information

MDPI

Article

Feature Extraction Network with Attention Mechanism for Data Enhancement and Recombination Fusion for Multimodal Sentiment Analysis

Qingfu Qi [1,2], Liyuan Lin [1] and Rui Zhang [1,2,*]

[1] College of Electronic Information and Automation, Tianjin University of Science & Technology, Tianjin 300222, China; qqf@mail.tust.edu.cn (Q.Q.); linly@tust.edu.cn (L.L.)
[2] School of Software and Communications, Tianjin Sino-German University of Applied Sciences, Tianjin 300222, China
* Correspondence: zhangrui@tsguas.edu.cn; Tel.: +86-186-2208-5255

Abstract: Multimodal sentiment analysis and emotion recognition represent a major research direction in natural language processing (NLP). With the rapid development of online media, people often express their emotions on a topic in the form of video, and the signals it transmits are multimodal, including language, visual, and audio. Therefore, the traditional unimodal sentiment analysis method is no longer applicable, which requires the establishment of a fusion model of multimodal information to obtain sentiment understanding. In previous studies, scholars used the feature vector cascade method when fusing multimodal data at each time step in the middle layer. This method puts each modal information in the same position and does not distinguish between strong modal information and weak modal information among multiple modalities. At the same time, this method does not pay attention to the embedding characteristics of multimodal signals across the time dimension. In response to the above problems, this paper proposes a new method and model for processing multimodal signals, which takes into account the delay and hysteresis characteristics of multimodal signals across the time dimension. The purpose is to obtain a multimodal fusion feature emotion analysis representation. We evaluate our method on the multimodal sentiment analysis benchmark dataset CMU Multimodal Opinion Sentiment and Emotion Intensity Corpus (CMU-MOSEI). We compare our proposed method with the state-of-the-art model and show excellent results.

Keywords: multimodal; natural language processing; sentiment analysis; time delay; hysteresis

Citation: Qi, Q.; Lin, L.; Zhang, R. Feature Extraction Network with Attention Mechanism for Data Enhancement and Recombination Fusion for Multimodal Sentiment Analysis. *Information* **2021**, *12*, 342. https://doi.org/10.3390/info12090342

Academic Editors: Salud María Jiménez-Zafra and Miguel Ángel García Cumbreras

Received: 14 July 2021
Accepted: 20 August 2021
Published: 24 August 2021

1. Introduction

With the development of virtual community [1] and multimedia platforms such as YouTube and Facebook, people tend to discuss topics in videos rather than individual texts or pictures [2]. They usually share their opinions, stories, and comments on these media sites in the form of videos. For sentiment analysis, the information displayed in the video is multimodal and is more intuitive than unimodal text information or picture information. We can clearly understand the user's emotions and intentions from the video, and this emotional information can help us understand user feedback and user preferences. Furthermore, when we perform public opinion discovery or product recommendation and emotional subtasks, the emotional information extracted from the video can better reflect its authenticity. Therefore, multimodal sentiment analysis has become an important research field in Natural Language Processing. It has also become the basic research content of other subtasks in the NLP field, for example, video description generation [3,4], visual question answering [5,6], multimodal machine translation [7], and visual dialog [8,9].

The method in which video provides content can be summarized into three information modes: visual, audio, and language. Traditional sentiment analysis research focuses more on unimodal information, such as visual information [10,11], audio information [12],

and language information [13], and has achieved satisfactory results. However, for some of the above subtasks, it is not appropriate to use unimodal information for sentiment analysis. Compared with the multimodal information given in the video, unimodal sentiment analysis obtains the user's sentiment information from a single information channel, ignoring the rest of the behavioral clues given in the video, so unilateral research has difficulty achieving the ideal result. Therefore, multimodal sentiment analysis is needed, and its main purpose is to extend sentiment analysis based on unimodal information to sentiment analysis based on multimodal (language, visual, audio) information. In videos, people express their emotions through the interaction of multiple modes of behavior, and these behaviors are intertwined with each other. When we perform sentiment analysis, it becomes crucial to accurately capture the relationship between multiple modalities. The core challenge of multimodal sentiment analysis is to process multimodal information simultaneously to model the information within the model and the interactive information between the models and to obtain a characteristic representation that can symbolize general information.

At present, the research methods that simultaneously consider the internal model information and the interactive information between the models [14–16] have achieved good results. Some methods [15,16] involve a multi-task learning framework that can obtain similar information between multiple tasks in the model by sharing internal parameters. Some methods [14] use hierarchical fusion strategies to gradually extract feature representations from multimodal signals and finally obtain fusion feature representations for sentiment analysis. However, the common feature of these research methods is that they all use feature-level cascade fusion when processing multimodal information at a single time step. That is, when the next operation is performed, multiple modal information is directly spliced on the feature dimension. When this strategy is used for multimodal information fusion in the middle layer, its potential meaning is to place each mode in an equally important position and give the same weight. This method does not highlight the important modal information, which results in ineffective use of the information in each modality, and it is difficult to obtain a representative multimodal feature representation.

Therefore, this paper proposes a multimodal sequence feature extraction network (MSFE), a model for human multimodal language. The model uses a method of enhancing the contextual information within the unimodal and the contextual information across time between multiple modalities to obtain embedding representations of different modal intensities. This method also fully considers the time delay and hysteresis between the interwoven multimodal signals. The core of the model is to process the multimodal sequence data with sequence reorganization and modal enhancement when dealing with this kind of non-synchronized multimodal data, and directly extracts the fusion feature representation of the information-enhanced multimodal sequence. To verify our proposed strategy, we conducted multimodal sentiment analysis and sentiment recognition experiments. Our model shows excellent performance on both tasks.

