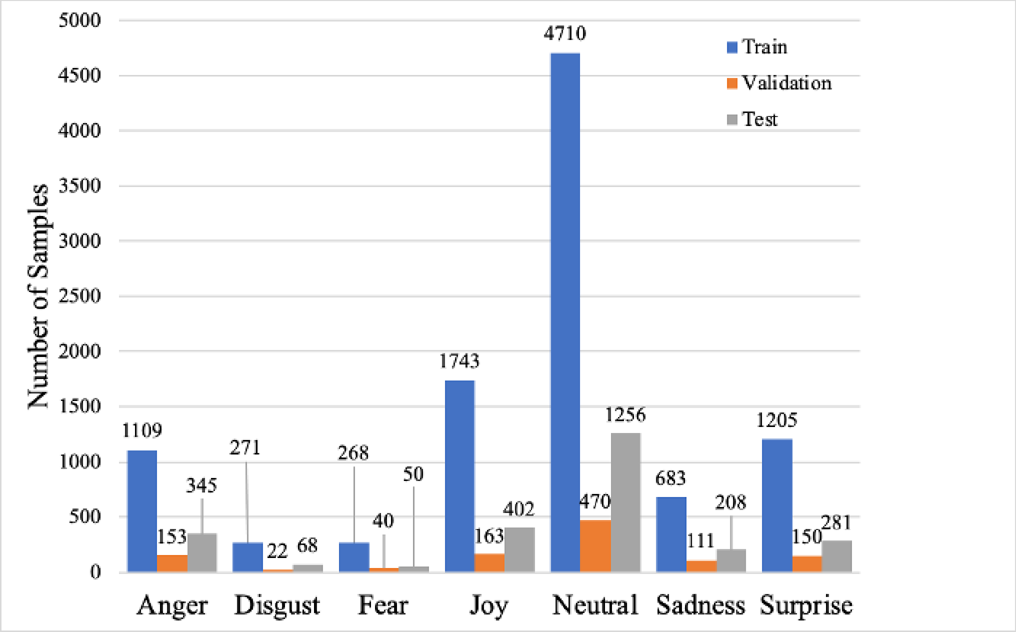


Figure 1. Number of samples per one emotion label for the training, validation, and test dataset.

4. Methods

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

4.1. Feature Extraction

Three different deep neural networks were 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 proposes first pre-training the model on a large-scale dataset, then for another smaller target dataset, the pre-trained model can be fine-tuned and achieve the desired performance efficiently. Therefore, the transfer learning paradigm was adopted in this study. In the stage of feature extraction, we first fine-tuned the domain-specific models for the emotion classification task with a single modality. Then, we extracted the features before the fully connected layer from the fine-tuned models, which were prepared 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 accepts the audio features from the sentence's audio clip, and the sequential face images from the video clip are 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 used the GPT [45] as our language model for extracting the features from the text modality. GPT is a multilayer transformer-based model, which was pre-trained on the 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 to their repository (https://github.com/huggingface/transfer-learning-conv-ai, accessed on 18 July 2021) for more details. To estimate the emotional states of the current sentence, we combined the history of the dialogues, the current sentence, and the reply sentence to generate sentence embeddings. Then, the position embeddings and segment embeddings were summed up to build a single input sequence for the transformer model to obtain the output sequence tokens.

