electronics

MDPI

Article

MMATERIC: Multi-Task Learning and Multi-Fusion for AudioText Emotion Recognition in Conversation

Xingwei Liang [1,2], You Zou [1], Xinnan Zhuang [1], Jie Yang [3], Taiyu Niu [2] and Ruifeng Xu [2,*]

[1] Konka Group Co., Ltd., Shenzhen 518053, China
[2] Joint Lab of HIT-Konka, Harbin Institute of Technology, Shenzhen 518055, China
[3] School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA
* Correspondence: xuruifeng@hit.edu.cn

Abstract: The accurate recognition of emotions in conversations helps understand the speaker's intentions and facilitates various analyses in artificial intelligence, especially in human–computer interaction systems. However, most previous methods need more ability to track the different emotional states of each speaker in a dialogue. To alleviate this dilemma, we propose a new approach, Multi-Task Learning and Multi-Fusion AudioText Emotion Recognition in Conversation (MMATERIC) for emotion recognition in conversation. MMATERIC can refer to and combine the benefits of two distinct tasks: emotion recognition in text and emotion recognition in speech, and production of fused multimodal features to recognize the emotions of different speakers in dialogue. At the core of MATTERIC are three modules: an encoder with multimodal attention, a speaker emotion detection unit (SED-Unit), and a decoder with speaker emotion detection Bi-LSTM (SED-Bi-LSTM). Together, these three modules model the changing emotions of a speaker at a given moment in a conversation. Meanwhile, we adopt multiple fusion strategies in different stages, mainly using model fusion and decision stage fusion to improve the model's accuracy. Simultaneously, our multimodal framework allows features to interact across modalities and allows potential adaptation flows from one modality to another. Our experimental results on two benchmark datasets show that our proposed method is effective and outperforms the state-of-the-art baseline methods. The performance improvement of our method is mainly attributed to the combination of three core modules of MATTERIC and the different fusion methods we adopt in each stage.

Keywords: emotion recognition in conversation; Multi-Task Learning and Multi-Fusion for AudioText Emotion Recognition in Conversation; multi-task learning; multimodal fusion

check for updates

Citation: Liang, X.; Zou, Y.; Zhuang, X.; Yang, J.; Niu, T.; Xu, R. MMATERIC: Multi-Task Learning and Multi-Fusion for AudioText Emotion Recognition in Conversation. *Electronics* **2023**, *12*, 1534. https://doi.org/10.3390/electronics12071534

Academic Editors: Chengwei Huang, Yongqiang Bao and Li Zhao

Received: 4 February 2023
Revised: 18 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, natural language processing technology has experienced rapid development [1,2]. Human–computer interactive systems have gained widespread attention and become a research hotspot in both academia and industry. As these systems continue to progress, humans have increasingly high expectations for them, hoping that machines can communicate more deeply with humans based on their attention to the content of the responses. Therefore, enabling machines to understand emotions in conversation has become a growing challenge for human–computer interactive systems. Emotion recognition in conversation (ERC) recognizes a person's emotional state by analyzing speech, visual, and physiological signals, and it utilizes the complementarity between multi-channel emotional information to improve the accuracy of emotion recognition [3].

With the development of the Internet, a massive amount of data, such as text, images, and videos, are becoming easier to obtain. Emotion recognition for such data is an essential task in artificial intelligence, and the research results can be used for product recommendation and public opinion monitoring. So far, speech emotion recognition has primarily been

limited to language itself, that is, it only pays attention to the linguistic system and semantic structure itself and its relationship with social cultural, and psychological cognition while ignoring other forms of meaningful expressions, such as images, sounds, colors, and animation. Multimodal discourse emotion recognition can significantly help people overcome these limitations. To this end, we build our framework to enable the integration of features from multiple modalities, thereby improving the accuracy of emotion recognition and lead to better applications.

Multi-task learning (MTL) is an inductive transfer mechanism of machine learning, where multiple tasks are learned in parallel, and the results influence each other [4,5]. In recent years, it has been widely used in computer vision [6], natural language processing [7,8], recommendation system [9] and other fields [10,11]. In this work, we propose a MTL framework, named MATTERIC, for conversational multimodal emotion recognition. The proposed framework leverages the interdependence of the two tasks (speech and text) to obtain better feature representation and improve the performance of each task. Moreover, a multi-task framework reduces complexity, as a single system can handle multiple tasks simultaneously. MMATERIC comprises two main components: feature networks (FeatureNets) and central knowledge processor (CKP). FeatureNets extract task-specific features from the input modality, and CKP combines these features to predict the emotional state. We employ MMATERIC for three tasks: ERC in text, ERC in speech, and ERC after the fusion of text and speech signals. The third task combines high-level textual information and low-level speech signals to learn emotional features from low-level speech signals.

FeatureNets is the crucial component of the emotion recognition task, which involves extracting features from speech using the open-source tool openSMILE [12], and from text using models such as Roberta [13]. However, standard long short-term memory (LSTM) [14] units are insufficient for remembering long-term changes in a particular speaker. To address this limitation, we propose a CKP component consisting of three core modules: an encoder with multimodal attention, a speaker emotion detection Unit (SED-Unit), and a decoder with speaker emotion detection Bi-LSTM (SED-Bi-LSTM). Together, those three modules can model the changing emotional state of a speaker in the conversation at a given time. To learn intra-modal representation, we incorporate pairwise intra-modal attention into the encoder with an attention module. We then use the SED-Unit to process the output of the encoder, which is subsequently fed into the SED-Bi-LSTM module. In the single-task process, the output is sent directly to the classification layer, while in the multi-task process, it is passed through the fusion block before being sent to the classification layer. To prevent overfitting, we use dropout, and a fully connected layer is used as our discriminator. In our framework, feature representations are first fed into pooling layers before being fed into the final fully connected layer (FC). Pooling layers help to reduce the size of the parameter matrix, which in turn reduces the number of parameters and speeds up computation while preventing overfitting. The fully connected layers act as "classifiers" in the multi-task learning framework. While convolution, pooling, and activation operations map the raw data to the hidden feature space, fully connected layers map the learned "distributed feature representation" to the sample label space.