This article is organized as follows. In Section 2, we introduce some related work on multimodal emotion recognition. In Section 3, we elaborate on the overall architecture of our model. In Section 4, we describe in detail the dataset used in the experiment and report the results and necessary analysis.In Section 5, the experimental results are shown and the performance is discussed. At the same time, the model is qualitatively analyzed on some samples. We conclude this paper with Section 6.

2. Related Work

In this part, we elaborated on the related work of multimodal sentiment analysis and the basic models used.

2.1. Multimodal Sentiment Analysis

Multimodal sentiment analysis models verbal and nonverbal information to analyze the emotions expressed by people. The current multimodal sentiment analysis is based on the following three modalities: language, visual, and audio. The earliest research work

on multimodal sentiment analysis was carried out on two modal pieces of information. In [17], the author combined audio and language modal signals for emotional research. In [18,19], the author combined audio and visual information for multimodal sentiment analysis research. The author of [20] proposed a new method of multimodal sentiment analysis that uses deep neural networks to combine vision and language. The methods mentioned above all show that the fusion research of two modal signals produces higher accuracy than any unimodal signal model.

With the emergence of three-modal signal datasets, scholars began to study three-modal fusion models. In the early research, researchers proposed many multimodal data fusion methods. They can be roughly divided into late fusion and early fusion. Late fusion [21,22] independently trains a monomodal classifier and perform decision voting through weighted average. Early fusion is the feature-level fusion mentioned above, which directly connects multimodal data in the time dimension. This method is used in many models. In [23], the author directly performs input-level feature fusion on multimodal data in the input stage and combines deep neural networks for sentiment analysis. In literature [24,25], the author first encodes each modal separately and then uses feature-level fusion in the middle layer to obtain multimodal embedding, which is static feature-level fusion. In [26,27], the author also encodes unimodal data, the difference is that unimodal encoding and feature-level fusion are carried out at the same time, which is dynamic feature-level fusion.

However, as mentioned before, the feature-level fusion defaults that each modal has are equally important, that is, each modal information is given the same weight at the same time step. With further research [28,29], people find that different modal information has different meanings to the final result, and the degree of importance is different. In [28], the author mentioned that verbal data will be affected by non-verbal data and cause word meaning shift. The author takes the verbal modality as the dominant and the non-verbal modality as the auxiliary for multimodal fusion. Based on this research idea, this paper proposes a multimodal sequence feature extraction network, which uses a new sequence fusion method and enhances different modal information to study the problem of multimodal emotion recognition.

2.2. Recurrent Neural Network

Recurrent Neural Network (RNN) is a type of chain-connected neural network. Compared with general neural networks, all RNNs have a chain form of repeated neural networks, which can process data that changes in sequence. However, due to the simple chain structure (for example, the Tanh layer) of ordinary RNNs, gradient disappearance and gradient explosion problems will occur during the training process, which makes it impossible to process too long sequence data and can only perform short-term memory. Therefore, Long Short-Term Memory (LSTM) [30] was proposed, which is a special RNN. The LSTM network combines short-term memory with long-term memory through sophisticated gate control, solves the problem of gradient disappearance to a certain extent, and can learn long-term dependent information. LSTM controls the state of the memory unit through input gates, forget gates, and output gates to achieve the effect of long-term memory. Gate Recurrent Units (GRU) [31] is a variant based on LSTM. It controls the state of the memory unit by reset gate and update gate. The reset gate determines how to combine the new input information with the previous memory, and the update gate defines the amount of the previous memory saved to the current time step. Through these two gating mechanisms, the information in the long-term sequence can be preserved. Compared with LSTM, GRU has fewer parameters, its convergence speed is faster, and the performance gap between the two is very small, which can greatly speed up the iterative process of the experiment. Therefore, this paper chooses to use the GRU network to solve the long-term dependencies within and between modalities.

3. Proposed Approach

In this section, we will explain in detail our Multimodal Sequence Feature Extraction Network (MSFE), which is mainly composed of the following three parts: (1) Context-Unimodal Information Extraction Module, which consists of multiple independent Gate Recurrent Units (GRU) composition, is used to encode the long-term dependencies within the modal, and obtain the characteristic representation with unimodal context information at every time step. (2) Information Enhancement and Data Reorganization Module, which enhances and reorganizes the acquired contextual information representation to obtain multimodal feature fusion representation. (3) Sentiment Analysis Layer, the fusion information feature representation obtained above is used for sentiment prediction and emotions classification through the fully connected layer. Figure 1 shows the overall architecture of MSFE.