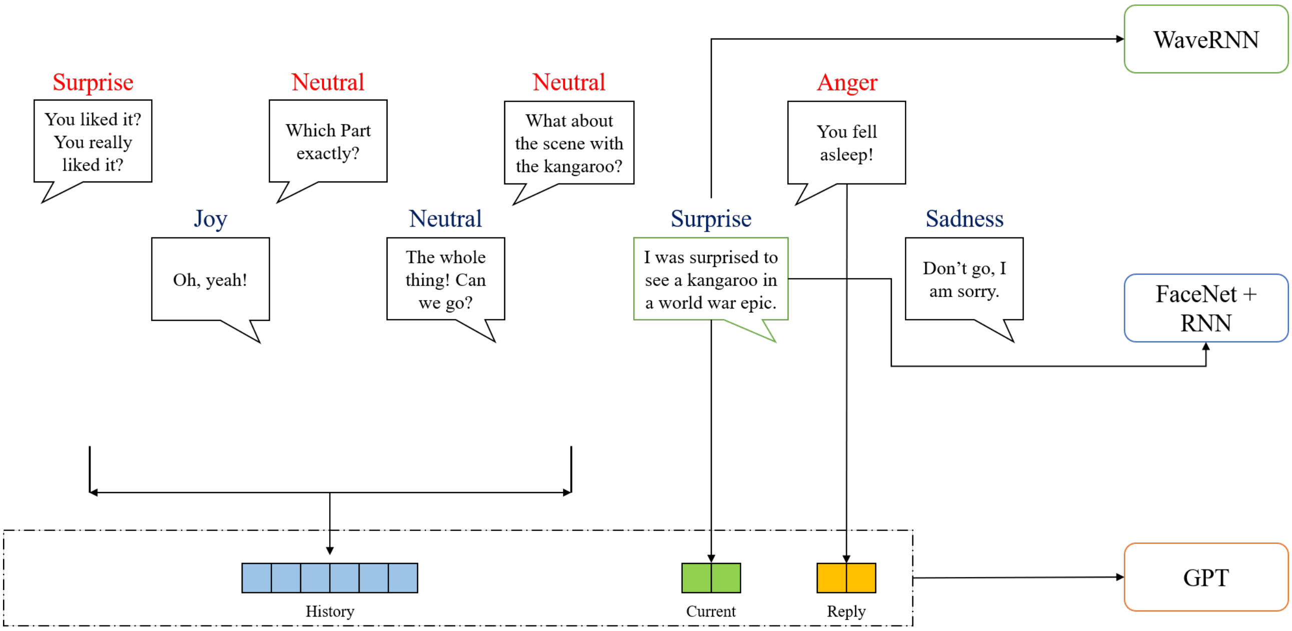


Figure 2. The crossmodality attention fusion transformer.

The optimization of the model was performed by multi-task learning. The fine-tuning was performed by introducing a combination of different loss functions: (1) next-sentence prediction loss, (2) language modeling loss, and (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 was to train a classifier to discriminate the gold reply (ground-truth sentence) from the randomly sampled distractor (the sentence other than the 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 multilayer classifier is introduced for classifying seven emotion labels from the MELD. 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:

$$L(total) = L(s) + L(l) + L(e). \tag{1}$$

4.1.2. Model for Audio Modality

The WaveRNN [48] model was used as an audio model to extract features from original audio clips. The original waveform of the audio and generated spectrogram of the signal are both inputs in the model. The pre-trained model used in this study was trained on LJSpeech dataset [49]. During the preprocessing, waveforms were sampled at a 16,000 Hz sample rate, and spectrograms were generated. Moreover, zero-padding for the batch inputs was applied on-the-fly during training. The input channels of the waveform and spectrogram have to be 1, so we averaged the values from both channels of the audio inputs. All the preprocessing was implemented via the torchaudio library (https://pytorch.org/audio/stable/index.html, accessed on 18 July 2021). As shown in Figure 2, the acoustic features from the time interval of one person's speech were 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 $[(n\_time - kernel\_size + 1) \times hop\_length =$ 47,200,512] before the fully connected layer of the WaveRNN model is extracted, where 512 is the default feature size of the WaveRNN output. The classifier consists of two consecutive linear layers at the end, giving us the predictions of the emotional states.

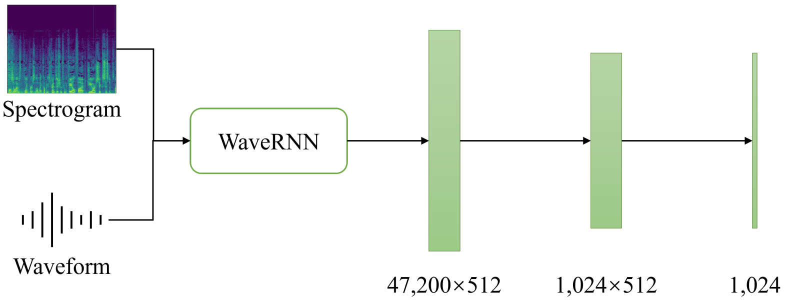


Figure 3. The classifier added to WaveRNN.