To demonstrate the effectiveness of our method, we conduct comprehensive experiments on two multimodal datasets for ERC analysis. The results show considerable improvements compared to well-known baseline methods.

The main contributions of our work are summarized as follows:

- We propose a central knowledge processor (CKP) that models speakers' changing emotional states during a conversation. The CKP comprises three core modules: an encoder with multimodal attention, a speaker emotion detection unit (SED-Unit), and a decoder with speaker emotion detection Bi-LSTM (SED-Bi-LSTM). The CKP compensates for the inability of LSTM, which is commonly used in ERC tasks, to remember long-term variations of specific speakers.
- By modifying the classic LSTM structure, we add the speaker state at the input to gain a deeper understanding of the speaker's emotions

- We employ a multi-task framework to process three interdependent tasks and use different fusion strategies at different processing stages to improve AudioText emotion recognition accuracy in conversation.
- A series of experiments on two benchmark datasets show that the proposed method achieves a considerable improvement over baseline models.

In the remaining paper, we describe related works on multi-task, multi-modal, multi-fusion, and conversational emotion recognition in Section 2. Next, we provide the details on the framework design in Section 3. Datasets and the experimental results are mentioned in Section 4, as well as the experimental analysis. In Section 5, we conduct ablation studies to measure the impact of removing different components on the system's performance. We conclude with a summary and mention our future work in Section 6.

2. Related Works

2.1. Multi-Task Learning

Multi-task learning is an approach that takes inspiration from human behavior of handling multiple tasks simultaneously. It involves using the domain information contained in the training signals of related tasks as an inductive bias to improve generalization [4]. This approach allows the model to extract features from associated tasks simultaneously, and different tasks can share relevant knowledge during the learning process. As a result, the correlation between tasks improves the model's performance on each task.

Recently, multi-task learning has been applied in the field of emotion recognition to learn the interdependence between different tasks and enhance the accuracy of single-task emotion recognition [15]. Researchers have developed multi-task frameworks that provide subspaces to learn modality-private and modality-shared features for emotion recognition tasks [16], as well as transformer-based models to learn multiple tasks across different domains, including sentiment analysis [17]. However, few studies have focused on using multi-task learning for ERC work. In this paper, we use the multi-task framework to resolve multimodal ERC problems by dividing the multi-task work into three sub-tasks: the audio task, the text task, and the combined audio and text task.

2.2. Multi-Modal

Humans operate in a multimodal environment, where they process information through various senses, such as speech, facial expressions, movement, and language. Each of these senses can be considered a separate modality, and they work together to provide a rich and immersive experience. Therefore, multimodal data are highly correlated and complementary.

The expression of human emotions is also a multimodal process, as emotions can be recognized from multiple modalities. To this end, researchers have proposed various multimodal models for emotion recognition, such as Chen et al.'s bimodal speech emotion recognition model that uses speech and text information [18]. Additionally, some researchers, such as Soleymani et al., argue that multimodal learning can simulate the way humans perceive the world, leading to more effective artificial intelligence [19]. Our study applies MTL not only to the single task of speech and text but also to the task of fusion of the two modalities.

2.3. Multimodal Features and Fusions

Multimodal data involve representing objects from various perspectives, such as text, audio, and image. When conducting a multimodal emotion recognition task, it is critical to determine how to integrate feature representations from various modalities to capture complementary information. The fusion of different modalities has been one of the main research areas in sentiment analysis. Gandhi et al. [20] discussed ten different fusion procedures, including early fusion, late fusion, hybrid fusion, model level fusion, tensor fusion, hierarchical fusion, bi-modal fusion, attention-based fusion, quantum-based fusion,

and word-level fusion. In contrast, we categorize fusion strategies into three types: early feature-based fusion, mid-stage model-based fusion, and late decision-based fusion.

Early feature fusion directly merges the underlying features of each modality, which can capture the correlation and interaction between them. For example, Qian et al. [21] combined speech and text features for fine-grained temporal alignment and cross-modal semantic interaction and used the same network to predict the final emotional state. However, early feature fusion approaches are prone to high feature dimensionality and out-of-sync features. Model-based fusion involves inputting various modal data into the network simultaneously and merging them based on the middle layers of the model. The advantage of model fusion is that the fusion position can be selected, and the interaction between modalities can also be realized. For instance, Gu et al. [22] proposed a multimodal framework for extracting features from text and audio, and they fused all features using DNN to learn cross-modal correlations. However, they overlooked word-level modal dynamic interactions when modeling sentence-level modal relations. Decision-level fusion works by weighting and combining the predictions from each modality. Decision-level fusion can also perform well when some modal data are missing, and the errors between different modalities do not affect each other. However, decision-level fusion requires more use of raw data information and capturing the correlation between shallow features. It also ignores the interaction between temporal modalities. For example, Rozgic et al. [23] built acoustic and text models to obtain their respective predictions, then obtained the final fusion prediction through ensemble learning.

Multimodal sentiment analysis has seen an increasing use of multi-level and multi-stage fusion strategies. For instance, Sun et al. [24] proposed a tensor fusion network followed by a decision-level fusion in a multi-level fusion setup. Liang et al. [25] employed the RNN framework to perform N stages of fusion. Atmaja et al. [26] combined early and late fusion techniques in a n-times fusion approach to enhance performance. In a similar vein, we incorporate multiple fusion strategies at different stages in our approach. Specifically, we leverage model fusion in the speaker emotion detection unit, and the optimal fusion model in the fusion block, to improve the accuracy of ERC.

2.4. Conversational Multimodal Emotion Recognition and Analysis

Conversational emotion recognition (ERC) is an active research area, and several ERC datasets have been published, such as IEMOCAP [27], AVEC [28], EmotionPush [29], DailyDialog [30], EmotionLines [31], and MELD [32]. The most common approach for ERC is context modeling in unimodal or multimodal settings using deep learning-based algorithms.