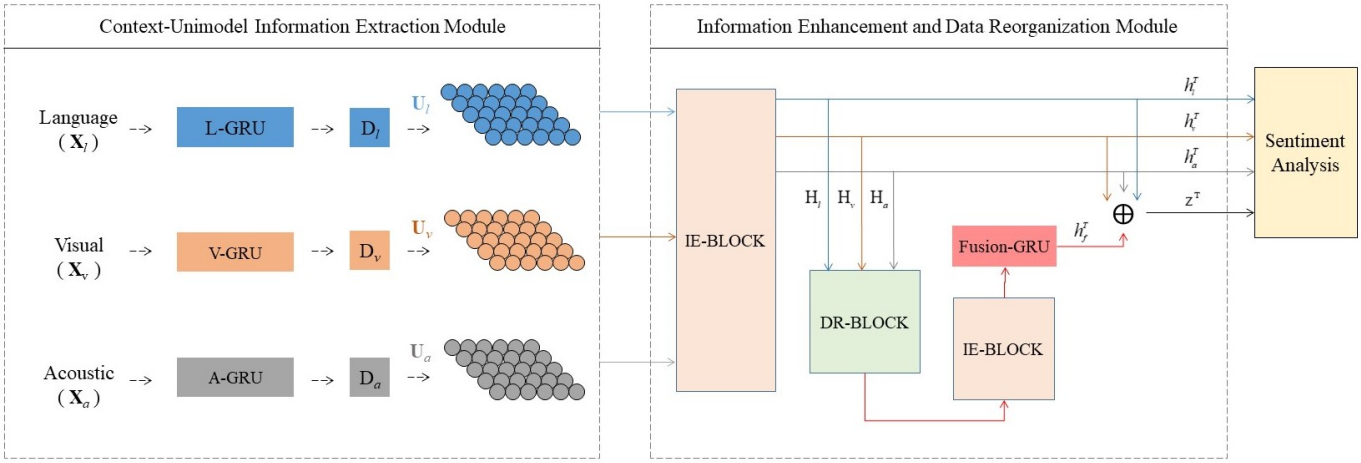

Figure 1. The overall architecture of the Multimodal Sequence Feature Extraction Network (MSFE). It consists of three modules: (1) Context-Unimodal Information Extraction Module is composed of three sub-networks containing GRU network, which are used to process contextual information in language, visual, and audio modalities. (2) Information Enhancement and Data Reorganization Module can be divided into information enhancement (IE) block, data reorganization (DR) block, and GRU network. (3) Sentiment Analysis Layer includes sentiment prediction and sentiment classification.

The MSFE input is N multimodal sequence data. In this article, the three modal data types of language, vision, and acoustics are involved, namely, N = {Language(l), Visual(v), audio(a)}. Let the sequence length of the multimodal data be T, and the feature vector of each modal at time step t can be expressed as $x_n^t \in \mathbb{R}^{d_{x_n}} (n \in \{l, v, a\})$. Among them, d_{x_n} is the feature dimension of modal n. The above three sequence modal information can be expressed as

$$X_n = \left[x_n^t : t \leqslant T, x_n^t \in \mathbb{R}^{d_{x_n}} \right] \tag{1}$$

3.1. Context-Unimodal Information Extraction Module

For the multimodal data obtained from the video, each modality X_n is sequence data containing a time stamp. In this module, we use three independent GRU networks to obtain internal-modal information over time for each modal sequence. Similar to the traditional Long Short-Term Memory Network (LSTM), the GRU network with reset gates and update gates is also specifically used to process sequence data with long-term dependencies according to the following GRU formula:

$$z^t = \sigma\left(W_z x^t + U_z h^{t-1} + b_z \right) \tag{2}$$

$$r^t = \sigma\left(W_r x^t + U_z h^{t-1} + b_r \right) \tag{3}$$

$$\tilde{h}^t = tanh\left(Wx^t + U\left(r^t \circ h^{t-1}\right) + b\right) \tag{4}$$

$$h^t = \left(1 - z^t\right) \circ h^{t-1} + z^t \circ \tilde{h}^t \tag{5}$$

In the above equation, $W_i \in \mathbb{R}^{h \times d}, U_i \in \mathbb{R}^{h \times h}$, and $b \in \mathbb{R}^h$ are the GRU training parameters, z^t, r^t are the update gate and reset gate, respectively. $\tilde{h}^t$ is a candidate activation, which accepts $\left[x^t, h^{t-1}\right]$ and $h^t \in \mathbb{R}^h$ is the GRU hidden state, where h represents the dimension of the GRU hidden state. $\circ$ is the Hadamard product, and σ is the sigmoid function.

We assign an independent GRU to each modal input sequence X_l, X_a, X_v, which helps to obtain the internal-modal feature representation $H_n = \left[h_n^t : t \leq T, h_n^t \in \mathbb{R}^h\right]$ with h denoting the dimensionality of the nth GRU hidden state.

$$H_l = L - GRU(X_l) \tag{6}$$

$$H_a = A - GRU(X_a) \tag{7}$$

$$H_v = V - GRU(X_v) \tag{8}$$

Finally, the acquired sequence information of the three modalities passes through the fully connected layer (D) of dimension d at each time step t. The purpose is to facilitate the consistency of the feature dimension direction when we reorganize the sequence data. The final output of the module is $U_l, U_a, U_v \in \mathbb{R}^{T \times d}$.

$$U_l = D_l(H_l) \tag{9}$$

$$U_a = D_a(H_a) \tag{10}$$

$$U_v = D_v(H_v) \tag{11}$$

3.2. Information Enhancement and Data Reorganization Module

In our proposed model, our goal is to use the contextual information within the modalities and the common contextual information between the modalities for multimodal sentiment analysis. As is known, the information that a piece of video conveys is multimodal. These multimodal signals are interwoven, and there is a certain delay and lag between the multimodal signals. Moreover, the importance of multimodal data is not the same, and it should not be potentially fused with the same weight. Especially as the sequence goes on, the weight of multimodal data at each time step is not static. Therefore, this paper introduces information enhancement and reorganization modules.

3.2.1. Information Enhancement (IE) Block

First, we perform information enhancement on the internal-modal feature representation $U_n (n \in \{l, a, v\})$ obtained above. We obtain an internal-modal matching matrix $E_n \in \mathbb{R}^{T \times T}$ through the following formula:

$$E_n = U_n \cdot U_n^T \tag{12}$$

Then, we calculate the correlation coefficient matrix $M_n \in \mathbb{R}^{T \times T}$ to the internal-modal matching matrix E_n using the softmax function. The element $M_n(i, j)$ in the matrix represents the degree of contextual information correlation between time i and time j in the model. Finally, the soft attention mechanism performs information enhancement at time t in each modality and obtains the information-enhanced sequence data $S_l, S_a, S_v \in \mathbb{R}^{T \times d}$.