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 showed that by combining the deep CNN and temporal models, one could 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 (https://github.com/timesler/facenet-pytorch, accessed on 18 July 2021) implemented with Pytorch designed for face verification, recognition, and feature extraction. The Multi-task CNN (MTCNN) [54] recognizes a face in the video frames and returns the sequence of the cropped images with the detected face from the video input. Afterward, we used the Inception ResNet (V1) model, which was pre-trained on the 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 obtain a sequence of face embeddings during the time interval of one person's speech.

Furthermore, we incorporated an RNN-based model to learn the temporal relations in the sequence of images. The Gated Recurrent Unit (GRU) [57] is similar to the LSTM [58] network with a forget gate, but is more efficient in 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 were downsampled 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 was set to 50 images. Oppositely, zero-padding was applied during data preprocessing to the sequences with a shorter length. Finally, during training, the last hidden state value, h_t, of the GRU outputs was extracted, and a classifier was introduced to learn the emotion classification task.

4.2. *Robust Crossmodality Fusion Transformer with EmbraceNet*

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

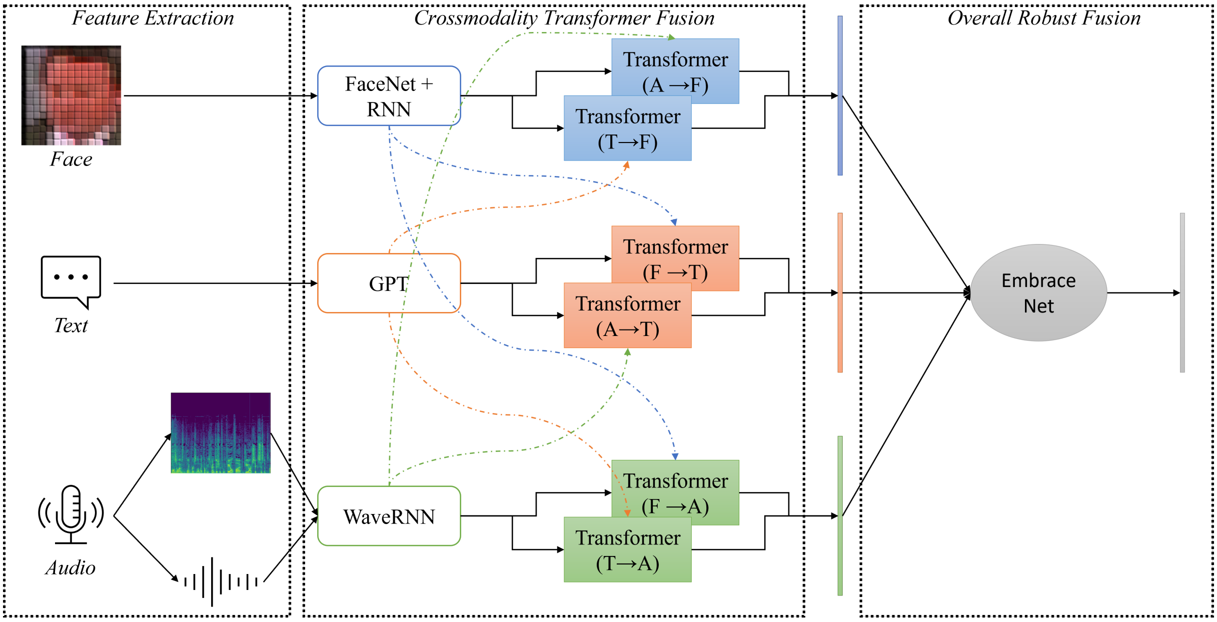


Figure 4. The crossmodal fusion transformer [40] with EmbraceNet [34].

4.2.1. Crossmodality 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 adapted 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 "$A \to B$". The multi-head attention was proposed in the work [35], where the attention function was mapping the query, keys, and values to the output.