Hazarika et al. [33] proposed a conversational memory network that utilizes contextual information from the conversation history to recognize emotions in dyadic dialogue videos. However, their work did not consider the order of conversation. Poria et al. [34] proposed an LSTM-based approach that considers interdependencies between utterances for multimodal emotion recognition, but it may suffer from long-term context propagation problems [35]. DialogueRNN [36] employed an attention mechanism that pools information about all or part of the dialogue for each target utterance, but it does not consider the speaker's role in utterance formation and the relative position of other words to the target utterance. To address these issues, DialogueGCN [37] is a graph convolutional neural network that utilizes interlocutor's self and inter-speaker dependencies to model dialogue context for emotion recognition. It solves the context propagation problem in RNN-based methods through graph networks. Shen et al. [38] combined the advantages of graph-based methods and cycle-based models by treating each conversation as a directed acyclic graph (DAG), providing a way to combine the information of long-term dialogue and adjacent context intuitively.

Despite significant progress, many studies have not considered multiple fusions in a multi-task framework at different stages. Furthermore, we propose to modify the classic LSTM by incorporating speaker information to better understand the speaker's emotion in ERC.

3. Proposed Method

This section presents our proposed framework, Multi-Task Learning and Multi-Fusion for AudioText Emotion Recognition in Conversation (MMATERIC), summarized in Figure 1. MMATERIC takes audio and text as input and is jointly trained on multiple tasks, such as audio emotion recognition, text emotion recognition, and fusion of the two.

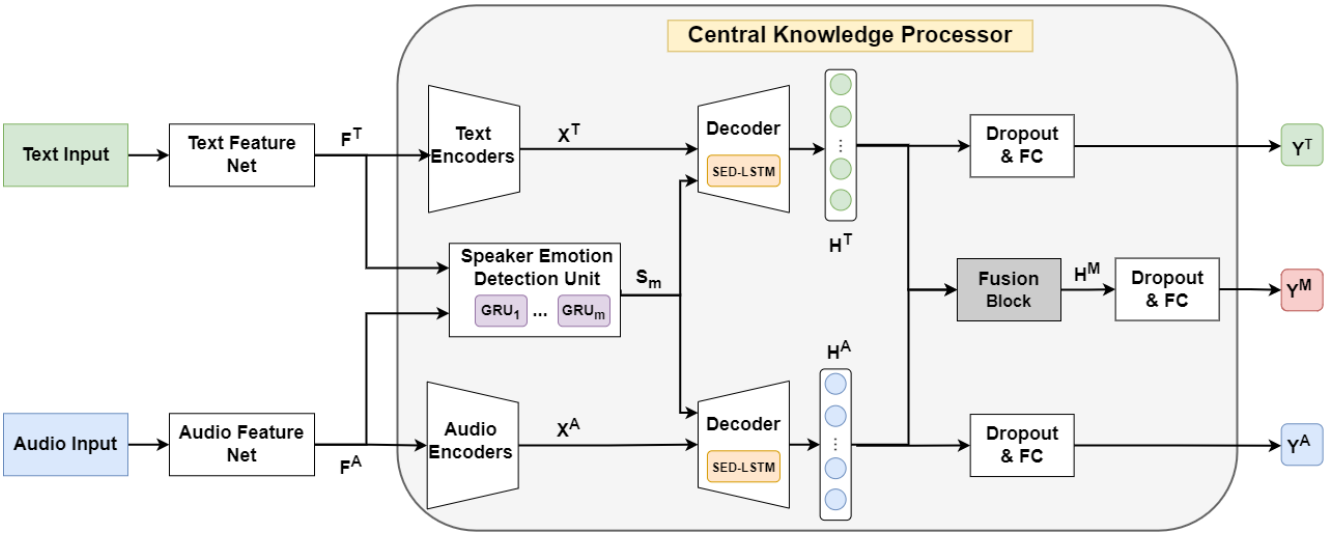


Figure 1. The architecture of the proposed MMATERIC framework. *A* represents audio modality, and *T* represents text modality.

3.1. Model Architecture

We use MMATERIC for multi-task learning across different modalities. It consists of individual feature networks (FeatureNets) for each input modality type, followed by a central knowledge processor (CKP).

CKP is a model that utilizes a transformer-based encoder to process both audio and text features. The resulting hidden states are combined using a fusion model and fed into a speaker emotion detection unit (SED-Unit). The fused hidden state and the previously encoded hidden state are then passed into a speaker emotion detection Bi-LSTM (SED-Bi-LSTM). In the next step, one or multiple tasks of the ERC (emotion recognition and classification) are performed depending on the task mode. For unimodal tasks, the decoded output from the SED-Bi-LSTM is fed into a task-specific head, such as a two-layer classifier that produces the final prediction. For multimodal tasks, the decoded output is sent to the fusion block, which combines the information from all modalities. The fused information is then passed through a two-layer classifier to make the final prediction.

Our CKP structure improves upon existing LSTM methods by addressing their limitations in remembering long-term variations of specific speakers. Our experimental results on two benchmark datasets show that CKP achieves higher accuracy than current well-known methods. MMATERIC, our model, is simple and can be easily extended to include additional modalities.

3.2. Feature Networks (FeatureNets)

FeatureNets are used to extract features from input data. In our work, we use both text and audio modalities. However, other modalities can be added depending on the specific task, such as video, smell, etc.

Audio Feature Network: Audio features are essential in providing information about a speaker's emotional state. We use the openSMILE open-source tool to extract audio features, which provides high-dimensional audio vectors. These vectors include loudness, Mel spectrum, MFCC, pitch, etc., totaling 6373 features. Subsequently, we perform dimensional reduction processing on the feature vector through a fully

connected layer, and the processed audio features are expressed as $F^T \in \mathbb{R}^{n*100}$.

Text feature network: The text input is extracted through Roberta. The model builds on BERT and modifies some key hyperparameters. It also removes the next sentence pretaining objective and trained with larger batches and learning rates. We tried feature embeddings using different pretained models, such as GloVe [39], and only Roberta pretained embedding led to significant performance gains. The textual features are represented as $F^T \in \mathbb{R}^{n*1024}$.