$$M_n(i, j) = \frac{e^{E_n(i,j)}}{\sum_{t=1}^{T} e^{E_n(i,j)}} \qquad for \;\; i, j = 1, \ldots, T \tag{13}$$

$$S_l = M_l \cdot U_l \tag{14}$$

$$S_a = M_a \cdot U_a \tag{15}$$

$$S_v = M_v \cdot U_v \tag{16}$$

3.2.2. Data Reorganization (DR) Block

Previously, when further processing the data of each time step, the feature-level cascade operation was used. When this method performs dimensional connection, multimodal data have the same weight by default, especially on a single time step. However, when people express emotions, the intensity of the modal signal is not the same, and it will change over time. Perhaps at time $t-1$, words with strong emotional meaning dominate, while at time t, the rich facial behavior suppresses other modal signals. Feature-level connections cannot solve this problem. Therefore, this paper uses the method of reorganizing multimodal sequences and combines the information enhancement module to assign different weights to the modal information of each time step. Considering that the language information at time t may have a certain connection with the acoustic (visual) at time $t-1$ and the acoustic (visual) at time $t+1$ or have multimodal contextual information, at each time step t, we arrange the sequence data after information enhancement in order of the language-visual-acoustic (Figure 2) to obtain new multimodal sequence data $F = \left[S_l^t; S_v^t; S_a^t : t \leq T, S_l^t; S_v^t; S_a^t \in \mathbb{R}^d\right]$. Then, the new sequence data after the arrangement pass through the **IE** block again to calculate the cross-modal matching matrix $E_f \in \mathbb{R}^{3T \times 3T}$ and the correlation coefficient matrix $M_f \in \mathbb{R}^{3T \times 3T}$. Further, strengthen the association between multimodal data, and obtain enhanced multimodal recombination sequences $S_f \in \mathbb{R}^{3T \times d}$.

$$E_f = F \cdot F^T \tag{17}$$

$$M_f(i,j) = \frac{e^{E_f(i,j)}}{\sum_{t=1}^{3T} e^{E_f(i,j)}} \qquad for \;\; i,j = 1,\ldots,3T \tag{18}$$

$$S_f = M_f \cdot F \tag{19}$$

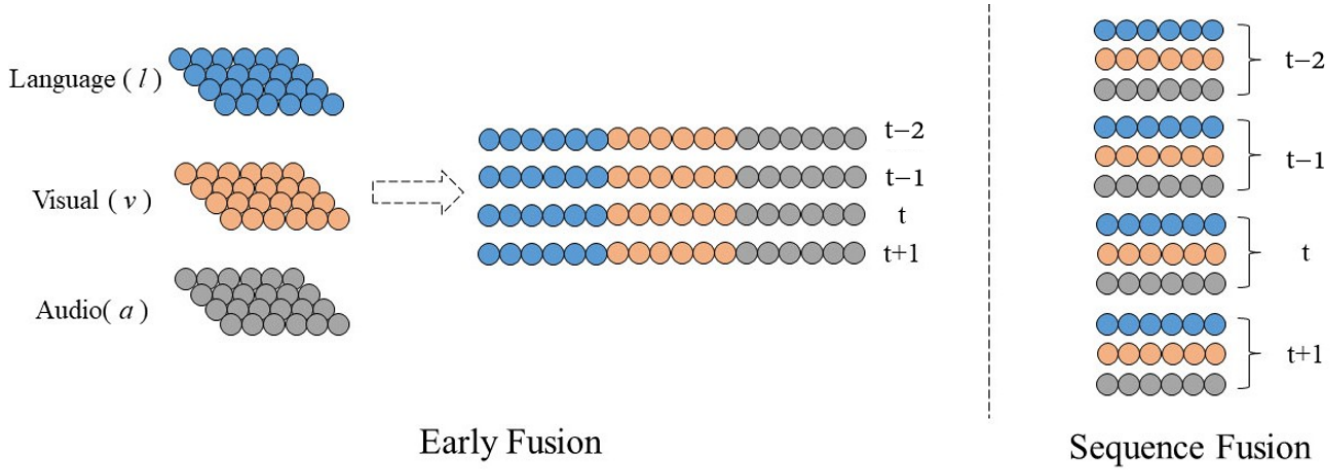

Figure 2. The left picture shows the traditional feature cascading method, and the right picture shows the new multimodal time series method.

This method prioritizes the enhancement of information within the unimodal before reorganizing the data and strengthening the degree of common association among multimodal data. The reason is that the degree of relevance of data within the unimodal should be greater than the degree of relevance of data between multimodal. After passing through the DR and IE blocks, the obtained new enhanced sequence data S_f is passed through a fusion GRU network to obtain a multimodal fusion embedding representation. Unlike the

Section 3.1 module, we obtain the contextual feature representation $h_f^T \in \mathbb{R}^{d_f}$ at the final moment T to represent the final representation output.

$$h_f^T = Fusion - GRU\left(S_f\right) \tag{20}$$

3.3. Sentiment Analysis Layer

After the information enhancement module and the information reorganization module, we finally obtain the common fusion feature representation among the multimodalities, which fully considers the time delay and lag between the multimodalities in the time dimension. At the task classification layer, we finally cascade the fusion feature representation h_f^T and the feature representation h_l^T, h_a^T, h_v^T within each modality at the final moment T.

$$Z^T = h_l^T \oplus h_a^T \oplus h_v^T \oplus h_f^T \tag{21}$$

After obtaining the complete fusion feature representation of language, visual, and acoustic modes, we performed sentiment prediction and emotion classification tasks. We use the ReLU activation function and the sigmoid layer for sentiment prediction, which is expressed as follows:

$$\hat{y} = \sigma\left(W_q \cdot \left(relu\left(W_p \cdot Z^T + b_p\right)\right) + b_q\right) \tag{22}$$

where $\hat{y}$ is the result of sentiment prediction; W_p and b_p re the weight and bias of the ReLU layer, respectively; and W_q and b_q are the weight and bias of the sigmoid layer, respectively.