Figure 5 shows the architecture of the crossmodal transformer. We denote the input features for Modalities A and B as $X_A \in \mathbb{R}^{T_A \times d_A}, X_B \in \mathbb{R}^{T_A \times d_B}$, where T and d are the 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 $K_B = X_B W_{K_B}$, and 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:

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

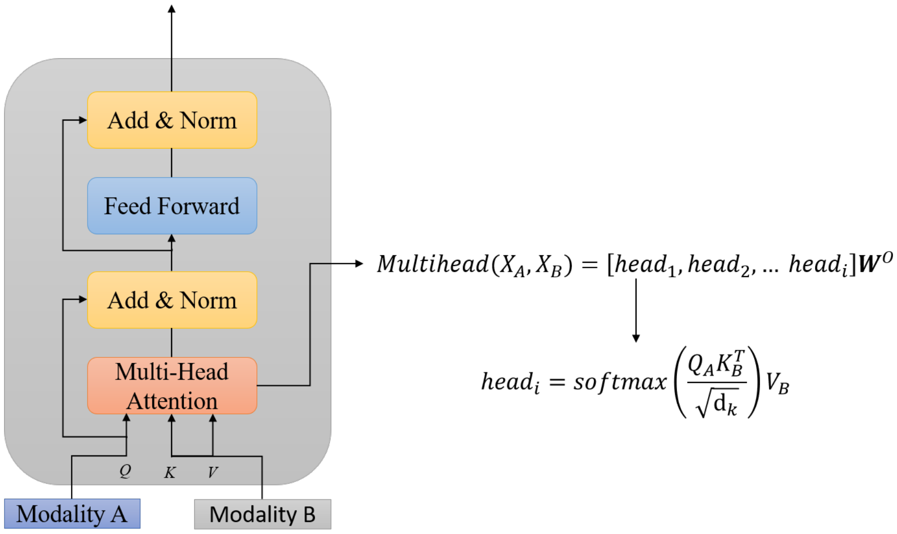

Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Q
K
V
Modality A
Modality B
$Multihead(X_A, X_B) = [head_1, head_2, \ldots\ head_i]\boldsymbol{W}^O$
$head_i = \mathit{softmax}\left(\frac{Q_A K_B^T}{\sqrt{\mathrm{d}_k}}\right)V_B$

Figure 5. The architecture of the crossmodal transformer.

Following the setting from the previous study [35], we also added a residual connection from the query to the attention output and layer normalization.

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

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

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

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

$$f_{x_a} = Linear(x_a) = x_a {A'_{x_a}}^\intercal + b_{xa}, \tag{10}$$

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 was used to obtain 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 the text modality, the final attention representation of the feature was the Hadamard product of two crossmodality 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, 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 the query, the target modality. The previous study showed 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 not only considered crossmodal fusion by using the transformer but also wanted to ensure the robustness of combining the multimodal outputs. We employed the EmbraceNet [34] in our network architecture, which focuses on dealing with crossmodal information carefully and avoids performance degradation due to the partial absence of data.