3.3. Central Knowledge Processor (CKP)

CKP is the core component of the MMATERIC framework, and it is composed of several modules that work together to enable multimodal emotion recognition. Specifically, the CKP module includes an encoder with multimodal attention, a speaker emotion detection unit (SED-Unit), a decoder with speaker emotion detection Bi-LSTM (SED-Bi-LSTM), and a feature fusion block. In this section, we present CKP in detail.

3.3.1. Encoder with Multimodal Attention

Based on the attention mechanism in neural machine translation, we developed a multimodal attention method to focus on the emotional aspect of the text and audio context. The self-attention mechanism in transformers is well suited to capture global temporal dependence in sequential data, and its parallel structure makes data processing more efficient. Therefore, we designed our encoder module using two transformer structures, as follows:

$$X^T = Tr_2(Tr_1(F^T)), \tag{1}$$

$$X^A = Tr_2(Tr_1(F^A)), \tag{2}$$

where Tr is the operation function of the transformer encoder, and X^T and X^A are the encoded feature representations for text and audio, respectively. In our transformer, we introduced a pair-wise intra-modal attention mechanism, which includes text-to-text intra-modal attention and audio-to-audio intra-modal attention, to learn the representation within each modality. The objective is to emphasize the contributing features by paying more attention to the respective and neighboring utterances.

3.3.2. Speaker Emotion Detection Unit (SED-Unit)

We use a gated recurrent unit (GRU) network to capture speaker information, with one network per speaker. To input data into the SED-Unit, we combine audio and text features using a fusion method we describe in Section 2.3. This approach involves extracting features from separate networks for audio and text, and then fusing them together.

In Section 3.3.4, we discuss our MultiFusion strategy in more detail. Specifically, we present the fusion block as our next stage fusion method. The GRU network encodes both the current utterance and the previous speaker states, generating the current speaker state as output. Suppose there are m speakers, for the current moment t:

$$S_{k,t} = GRU(concat(F_t^T, F_t^A),\ S_{k,t-1}), \tag{3}$$

where $k \in [1, m]$, $t \in [1, n]$, n is the number of utterances in a conversations. $S_{k,t-1}$ and $S_{k,t}$ are the previous and current speaker state.

3.3.3. Decoder with Speaker Emotion Detection Bi-LSTM (SED-Bi-LSTM)

We use bidirectional LSTM as the decoder. The conventional LSTM can capture long-term dependencies in the sequence and solve the gradient vanishing problem. However, it only considers the output related to the previous states at the current time step, not associating it with the future states. In comparison, Bi-LSTM handles the complexity of emotional changes in conversation in the previous and future states. The Bi-LSTM

structure of the decoder is schematically shown in Figure 2. The inputs to the decoder include the speaker emotion detection unit's output and the encoder modules' outputs. The decoder produces feature representations for classification in single tasks and multi-task fusion scenarios.

The output of the text modality and the audio modality from the encoder unit ($X^T = (x_1^T, x_2^T, ..x_t^T .., x_n^T)$, $X^A = (x_1^A, x_2^A, ..x_t^A .., x_n^A)$) are used as input x_t to the decoder respectively, and the LSTM unit operation is formulated as

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f s_t + b_f), \tag{4}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i s_t + b_i), \tag{5}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o s_t + b_o), \tag{6}$$

$$\hat{c}_t = \tanh(W_c x_{t-1} + U_c h_{t-1} + V_c s_t + b_c), \tag{7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t, \tag{8}$$

$$h_t = o_t \odot \tanh(c_t), \tag{9}$$

where f_t, i_t, o_t, and c_t are the forget gate, input gate, output gate, and cell state at time t. $W_{(.)}, U_{(.)}, V_{(.)}, b_{(.)}$ are learned parameters. σ is the sigmoid activation function. h_{t-1} and c_{t-1} are the output and cell state at time $t-1$. s_t is the state information of the current speaker at time t. $\odot$ denotes the element-wise product. The modal superscript is omitted for simplicity. The memory cell c_t is the core of the LSTM. It records the input history. The three gates are functions of the current time input x_t and previous output state h_{t-1}. The input gate i_t moderates the input information flow. The forget gate f_t controls if the LSTM forgets its previous memory cell c_{t-1}. The output gate manages the amount of information from c_t to the output state h_t. We modified the classic LSTM to add a speaker state to better comprehend the speaker's emotion in the ERC. In this way, the combined three gates and a memory cell, with enhanced LSTM capability, make SED-Bi-LSTM a better decoder model fit to our proposed framework.

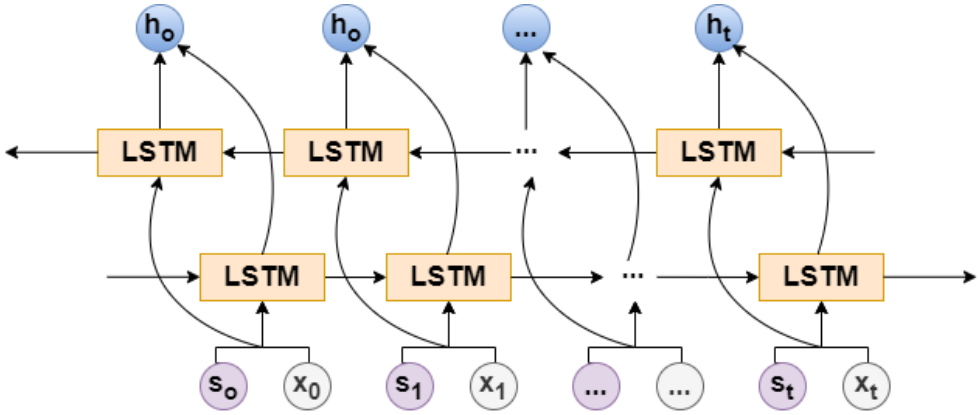


Figure 2. Network architecture of the speaker emotion detection Bi-LSTM (SED-Bi-LSTM), S_t and X_t are speaker state and encoded audio or text representation.