Similarly, we use the ReLU activation function and softmax layer for multilabel emotion classification for each category, which is expressed as follows:

$$\hat{y}^n = softmax\left(W_q \cdot \left(relu\left(W_p \cdot Z^T + b_p\right)\right) + b_q\right) \tag{23}$$

where $\hat{y}^n \in \mathbb{R}^2$represents the probability prediction value of the nth category in the sentiment classification, W_p and b_p are the weight and bias of the ReLU layer, respectively, and W_q and b_q are the weight and bias of the sigmoid layer, respectively. We show the proposed method in Algorithm 1 .

Algorithm 1 Multimodal Sequence Feature Extraction Network (MSFE)

Require: Multimodal sequence information, including language (X_l), visual(X_v), audio(X_a)
Ensure: Sentiment prediction; Emotion classification

procedure MSFE(l, v, a)
  for each $n \in [l, v, a]$, $N \in [L, V, A]$ **do**
    $H_n \leftarrow N - GRU(X_n)$ ▷$h_n^T \in H_n$
    $U_n \leftarrow D_n(H_n)$
    $S_n \leftarrow IE - Block(U_n)$
  end for
  $F \leftarrow DR - Block(S_l, S_v, S_a)$
  $S_f \leftarrow IE - Block\left(U_f\right)$
  $h_f^T \leftarrow Fusion - GRU\left(S_f\right)$
  $Z^T \leftarrow h_l^T \oplus h_v^T \oplus h_a^T \oplus h_f^T$
  $prediction \leftarrow Sentiment\left(Z^T\right)$
  $classification \leftarrow Emotion\left(Z^T\right)$
  return Sentiment; Emotion
end procedure

Algorithm 1 *Cont.*

procedure $IE - Block(X)$
  /*internal-modal matching matrix*/
  $E \leftarrow X \cdot X^T$
  /*correlation coefficient matrix*/
  for $i, n \in 1, \cdots, U$ **do** $U \in (T, 3T)$
    $M(i,j) \leftarrow \frac{e^{E(i,j)}}{\sum_{t=1}^{U} e^{E(i,j)}}$
    $Y \leftarrow M \cdot X$
  end for
  return Y
end procedure
procedure $DR - Block(X)$
  $F \leftarrow [\,]$
  for $t \in 1, \cdots, T$ **do**
    $F^t \leftarrow \left[S_l^t; S_v^t; S_a^t\right]$
    $F \leftarrow F \oplus F^t$
  end for
  return F
end procedure

4. Experiments

In this part, to verify our proposed method, we describe the datasets used in the experiment and report the results and the necessary analysis.

4.1. Datasets

We evaluate the proposed model on the benchmark dataset of sentiment and sentiment analysis, namely, the CMU Multimodal View Sentiment and Mood Intensity (CMU-MOSEI) dataset [32]. The CMU-MOSEI dataset contains 3228 videos from more than 1000 online YouTube speakers with a total of 22,413 utterances. We divide the data into training, validation, and test sets to contain 15,290, 2291, and 4832 utterances, respectively. Each utterance is continuous for the sentiment label $[-3, +3]$ from highly negative (-3) to highly positive $(+3)$, and value < 0 and value ≥ 0 represent negative and positive emotions, respectively. The output layer of our model uses a sigmoid activation function for target prediction. In contrast, the emotion label contains 6 categories: anger, disgust, fear, happiness, sadness, and surprise. We consider the sample with an emotional intensity value of 0 as a negative example that does not contain emotion, and an emotional intensity value greater than 0 is a positive example of emotion in the sentence. We treat each category as a binary classification label and use 6 softmax functions for multilabel sentiment classification tasks. Table 1 shows the detailed statistics of the CMU-MOSEI dataset.

The dataset includes three modal data of language, visual and audio. Among them, The language feature is 300-dimensional GloVe word vectors [33]. The visual feature first cuts the video at a frequency of 30 frames and then extracts the face embedding by the commonly used facial recognition module [34–36]. Audio features are extracted using COVAREP software [37], and they are all related to mood and speech intonation. The word-level alignment method is used to align all content with the GloVe vector modality, that is, the time interval between the visual information and acoustic information of multiple frames corresponding to the corresponding words. Therefore, the sequence length of all modal data is 20.

Table 1. Dataset statistics for CMU-MOSEI. Each utterance contains multimodal information.

Statistics	Train	Validation	Test
videos		3228	
speakers		1000	
Utterance	15,290	2291	4832
Positive	10,887	1673	3360
Negative	4403	618	1472
Anger	3433	427	971
Disgust	2720	352	922
Fear	1319	186	332
Happy	8147	1313	2522
Sad	3906	576	1334
Surprise	1562	201	479

4.2. Baselines

In order to verify the performance of MSFE, we compared it with the following benchmark model for multimodal language analysis.

EF-LSTM. Early Fusion LSTM [25] concatenates the inputs from different modalities at each time-step and uses that as the input to a single LSTM.

TFN. The Tensor Fusion Network [16] creates a multidimensional tensor to capture unimodal, bimodal, and trimodal interactions, and explicitly model intra-modal and inter-modal dynamics.

G-MFN. The Graph Memory Fusion Network [32] models n-modal interaction through a set of parallel LSTM and dynamic fusion graph components. The dynamic fusion graph components can dynamically alter its structure and choose the proper fusion graph based on the importance of each n- modal dynamics during inference.

MTMM-ES. The Multi-task Multimodal Emotion Recognition and Sentiment Analysis [38] uses multimodal and contextual information to simultaneously predict the emotion and emotion of the utterance in a multi-task learning framework through a method of sharing parameters.

TBJE. The Transformer-based joint-encoding Model [29] uses the transformer network to transfer each unimodal information to the rest of the modalities, and mainly relies on the attention mechanism and the feedforward neural network (FFN) to draw the global dependency between input and output.