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

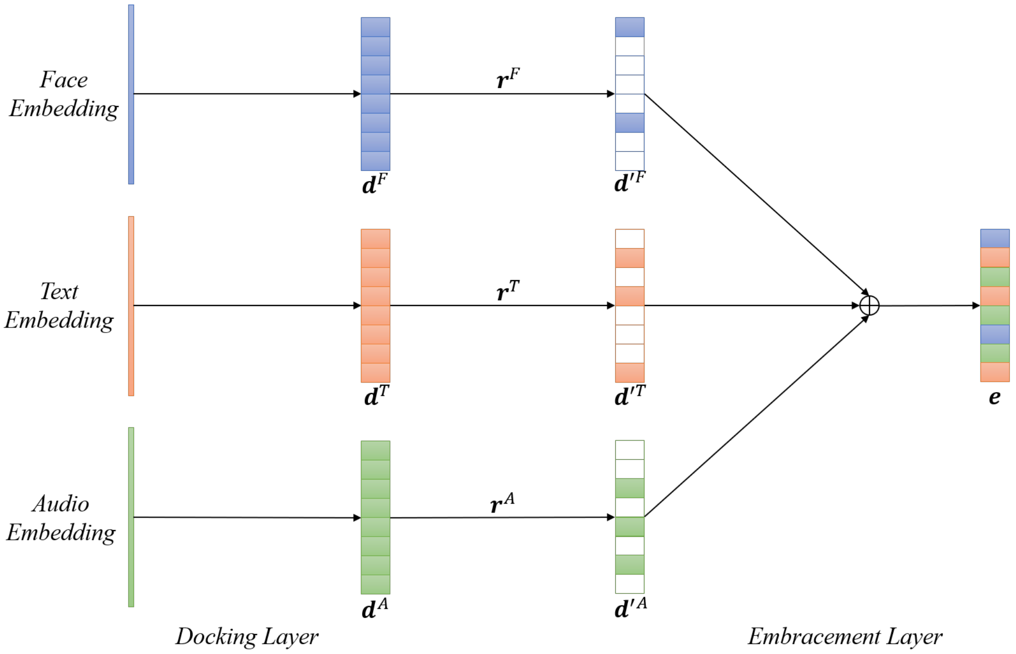


Figure 6. The EmbraceNet for multimodal fusion.

4.2.3. Docking Layers

Three different deep learning models were used to obtain the feature vectors from the 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 to 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 was 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 the probability 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 of the 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 combines all the features to generate a fused embedding vector $\mathbf{e} = [e_1, e_2, ..., e_c]^\intercal$, and:

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

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

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

As stated in [34], the docking layers that consider the correlations between different modalities during training cause only part of the feature embedding vector for each modality to 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 CUDA Version 10.2 was utilized to develop the model and evaluate the performance of the MELD. The training of the model was run on two Nvidia GeForce GTX 2080 Ti graphic cards with 11 GB of memory.

5.2. Training Details

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

Furthermore, for the proposed multimodal fusion model, we combined all of the extracted features from the domain-specific models before the last fully connected layers. We used 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 evaluated the performance of multimodal emotion recognition task using the following evaluation metrics: *Accuracy*, *Balanced Accuracy*, *Precision*, *Recall*, and *F1*-score.

Using the notions of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the expression of these metrics is 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 was believed that it would be effective for evaluating the MELD. *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. The *F1*-score takes both the Precision and Recall into account, which is the harmonic mean of the *Precision* and *Recall*.

5.4. Performance Evaluation

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

6. Results and Discussion

6.1. Performance Comparison between Single Modality and Multiple Modalities

Since the MELD has already been divided into training, validation, and test sets, we built our crossmodality fusion model, tuned the training hyperparameters based on the training and validation set, and evaluated the model on the test set to obtain the final results. Table 1 shows the performances of the single modality and the proposed multimodal fusion model. The fusion evaluation was performed by evaluating the classification results of the fused representation vectors. The fusion model outperformed 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 contributed the most to the final fusion results, achieving a weighted average *F1*-score of 61.8%.

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

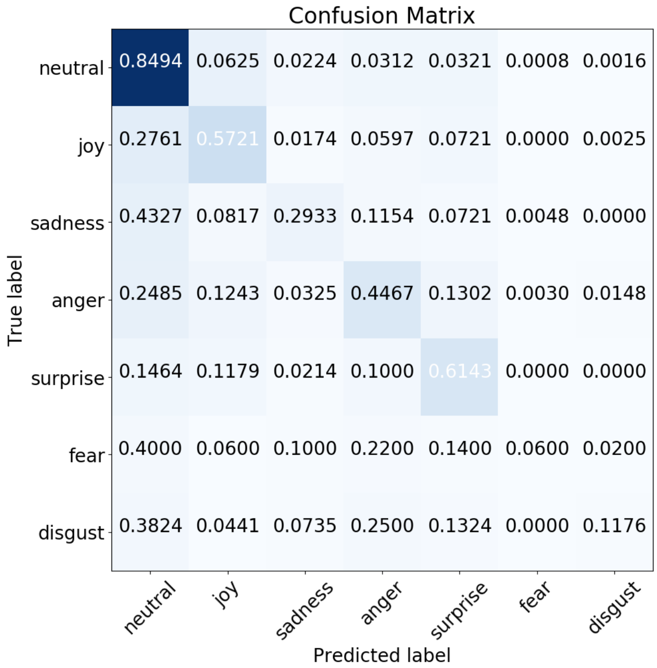
Table 1. The performance of the classification results for the MELD (PRE: Precision, REC: Recall, F1: F1-score) (Unit = %).