3.3.4. Feature Fusion Block

By understanding how text, audio, and other modalities interact in multimodal language, we can improve our ability to recognize the speaker's intentions. Multimodal fusion allows us to leverage the complementary information available in multimodal data, which can reveal how information depends on different modalities. To maximize this discovery, we can use a combination of fusion methods.

In our approach, we use model fusion in the SED-Unit and experiment with various fusion strategies in the fusion block to identify the most suitable one for our overall architecture. The fusion block comprises four fusion models: later fusion DNN (LF-DNN), tensor fusion network (TFN), low-rank multimodal fusion (LMF), and weight fusion (WF).

We compare the performance of each fusion strategy within the block to determine the optimal one.

LF-DNN: This model learns features separately for audio and text, and then concatenates them before the final classification.

TFN [40]: TFN first expands the dimensions for each modality, and then applies the outer product on the expanded features to create a multi-dimensional tensor. This approach captures both intra-modal and inter-modal dynamics. However, the high dimensionality of the resulting feature vector makes it challenging to learn and computationally expensive.

LMF [41]: LMF aims to improve the efficiency of tensor-based multimodal fusion methods by using modality-specific low-order factors. By decomposing tensors and weights in parallel, it performs multimodal fusion to learn both modality-specific and cross-modal interactions.

WF: Audio and text modalities may convey different emotional expressions, and a weighted fusion strategy is used in this model to capture these differences. By modeling the features from different modalities with appropriate weights, it aims to identify the speaker's emotion accurately:

$$H^M = \omega_1 * H^T + \omega_2 * H^A, \tag{10}$$

where ω_1 and ω_2 are trainable parameters, and $\omega_1 + \omega_2 = 1$.

3.4. Optimization Objective

We used cross-entropy loss as the model loss function. It is defined as follows:

$$\text{loss} = \min(\frac{1}{N}\sum_{n=1}^{N}\sum_{i} L(y_n^i, \hat{y}_n^i)) \tag{11}$$

where N is the total number of samples in a batch. $i \in \{\text{M, T, A}\}$ "M" represents a multimodal task, "T" and "A" represent a text task and an audio task, respectively. $L(y_n^i, \hat{y}_n^i)$ represents modal i's cross-entropy losses in the n-th sample.

3.5. Multi-Task Learning

The shared representation receives error gradients from all three tasks, enabling it to adjust the model's weights. As a result, the shared representations are not biased toward any particular task, which helps the model achieve generalization across multiple tasks. For instance, when training the model on task A, we aim to obtain a robust representation of task A while disregarding noise in the data and improving the generalization performance. Since different tasks have varying levels of noise, training two or more tasks simultaneously can lead to a more generalized representation.

4. Datasets, Experiments, and Analysis

To evaluate the performance of our proposed framework, we conduct experiments on the public emotion dataset—IEMOCAP and MELD. In this section, we first detail the datasets we use and their properties. Further, we present our experiment baselines. Then the experimental setup is described, followed by the experimental results and analysis of the proposed framework.

4.1. Datasets

IEMOCAP [27]: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is a standard English dataset. It consists of 136 rounds of conversations collected by a total of

10 participants. There are a total of 10 speakers grouped into five pairs. The first four pairs of conversations are used as the training set, and the remaining is used as the test set.
MELD [32]: The Multimodal Emotionlines dataset (MELD) has 13.7 h of conversations extracted from the Friends TV series. MELD contains multiparty conversation. There is a total of 1400 rounds of conversation and 13,000 utterances. To evaluate the performance of our proposed framework, we conduct experiments on the public emotion dataset—IEMOCAP and MELD. In this section, we first detail the datasets we use and their properties. Further, we present our experiment baselines. Then the experimental setup is described, followed by the experimental results and analysis of the proposed framework.

4.2. Baseline and Other Well-Known Methods

To provide a comprehensive evaluation of MMATERIC, we compared our model with several baseline methods:

ICON [42]: It is a multimodal sentiment detection framework based on interactive session memory network (ICON). By connecting the historical dialogue between two people with the same layer of GRU and using GRU to model the following multimodal input in each memory network, the emotional influence of the self-speaker and the inter-speaker is modeled hierarchically into the global memory.

DialogueRNN [36]: It is based on recursive neural networks that track the state of the parties throughout the session and uses this information for emotion classification.

DialogueGCN [37]: This method is based on a graph convolutional neural network. It resolves the problem of context transmission in the current RNN-based method by using interlocutor self-dependence and interspeaker dependency to model the context of dialogue in emotion recognition.

COIN [43]: COIN is a conversational interaction model that focuses on the interactions of different speakers at the speaker-utterance level by applying state interactions in historical contexts and introduces a stacked global interaction module that captures context and interdependent representations in a hierarchical manner. In addition, small perturbations are added to the input of multimodal features to generate adversarial examples to improve the robustness and generalization of the training process.

Iterative [44]: An iterative emotion interaction network is proposed. The model uses the emotion label of iterative prediction to replace the actual emotion label. In the iterative process, the prediction is constantly corrected. The input is feedback to realize gradually enhanced explicit emotion interaction, which effectively retains the performance advantage of explicit modeling and can realize effective prediction correction in the iterative process.

COSMIC [45]: COSMIC is a commonsense-guided framework for ERC. By building an extensive knowledge base of common sense, the proposed framework captures complex interactions between personalities, events, mental states, and intentions better to understand emotional dynamics and other aspects of the conversation.

EmoBERTa [46]: EmoBERTa is an abbreviation of "Speaker conscious emotion Recognition in Conversation with RoBERTa". By simply placing the speaker's name before the utterance and inserting a separator symbol between the utterance in the conversation, EmoBERTa can learn the state and context within and between the speaker to predict the current speaker's mood in an end-to-end manner.

DAG-ERC [38]: It encodes dialogue through the directed acyclic graph. It combines the traditional graphics-based neural network model with the recursion-based neural

model. This new encoding method can better combine the information of long-term dialogue and the information of adjacent context.

Our experiment report is shown in Table 1. It compares the performance of MMATERIC on the test data to other well-known methods.