4.3. Experimental Settings

We conducted experiments on two tasks: sentiment binary classification prediction, and emotional multi-label classification. MSFE contains three independent unimodal coding sub-networks, a multimodal fusion coding network and a sentiment analysis layer. According to the different tasks of the experiment, the optimal parameters obtained are also different. For sentiment binary classification prediction, each sub-network contains a single-layer GRU and a fully connected layer. The number of three GRU neurons is 128, 64, 64, and the number of neurons in the fully connected layer is 128. The multimodal fusion coding network includes an information enhancement module, a data recombination module, and a multimodal single-layer fusion GRU with 64 neurons. For emotional multi-label classification, the number of neurons in the GRU and fully connected layer in each sub-network is 128, 32, 64, and 128, respectively. The number of neurons in the multimodal fusion GRU is 128. In the sentiment analysis layer, different tasks pass through the transformation layer of the same dimension of the fusion GRU and the independent output layer to obtain the final prediction results.

In the training phase, we use the Adam optimizer to train the model. The two subtasks have different learning rates of 1×10^{-4} and 1×10^{-3}, and each sub-network is assigned a different weight attenuation coefficient to prevent overfitting. On the issue of the number of iterations, we use the early stopping method for the training indicator Loss, and select the

epoch network with the smallest verification Loss for the evaluation process.In the process of model training, the loss functions of sentiment prediction and sentiment classification were cross-entropy loss functions with weight attenuation:

$$Loss(\hat{y}, y) = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{l=1}^{L} y_l^c \cdot \log \hat{y}_l^c + \beta^s \|w^s\| \quad (24)$$

where y is the true label, $\hat{y}$ is the probability predicted by the model, N is the total number of training samples, L represents the number of labels (1 for sentiment prediction and 6 for sentiment classification), and C is the number of categories. β^s (s is the number of GRU networks) is the decay coefficient of the weight value. Table 2 shows the main configuration of the model.

Table 2. Model configurations.

Parameters	Sentiment Params	Emotion Params
$T-GRU$	128 neurons	128 neurons
$V-GRU$	64 neurons	32 neurons
$A-GRU$	64 neurons	64 neurons
DenseLayer	128 neurons	128 neurons
$Fusion-GRU$	64 neurons	128 neurons
Optimizer	Adam ($lr = 1 \times 10^{-4}$)	Adam ($lr = 1 \times 10^{-3}$)
Output	Sigmoid	Softmax
Loss	*Binary crossentropy*	*Categorical crossentropy*
Batch	16	32
Activations	ReLu	
Epochs	Early Stopping (8)	

5. Results and Discussion

In this section, we have conducted a detailed analysis and discussion on the experimental results of CMU-MOSEI.

5.1. Results

We compared the experimental results with the benchmark model and the latest TBJE model. We compared with the benchmark model on the sentiment two-classification prediction task, and compared with the TBJE model with the same accuracy on the sentiment multi-label six-classification task. Table 3 shows the comparison of different models.

Table 3. Results on the test set. Please note that the F1 score and the binary classification accuracy of sentiment prediction are weighted to be consistent with the previous state-of-the-art technology. In the six categories of emotion labels, the F1 score is consistent with the previous technology, and the accuracy rate is the same as the latest technology TBJE, which is the standard accuracy rate. The Macro-f1 score is the average of the F1-score of six categories. * Values are taken from in [32] . $^+$ Values are taken from in [38].

Test Set	Sentiment		Emotion												
	2-Class		Happy		Sad		Angry		Fear		Disgust		Surprise		Avg
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Macro-F1
EF-LSTM *	-	-	57.8	-	59.2	-	-	-	56.7	-	-	-	-	-	-
TFN *	-	-	66.5	66.6	58.9	-	60.5	-	-	-	-	-	52.2	-	-
G-MFN *	76.9	77.0	66.3	66.3	60.4	66.9	62.6	72.8	62.0	89.9	69.1	76.6	53.7	85.5	-
MTMM-ES (S) $^+$	79.8	77.6	61.6	59.3	65.4	72.4	64.5	75.6	51.5	87.7	72.2	81.0	53.0	86.5	-
MTMM-ES (M) $^+$	80.5	78.8	53.6	67.0	61.4	72.4	66.8	75.9	62.2	87.9	72.7	81.9	60.6	86.0	-
TBJE (LA)	**82.4**		66.0	65.5	**73.9**	67.9	**81.9**	**76.0**	89.2	87.2	**86.5**	**84.5**	90.6	86.1	77.5
TBJE (LAV)	81.5		65.0	64.0	72.0	**67.9**	81.6	74.7	89.1	84.0	85.9	83.6	90.5	**86.1**	76.7
MSFE (LAV)	79.7	**81.8**	**67.7**	**67.8**	**72.1**	67.0	81.1	74.4	**93.1**	**89.8**	80.7	78.9	**90.7**	85.4	**77.2**

For different tasks, our model has achieved good results in some performance indicators. On the sentiment two-category prediction task, the weighted F1 score reached 81.8%, which is 3% higher than the multi-task learning model (MTMM-ES) overall, which proves that our model has a good learning ability. On the emotional multi-label six-classification task, we first compared the experimental results with the three-modality of the TBJE model. It can be seen that the accuracy and F1 score on the happiness label have improved significantly. Compared with the optimal model, the accuracy increased by 2.7%, and the F1 score increased by 3.8%. Moreover, the model is relatively sensitive to the fear label among the six labels, with an accuracy rate of 93.1%, an increase of 4%, and the weighted F1 score also increased by 5.8% to 89.8%.