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. The overall classification results of the single modality and multimodal methods (Unit = %).

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 caused 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 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 are labeled as *Neutral*. Therefore, the model is influenced by the data imbalance and learns more weights for the *Neutral* emotional class.



	neutral	joy	sadness	anger	surprise	fear	disgust
neutral	0.8494	0.0625	0.0224	0.0312	0.0321	0.0008	0.0016
joy	0.2761	0.5721	0.0174	0.0597	0.0721	0.0000	0.0025
sadness	0.4327	0.0817	0.2933	0.1154	0.0721	0.0048	0.0000
anger	0.2485	0.1243	0.0325	0.4467	0.1302	0.0030	0.0148
surprise	0.1464	0.1179	0.0214	0.1000	0.6143	0.0000	0.0000
fear	0.4000	0.0600	0.1000	0.2200	0.1400	0.0600	0.0200
disgust	0.3824	0.0441	0.0735	0.2500	0.1324	0.0000	0.1176

Figure 7. The confusion matrix for the multimodal results.

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. Most of the previous studies [36,39] only considered the audio and text modalities. However, the study by Siriwardhana et al. [38] proposed a multimodal fusion model for combining the modality of audio, face, and text and achieved state-of-the-art results. The results from the previous studies demonstrated that both the *Accuracy* and *F1*-scores were improved by combining multiple modalities compared with a single modality. The previous results 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. Compared to 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.

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

Model	Modality	Acc	F1
N. Ho et al. [36]	A	48.8	45.3
	T	61.7	59.0
	Multimodal (A + T)	63.3	60.6
Z. Lian et al. [39]	Audio	46.9	38.2
	Text	60.6	58.3
	Multimodal (A + T)	62.0	60.5
S. Siriwardhana et al. [38]	Multimodal (A + F + T)	64.3	63.9
Robust Crossmodality 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**

6.3. *t-SNE Visualization for the MELD*

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, this is not particularly surprising given the fact that the *Neutral* emotion may contain characteristics from either emotion.

It can also be seen that the cluster of the *Surprise* emotion is far away from the cluster of *Neutral* emotion data points, meaning the model generates distinct features for this class, which are also reflected in Table 1, showing that the *Surprise* emotion obtains the highest *F1*-score among the other emotion classes, except for *Neutral*.

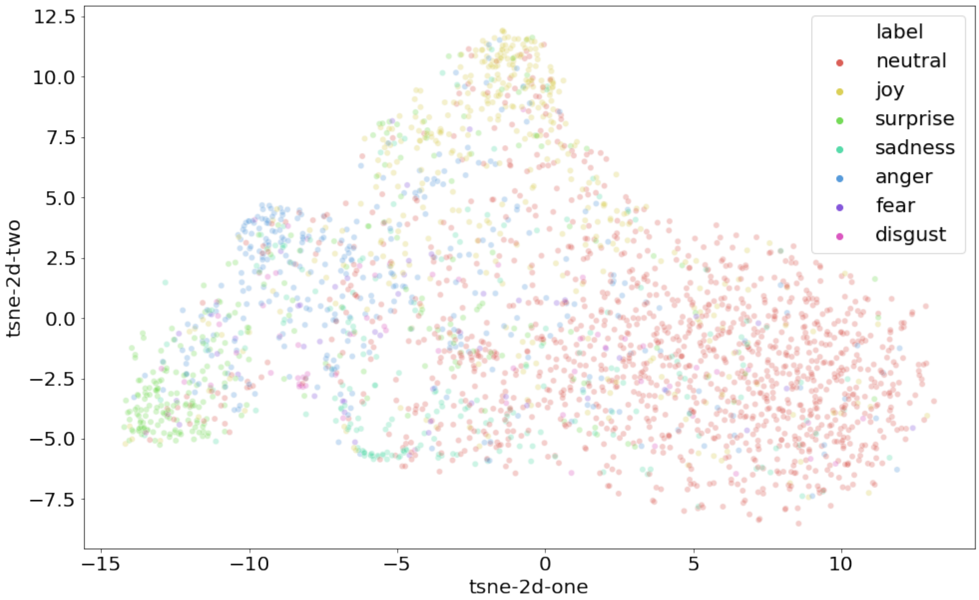


Figure 8. The t-SNE visualization for the MELD.