Table 1. Comparison results on MELD and IEMOCAP datasets. Our model results are obtained for a single task combined with weight fusion. The results of the baselines are retrieved from the original papers. The scores in italics are recovered from our experiments using the original models.

Methods	IEMOCAP (6-Cls)		MELD (7-Cls)	
	Accuracy	Weighted Average F1	Accuracy	Weighted Average F1
ICON	64.0	63.5	-	-
DialogueRNN	63.4	62.75	59.54	57.03
DialogueGCN	65.25	64.18	59.46	58.1
DialogueCRN	66.05	66.20	60.73	58.39
COIN	66.05	65.37	-	-
Iterative	-	64.37	-	60.72
COSMIC	*64.14*	*64.02*	*66.21*	*64.52*
EmoBERTa	-	67.42	-	**65.61**
DAG-ERC	*67.82*	*67.89*	*63.98*	*63.63*
MMATERIC	**70.55**	**70.52**	**66.67**	65.27

4.3. Experiment Setup

We conducted our experiments using PyTorch to implement all the networks. We set the hidden layer dimension of the LSTM and GRU units to 256, and used a bidirectional structure for the LSTM and a unidirectional structure for the GRU. The loss function we used was cross entropy, and we initialized the learning rate to 1×10^{-3}, with a decay rate of 0.97. The batch size was set to 60. To evaluate the performance of our model, we used accuracy and the weighted-average F1 score as metrics.

4.4. Experiments

4.4.1. Comparison with Other Models

Table 1 presents the results of our MMATERIC framework in comparison to other methods on the IEMOCAP and MELD datasets. Our approach consistently outperforms the baseline models in terms of accuracy and F1 score, except for the weighted average F1 score on the MELD dataset for EmoBERTa, where EmoBERTa achieves a slightly better F1 score. It is worth noting that EmoBERTa's approach involves considering past, current, and future conversations to determine the emotion of the current conversation. This context-based approach improves their emotion recognition performance. However, in practical applications, the use of future conversations is often impossible, as they have yet to happen. In their experiments, when only considering past and current conversations, EmoBERTa achieves an F1 score of 64.55 on the MELD dataset, which is lower than the performance of our proposed MMATERIC model in the same settings.

Our approach achieves better performance than EmoBERTa and other well-known models mainly due to our CKP structure, which can remember long-term variations of specific speakers. Additionally, we incorporated speaker information, conducted multi-stage fusion, and utilized MTL to share relevant knowledge during the learning process. The correlation in MTL between tasks improves the model's performance, resulting in the observed performance gain over other models.

4.4.2. Comparison of Single-Task and Multi-Task

We compared the performance of our model, combining different fusion approaches in single-task and multi-task; results are shown in Table 2. Compared with single-task

models, multi-task models perform better in the IEMOCAP dataset, while the reverse is true in the MELD dataset. We speculate that this is due to the small number of participants in the IEMOCAP dataset, which allows our model to differentiate better. In contrast, the dialogues in the MELD dataset involve too many plot contexts, making emotion recognition more challenging.

4.4.3. Comparison of Multi-Fusion

Table 2 shows the different fusion strategies on both single-task and multi-task experiments. Different fusion strategies perform differently for single-task and multi-task learning. In particular, late fusion methods, such as LFDNN (late fusion DNN) and WF (weighted fusion), generally achieve enhanced accuracy and F1 score on both datasets in a single task. For multi task, TFN (tensor fusion network) and WF have better performance on both datasets.

Compared to other complex models, our model, shown in bold in Table 2, outperforms other baselines using weighted fusion. Late fusion methods, such as LFDNN (late fusion DNN) and WF (weighted fusion) generally achieve enhanced accuracy and F1 score on both datasets because they combine the predictions of multiple models or features, which allows them to capture more information and make more accurate predictions. TFN is a fusion technique that uses a tensor decomposition approach to combine different modalities of data. Like TFN, LMF is a fusion technique that uses low-rank matrix decomposition to combine different modalities of data. While the performance of those four different fusion models depends on our experiment settings, WF is the best from our experiments. Therefore, it is not necessarily to say that any other fusion techniques is universally not better than WF. Furthermore, our experiments found that the results obtained from the IEMOCAP dataset generally perform better than those of the MELD dataset. Compared with the MELD dataset, the IEMOCAP dataset is more standardized. The MELD dataset is from the TV series Friends. Many dialogues are related to the daily conversation context, so the standardization may take time to formulate.

Table 2. Comparison of single-task and multi-task results with different fusion strategies on MELD and IEMOCAP datasets. The models with ▷ are multi-task models. The symbol ⇑ means improvements or reductions in multi-task compared to single-task models in the current evaluation metric. green mean positive upper direction arrow, a gain or an improvements in performance for multi-task models compared to single-task models. red means negative direction arrow, a decrease in performance for multi-task models compared to single-task models

Methods	IEMOCAP (6-Cls)		MELD (7-Cls)	
	Accuracy	Weighted Average F1	Accuracy	Weighted Average F1
LMF	66.36	66.38	65.94	64.52
MLMF ▷	70.73	70.80	65.59	64.97
⇑	+4.37	+4.42	−0.35	+0.45
TFN	69.69	69.72	65.96	65.01
TFN ▷	71.16	70.91	65.82	65.67
⇑	+1.47	+1.19	−0.14	+0.66
LFDNN	69.69	70.06	66.57	65.12
LFDNN ▷	70.43	70.71	65.13	64.98
⇑	+0.74	+0.65	−1.44	−0.14
WF	**70.55**	**70.52**	**66.67**	**65.27**
WF ▷	70.60	70.71	65.79	65.17
⇑	+0.05	+0.19	−0.88	−0.10

4.4.4. Representations Differences

We use T-SNE [47], a dimensionality reduction visualization tool, to project highly dimensional features into a two-dimensional space. Specifically, our original data size is [batch size, dimension, length]. After using the T-SNE tool, it becomes two-dimension data, which are [batch size, process-dimension]. As a result, the data can be projected into a two-dimensional space.

Figure 3 shows the data projection in the two-dimensional space. It presents a progressive processing of feature presentation in both single-task multimodal and multi-task multimodal scenarios.