However, the performance index of the TBJE model on the task of bimodal sentiment analysis is generally better than the experimental results of its three-modality version. This is because TBJE mainly uses the transformer network to directionally encode information from one modal to another modal, focusing on the mutual information and long-term dependence between the two modalities. Make the model more prefer to deal with bimodal information. Our model focuses on the importance of the difference in the time dimension between the three-modal information, and at the same time reorganizes and encodes the three-modal data. The macro-f1 score in the table also proves this point very well. Moreover, we found that MSFE performs better than sentiment prediction on sentiment classification tasks. In order to further verify the sensitivity of the model to six categories of emotions, and to verify whether the model has better performance capabilities for three modalities. We also performed a bimodal (L-V, L-A, V-A) emotional multi-label six classification task. Table 4 shows the bimodal experimental results of MSFE on the six-labels emotion classification task.

Table 4. Display of results of bimodal emotion classification. L: Language, V: Visual, A: Acoustic.

Test Set	**Emotion**												
	Happy		**Sad**		**Angry**		**Fear**		**Disgust**		**Surprise**		**Avg**
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Macro-F1**
L-V-A	**67.7**	**67.8**	72.1	**67.0**	81.1	**74.4**	93.1	89.8	80.7	**78.9**	**90.7**	85.4	**77.2**
L-V	64.5	64.2	73.0	64.5	**83.1**	73.2	**93.4**	**89.9**	79.8	75.9	89.4	**87.9**	76.0
L-A	66.1	66.1	**73.6**	64.4	80.4	74.1	93.1	89.8	80.6	78.5	90.1	85.5	76.4
V-A	65.0	65.2	68.8	61.7	79.7	71.8	93.2	89.8	**81.5**	75.0	90.5	85.9	74.9

It can be seen in the results that our proposed model also achieves good results in the bimodal emotion 6 classification. In the dual mode, L-V achieves good results compared to L-A and V-A. It has an accuracy of 93.4% on the fear label and an accuracy of 83.1% on the anger label, which is higher than that of the three modalities (L-V-A). The accuracy of the experiment increased by 2%. We noticed that the fear and anger tags focused more on visual feature information than the added acoustic modal signal. However, judging from the overall results, our model has better performance capabilities for three-modal signals. It is because our model can effectively enhance and weaken the three-modal signal through the data enhancement module and the data reorganization module at the same time, so as to make better use of the three-modal information.In summary, this fully demonstrates the effectiveness of our proposed method in bimodal sentiment classification.

5.2. Qualitative Analysis

In order to show the more intuitive performance of MSFE in six emotion classification tasks, in Table 5 we selected multimodal samples from the MOSEI data set to display the results. We apply the model to the test samples to output the prediction results and fit the true values, respectively. In the CMU-MOSEI dataset, the sentiment label $[-3, +3]$ ranges from highly negative (-3) to highly positive $(+3)$, and the value < 0 and the value ≥ 0 represent negative and positive sentiment, respectively. Among them, the samples between the negative sentiment label $[-3, -1]$ and the positive sentiment label $[1, 3]$ are called

samples with strong sentiment attributes. This type of sample also has multiple emotional attributes in the emotional multi-label, and each attribute is biased towards the same sentiment (positive or negative). For example, Case 1 in Table 5 has three negative labels: Sad, Angry, and Disgust, showing strong negative sentiment. Case 2 showed strong positive sentiment. From the results, our model can obtain more accurate results for this type of sample. However, for some samples without emotional labels, some unmatched emotional labels will always be assigned to the sample. Case 3 shows the most undesirable result. For samples with weaker sentiment attributes, in addition to correctly predicting the corresponding emotional label, they will also be assigned other labels of the same sentiment (negative or positive). In addition, in the samples with the Disgust label, MSFE correctly predicted the label. The experimental results further confirmed the sensitivity of the MSFE model to emotional labels and its excellent learning ability in emotional six-label classification tasks.

Table 5. Example from the CMU-MOSEI dataset. The true emotional label lies between the six emotional labels. MSFE outputs the predicted label type. Yellow, Lime and Orange represent text information, visual information and audio information, respectively. Red indicates misclassified labels.

	Language + **Visual** + **Audio**		**Truth**	**MSFE**
1	But, I mean, if you're going to watch a movie like that, go see Saw again or something, because this movie is really not good at all.	+ frown + rapid	Sad Angry Disgust	Sad Angry Disgust Fear
2	It's one of the best action blockbuster I've seen this whole summer and I highly recommend you guys seeing it.	+ smile + excited	Happy Surprise	Happy Surprise
3	Bruce Willis is your old, your (umm) your (stutter) old typical cop but basically this time he's fighting internet (umm) terrorists.	+ calm + smooth	No class	Fear Disgust Angry Sad
4	(uhh) the story's just kind of a rehash of the previous movie and it overall just feels very forced	+ frown + slight	Disgust	Disgust Angry Happy Sad

6. Conclusions

In this paper, we propose a new method of processing multimodal data, which aims to fully consider the lag and time delay of multimodal data in the time dimension as well as the interembedding relationship within multimodal data. We evaluated our proposed method on the recently released benchmark dataset on multimodal sentiment and sentiment analysis (MOSEI). Experimental results show that compared with baseline data and the accuracy of the latest model, our method effectively enhances the correlation of relevant information between multimodal signals. In the future, we will introduce a multi-task joint learning framework into our model to predict sentiment analysis and emotional classification at the same time from end to end. At the same time, in order to better reflect the advantages of multi-modality, we will further consider unimodal sentiment analysis research.