7. Conclusions

This study demonstrated a robust multimodal emotion classification architecture, which included crossmodal transformer fusion to combine three different modalities of the input information. The architecture considers both the joint relations among the modalities and robustly fuses different sources of the representation vector. Three separate prediction models were adapted to identify emotion from audio, visual, and textual inputs. Text, audio, and image inputs were trained by GPT, WaveRNN, and FaceNet+GRU, respectively. The designed transformer-based fusion mechanism with EmbraceNet 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 and embraces them to build a fused representation of the emotion embedding vectors. The experiment's metrics showed 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 demonstrated the effectiveness of the proposed method. The performance of our method can reach the previous state-of-the-art strategies [38] (with a 0.7% performance improvement on *Accuracy*). The performance of the *F1*-score is equivalent. Nevertheless, our proposed network architecture extends the previous studies of multimodal emotion recognition with the crossmodal transformer [38,40], and the structure of the network can also be robustly expanded for a higher number of input modalities. For future studies, 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 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 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.

Data Availability Statement: The MELD dataset used in this study can be found at https://affective-meld.github.io/ (accessed on 18 July 2021).

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. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 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 Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
3. Nediyanchath, A.; Paramasivam, P.; Yenigalla, P. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 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. *Comput. Hum. Behav.* **2019**, *93*, 309–317. [CrossRef]
5. Batbaatar, E.; Li, M.; Ryu, K.H. Semantic-emotion neural network for emotion recognition from text. *IEEE Access* **2019**, *7*, 111866–111878. [CrossRef]
6. Tarnowski, P.; Kołodziej, M.; Majkowski, A.; Rak, R.J. Emotion recognition using facial expressions. *Procedia Comput. Sci.* **2017**, *108*, 1175–1184. [CrossRef]
7. Cohen, I.; Garg, A.; Huang, T.S. Emotion recognition from facial expressions using multilevel HMM. In *Neural Information Processing Systems*; Citeseer: State College, PA, USA, 2000; Volume 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. *Cogn. Emot.* **2012**, *26*, 995–1014. [CrossRef] [PubMed]
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. [CrossRef] [PubMed]
10. Jessen, S.; Kotz, S.A. The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *Neuroimage* **2011**, *58*, 665–674. [CrossRef] [PubMed]
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. [CrossRef]
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 Trans. Robot.* **2007**, *23*, 840–851. [CrossRef]
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 Bi-directional Communication. *IEEE Trans. Cybern.* **2020**. [CrossRef]
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. In Proceedings of the 2017 IEEE 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Jeju, Korea, 28 June–1 July 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. *Appl. Sci.* **2020**, *10*, 902. [CrossRef]
16. Maat, L.; Pantic, M. Gaze-X: Adaptive, affective, multimodal interface for single-user office scenarios. In *Artifical Intelligence for Human Computing*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 251–271.
17. Kapoor, A.; Burleson, W.; Picard, R.W. Automatic prediction of frustration. *Int. J. Hum. Comput. Stud.* **2007**, *65*, 724–736. [CrossRef]
18. Murray, I.R.; Arnott, J.L. Synthesizing emotions in speech: Is it time to get excited? In Proceedings of the IEEE Fourth International Conference on Spoken Language Processing (ICSLP'96), Philadelphia, PA, USA, 3–6 October 1996; Volume 3, pp. 1816–1819.
19. Walker, M.A.; Cahn, J.E.; Whittaker, S.J. Improvising linguistic style: Social and affective bases for agent personality. In Proceedings of the First International Conference on Autonomous Agents, Marina Del Rey, CA, USA, 5–8 February 1997; pp. 96–105.
20. Schröder, M. Emotional speech synthesis: A review. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 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 Trans. Biomed. Eng.* **2000**, *47*, 829–837. [CrossRef]