Our motivation for proposing multi-task-based multimodal learning is that in multimodal learning, if the representations of different modalities are very similar, there will be a lack of complementarity between different modalities, thus affecting the performance of multimodal learning.

F^T and F^A in Figure 3 are the extracted feature representation from FeatureNets. X^T and X^A are the feature representation from the output of the text and audio encoders. H^T and H^A are the feature representation from the output of the text and audio decoders.

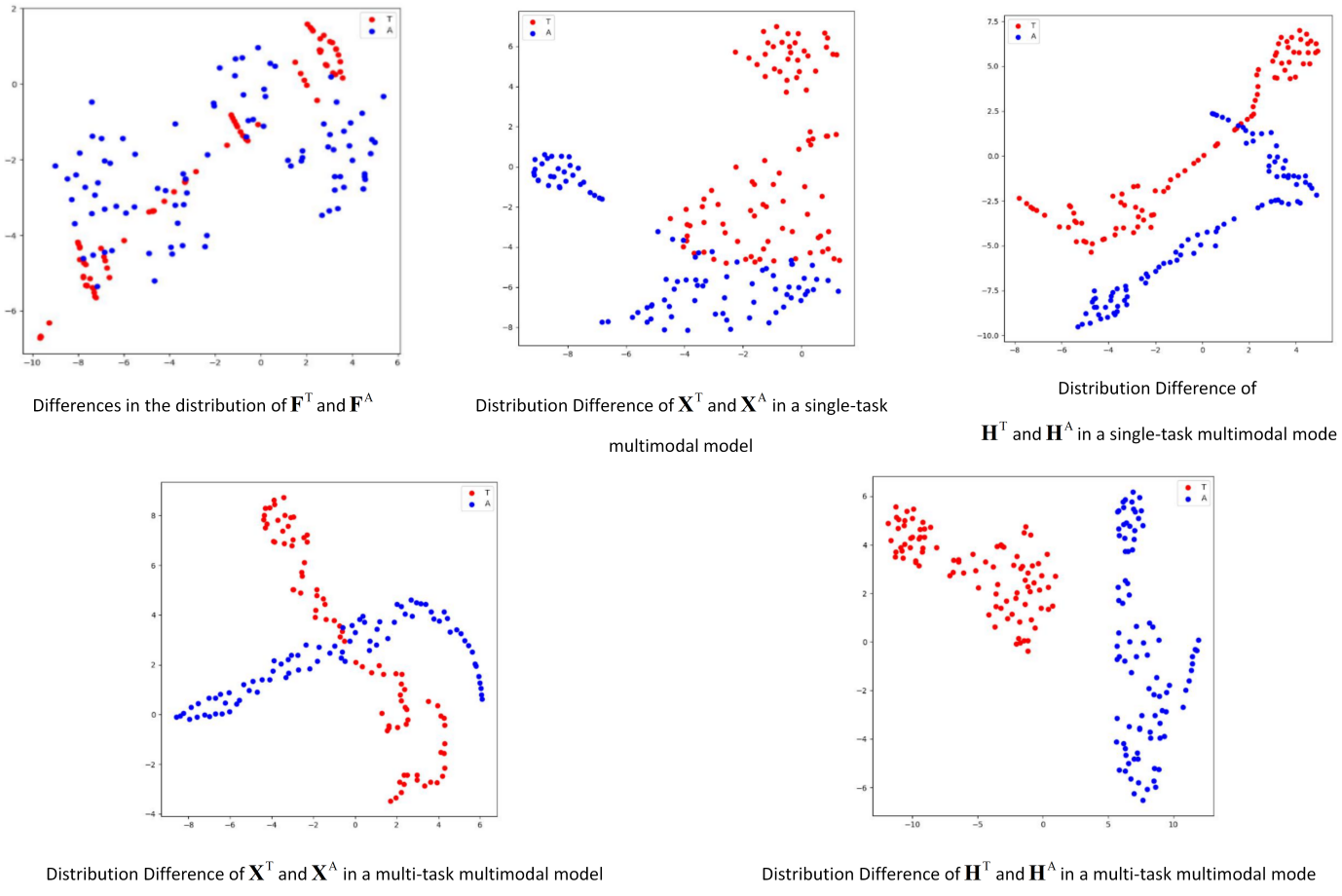


Figure 3. The different modal representations obtained at different stages of the single-task and multi-task multimodal models; the red dots and blue dots represent the single-modal characteristics of text and audio respectively.

The distribution difference of X^T and X^A, H^T and H^A in a multi-task multimodal model is more significant than that in a single-task multimodal model, which means the multi-task multimodal learning has a better representation of different modalities and in different stages than single-task multimodal learning. Therefore, multi-tasking multimodal learning has better results than single-tasking multimodal learning.

5. Ablation Studies

In order to find out the impact of a sub-task of a single modal on the multimodal, we combined different single-modal tasks with the multimodal task to experiment with multi-task learning. The experimental results can be seen in Table 3, which shows how the multimodal task "M" is impacted by single modal task "T" and/or "A".

Table 3. Results for multimodal sentiment analysis with different tasks using MLMF. "M" is the main task, and "A, T" are auxiliary tasks. Only the results of task "M" are reported.

Methods	IEMOCAP (6-Cls)		MELD (7-Cls)	
	Accuracy	Weighted Average F1	Accuracy	Weighted Average F1
M	66.36	66.38	65.94	64.52
M, A	69.62	69.75	66.36	65.11
M, T	69.93	69.62	66.59	65.29
M, A, T	70.73	70.80	65.59	64.97

From the experiment results, after adding single modal tasks "A" and "T", the multimodal task "M" performance is enhanced slightly. When it only has multimodal "M", the text encoder and audio encoder are only supervised by multimodal loss; therefore, they have close results in representation though they have different modalities because those modalities lack complementarity. While adding single modal tasks "T" and "A" for multi-task learning, the text encoder and audio encoder are supervised not only by their joint loss but also by their individual loss. Hence, the representation differences between different modalities are increased. It, therefore, improves the complementarity among different modalities, improving the performance on multimodal tasks.