Author Contributions: Conceptualization, Q.Q. and R.Z.; Data curation, Q.Q.; Formal analysis, Q.Q.; Methodology, Q.Q.; Project administration, R.Z.; Resources, R.Z.; Software, Q.Q.; Supervision, L.L.;

Validation, Q.Q. and R.Z.; Visualization, Q.Q.; Writing—original draft, Q.Q.; Writing—review & editing, Q.Q., L.L., and R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tianjin Sino-German University of Applied Sciences Technology project grant number ZDKT2018-006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Previously reported CMU Multimodal View Sentiment and Mood Intensity (CMU-MOSEI) data were used to support this study and are available at https://github.com/A2Zadeh/CMU-MultimodalSDK (accessed on 19 August 2021). These prior studies (and datasets) are cited at relevant places within the text as references.

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

References

1. Korobiichuk, I.; Syerov, Y.; Fedushko, S. The Method of Semantic Structuring of Virtual Community Content. In *Mechatronics 2019: Recent Advances Towards Industry 4.0. MECHATRONICS 2019*; Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2020; Volume 1044; pp. 11–18.
2. Cambria, E.; Hazarika, D.; Poria, S.; Hussain, A.; Subramanyam, R.B.V. Benchmarking multimodal sentiment analysis. *Computational Linguistics and Intelligent Text Processing*; Springer: Cham, Switzerland, 2018; Volume 10762, pp. 166–179.
3. Reiter, E.; Dale, R. *Building Natural Language Generation Systems*, 1st ed.; Cambridge University Press: Cambridge, MA, USA, 2000.
4. De Mulder, W.; Bethard, S.; Moens, M.-F. A survey on the application of recurrent neural networks to statistical language modeling. *Comput. Speech Lang.* **2015**, *30*, 61–98. [CrossRef]
5. Harabagiu, A.M.; Paca, M.A.; Maiorano, S.J. Experiments with Open-Domain Textual Question Answering. *Coling* **2000**, *2*, 292–298.
6. Strzalkowski, T.; Harabagiu, S. Advances in Open Domain Question Answering. *Comput. Linguist.* **2007**, *33*, 597–599.
7. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2015**, arXiv:1409.0473v7
8. Dodge, J.; Gane, A.; Zhang, X.; Bordes, A.; Chopra, S.; Miller, A.H.; Szlam, A.; Weston, J. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv* **2016**, arXiv:1511.06931
9. Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; Jurafsky, D. Deep Reinforcement Learning for Dialogue Generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1192–1202.
10. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
11. Jiang, H.; Learned-Miller, E. Face Detection with the Faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 650–657.
12. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service, PlatCon 2017, Busan, Korea, 13–15 February 2017.
13. Thongtan, T.; Phienthrakul, T. Sentiment Classification Using Document Embeddings Trained with Cosine Similarity. In Proceedings of the ACL 2019—57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop, Florence, Italy, 28 July–2 August 2019; pp. 407–414.
14. Pham, H.; Manzini, T.; Liang, P.P.; Poczos, B. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. *arXiv* **2018**, arXiv:1807.03915.
15. Akhtar, M.S.; Chauhan, D.S.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In Proceedings of the NAACL HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 370–379.
16. Huang, Y.; Wang, W.; Wang, L.; Tan, T. Multi-task deep neural network for multi-label learning. In Proceedings of the 2013 IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, 15–18 September 2013; pp. 2897–2900.
17. Yoon, S.; Byun, S.; Jung, K. Multimodal Speech Emotion Recognition Using Audio and Text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, 18–21 December 2018; pp. 112–118.
18. Glodek, M.; Tschechne, S.; Layher, G.; Schels, M.; Brosch, T.; Scherer, S.; Kachele, M.; Schmidt, M.; Neumann, H.; Palm, G.; et al. Multiple Classifier Systems for the Classification of Audio-Visual Emotional States. In *Computational Linguistics and Intelligent Text Processing*; Springer: Cham, Switzerland, 2011; Volume 6975, pp. 359–368.
19. Ghosh, S.; Laksana, E.; Morency, L.-P.; Scherer, S. Representation Learning for Speech Emotion Recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 3603–3607.

20. Hu, A.; Flaxman, S. Multimodal Sentiment Analysis To Explore the Structure of Emotions. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018; pp. 350–358.
21. Wang, H.; Meghawat, A.; Morency, L.P.; Xing, E.P. Select-additive learning: Improving generalization in multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 949–954.
22. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.
23. Williams, J.; Kleinegesse, S.; Comanescu, R.; Radu, O. Recognizing emotions in video using multimodal dnn feature fusion. In Proceedings of the Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Melbourne, Australia, 20 July 2018; pp. 11–19.
24. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [CrossRef]
25. Poria, S.; Mazumder, N.; Cambria, E.; Hazarika, D.; Morency, L.-P.; Zadeh, A. Context-Dependent Sentiment Analysis in User-Generated Videos. In Proceedings of the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics ACL 2017, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883.
26. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.-P. Multimodal Language Analysis with Recurrent Multistage Fusion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2020; pp. 150–161.
27. Zadeh, A.; Poria, S.; Liang, P.P.; Cambria, E.; Mazumder, N.; Morency, L.-P. Memory Fusion Network for Multi-view Sequential Learning. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LS, USA, 2–8 February 2018; pp. 5634–5641.
28. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence AAAI, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7216–7223.
29. Delbrouck, J.-B.; Tits, N.; Brousmiche, M.; Dupont, S. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In Proceedings of the Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), Seattle, WA, USA, 10 July 2020; pp. 1–7.
30. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
31. 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.
32. Zadeh, A.; Liang, P.P.; Vanbriesen, J.; Poria, S.; Tong, E.; Cambria, E.; Chen, M.; Morency, L.-P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the ACL 2018, Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
33. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
34. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
35. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
36. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6738–6746.
37. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the ICASSP, Florence, Italy, 4–9 May 2014; pp. 960–964.
38. Akhtar, M.S.; Chauhan, D.S.; Ghosal, D.; Poria, S.; Ekbal, A.; Bhattacharyya, P. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv* **2019**, arXiv:1905.05812.