22. Edwards, J.; Jackson, H.J.; Pattison, P.E. Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review. *Clin. Psychol. Rev.* **2002**, *22*, 789–832. [CrossRef]
23. Streit, M.; Wölwer, W.; Gaebel, W. Facial-affect recognition and visual scanning behaviour in the course of schizophrenia. *Schizophr. Res.* **1997**, *24*, 311–317. [CrossRef]
24. Sebe, N.; Cohen, I.; Huang, T.S. Multimodal emotion recognition. In *Handbook of Pattern Recognition and Computer Vision*; World Scientific: Singapore, 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. *J. Multimodal User Interfaces* **2010**, *3*, 33–48. [CrossRef]
26. Samani, H.A.; Saadatian, E. A multidisciplinary artificial intelligence model of an affective robot. *Int. J. Adv. Robot. Syst.* **2012**, *9*, 6. [CrossRef]
27. Barros, P.; Magg, S.; Weber, C.; Wermter, S. A multichannel convolutional neural network for hand posture recognition. In *International Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 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. *Front. Robot. AI* **2020**, *7*, 43. [CrossRef] [PubMed]
29. Wu, C.H.; Lin, J.C.; Wei, W.L. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, e12. [CrossRef]
30. Schuller, B.; Valster, M.; Eyben, F.; Cowie, R.; Pantic, M. Avec 2012: The continuous audio/visual emotion challenge. In Proceedings of the 14th ACM International Conference on Multimodal Interaction, Santa Monica, CA, USA, 22–26 October 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. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, 23–27 October 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. In Proceedings of the 16th International Conference on multimodal interaction, Istanbul, Turkey, 12–16 November 2014; pp. 494–501.
33. Chen, S.; Jin, Q. Multi-modal conditional attention fusion for dimensional emotion prediction. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 571–575.
34. Choi, J.H.; Lee, J.S. EmbraceNet: A robust deep learning architecture for multimodal classification. *Inf. Fusion* **2019**, *51*, 259–270. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
36. Ho, N.H.; Yang, H.J.; Kim, S.H.; Lee, G. Multimodal approach of speech emotion recognition using multi-level multihead fusion attention-based recurrent neural network. *IEEE Access* **2020**, *8*, 61672–61686. [CrossRef]
37. Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Niu, M. Multimodal transformer fusion for continuous emotion recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 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. [CrossRef]
39. Lian, Z.; Liu, B.; Tao, J. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 985–1000. [CrossRef]
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. *Proc. Conf. Assoc. Comput. Linguist. Meet.* **2019**, *2019*, 6558–6569.
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* **2018**, arXiv:1810.02508.
42. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Ku, L.W. Emotionlines: An emotion corpus of multi-party conversations. *arXiv* **2018**, arXiv:1802.08379.
43. Bower, G.H. How might emotions affect learning. *Handb. Emot. Mem. Res. Theory* **1992**, *3*, 31.
44. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
45. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 18 July 2021).
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. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 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* **2019**, arXiv:1901.08149.
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. In International Conference on Machine Learning. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.

49. Ito, K.; Johnson, L. The LJ Speech Dataset. 2017. Available online: https://keithito.com/LJ-Speech-Dataset/ (accessed on 18 July 2021).
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. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 577–582.
51. Abdulsalam, W.H.; Alhamdani, R.S.; Abdullah, M.N. Facial emotion recognition from videos using deep convolutional neural networks. *Int. J. Mach. Learn. Comput.* **2019**, *9*, 14–19. [CrossRef]
52. Leong, F.H. Deep learning of facial embeddings and facial landmark points for the detection of academic emotions. In Proceedings of the 5th International Conference on Information and Education Innovations, London, UK, 26–28 July 2020; pp. 111–116.
53. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 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 Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
55. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74.
56. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
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* **2014**, arXiv:1406.1078.
58. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
59. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
60. Franzoni, V.; Vallverdù, J.; Milani, A. Errors, biases and overconfidence in artificial emotional modeling. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume, Thessaloniki, Greece, 14–17 October 2019; pp. 86–90.