6. Conclusions

In this paper, we present a novel framework for the emotion recognition in conversation (ERC) task, called multi-task learning and multi-fusion (MMATERIC). Our framework employs multi-task learning and multimodal fusion to capture emotional information from different modalities in conversation. Firstly, we leverage the feature networks component to model the text and speech conversation information; then, we employ the central knowledge processor to fuse multimodal representation. Specifically, we introduce speaker emotion detection unit feeds into the LSTM sequentially, allowing for better tracking of the emotional states of the different speakers and utilizing multimodal emotional features during conversation modeling. Further, a multimodal fusion mechanism is employed to extract vital information from different modalities, thereby improving the ability for multimodal emotion learning. Experimental results on two benchmark datasets show that our proposed MMATERIC framework outperforms previous state-of-the-art methods in the ERC task.

However, we acknowledge that our work does not currently incorporate the speaker's background knowledge, which could further enhance emotion recognition in ERC. To address this limitation, we plan to incorporate speaker knowledge grounded conversation with a global emotion state in our future work. Additionally, our model updates the utterance state using only limited information from recent utterances through a recurrence-based model. We believe that by combining this model with a graph-based model, we can gather information from nearby as well as distant utterances to better model the conversation context. Finally, we also plan to extend our framework to include video, audio, and text modalities in our future work. Overall, we believe that these extensions will enhance the performance of our MMATERIC framework and enable better emotion recognition in conversation.

Author Contributions: Conceptualization and writing, X.L.; formal analysis and data curation, Y.Z., X.Z., T.N., and J.Y.; supervision, R.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China (62006062, 62176076), Shenzhen Foundational Research Funding JCYJJCYJ20210324115614039, and Joint Lab of HITSZ and Konka.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef] [PubMed]
2. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]
3. Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access* **2019**, *7*, 100943–100953. [CrossRef]
4. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
5. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
6. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
8. Chen, S.; Zhang, Y.; Yang, Q. Multi-task learning in natural language processing: An overview. *arXiv* **2021**, arXiv:2109.09138.
9. Xiao, Y.; Li, C.; Liu, V. DFM-GCN: A Multi-Task Learning Recommendation Based on a Deep Graph Neural Network. *Mathematics* **2022**, *10*, 721. [CrossRef]
10. Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively multitask networks for drug discovery. *arXiv* **2015**, arXiv:1502.02072.
11. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8599–8603.
12. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
13. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
14. Long, M.; Cao, Z.; Wang, J.; Yu, P.S. Learning multiple tasks with multilinear relationship networks. *arXiv* **2017**, arXiv:1506.02117.
15. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35; pp. 10790–10797.
16. Wang, Y.; Chen, Z.; Chen, S.; Zhu, Y. MT-TCCT: Multi-task Learning for Multimodal Emotion Recognition. In Proceedings of the International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 429–442.
17. Hu, R.; Singh, A. Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1439–1449.
18. Chen, M.; Zhao, X. A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 374–378.
19. Soleymani, M.; Garcia, D.; Jou, B.; Schuller, B.; Chang, S.F.; Pantic, M. A Survey of Multimodal Sentiment Analysis. *Image Vis. Comput.* **2017**, *65*, 3–14. [CrossRef]
20. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2022**, *91*, 424–444. [CrossRef]
21. Chen, B.; Cao, Q.; Hou, M.; Zhang, Z.; Lu, G.; Zhang, D. Multimodal emotion recognition with temporal and semantic consistency. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3592–3603. [CrossRef]
22. Gu, Y.; Chen, S.; Marsic, I. Deep mul timodal learning for emotion recognition in spoken language. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5079–5083.

23. Rozgić, V.; Ananthakrishnan, S.; Saleem, S.; Kumar, R.; Prasad, R. Ensemble of svm trees for multimodal emotion recognition. In Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Hollywood, CA, USA, 3–6 December 2012; pp. 1–4.
24. Sun, J.; Yin, H.; Tian, Y.; Wu, J.; Shen, L.; Chen, L. Two-level multimodal fusion for sentiment analysis in public security. *Secur. Commun. Netw.* **2021**, *2021*, 6662337. [CrossRef]
25. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.P. Multimodal language analysis with recurrent multistage fusion. *arXiv* **2018**, arXiv:1808.03920.
26. Atmaja, B.T.; Akagi, M. Multitask learning and multistage fusion for dimensional audiovisual 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. 4482–4486.
27. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
28. 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.
29. Zahiri, S.M.; Choi, J.D. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv* **2017**, arXiv:1708.04299.
30. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv* **2017**, arXiv:1710.03957.
31. 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.
32. 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.
33. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.P.; Zimmermann, R. Conversational memory network for emotion recognition in dyadic dialogue videos. *Proc. Conf.* **2018**, **2018**, 2122–2132. [PubMed]
34. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883.
35. Bradbury, J.; Merity, S.; Xiong, C.; Socher, R. Quasi-recurrent neural networks. *arXiv* **2016**, arXiv:1611.01576.
36. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33; pp. 6818–6825.
37. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv* **2019**, arXiv:1908.11540.
38. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed acyclic graph network for conversational emotion recognition. *arXiv* **2021**, arXiv:2105.12907.
39. 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.
40. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
41. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
42. Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; Zimmermann, R. Icon: Interactive conversational memory network for multimodal emotion detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2594–2604.
43. Zhang, H.; Chai, Y. COIN: Conversational Interactive Networks for Emotion Recognition in Conversation. In Proceedings of the Third Workshop on Multimodal Artificial Intelligence, City, Country, Day Month 2021; pp. 12–18.
44. Lu, X.; Zhao, Y.; Wu, Y.; Tian, Y.; Chen, H.; Qin, B. An iterative emotion interaction network for emotion recognition in conversations. In Proceedings of the 28th International Conference on Computational Linguistics, Virtual, 8–13 December 2020; pp. 4078–4088.
45. Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; Poria, S. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv* **2020**, arXiv:2010.02795.
46. Kim, T.; Vossen, P. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv* **2021**, arXiv:2108.12009.
47. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.