Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures

Fazliddin Makhmudov , Alpamis Kultimuratov , Young-Im Cho *

Posted Date: 24 April 2024

doi: 10.20944/preprints202404.1574.v1

Keywords: Deep learning; CNN; BERT; Emotion recognition

 doi:10.20944/preprints202404.1574.v1

Article

Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures

Fazliddin Makhmudov [1], Alpamis Kultimuratov [2] and Young-Im Cho [2,*]

[1] Department of Computer Engineering, Gachon University, Seongnam 1342, Republic of Korea; fazliddin12@gachon.ac.kr
[2] Tashkent State University of Economics, Tashkent 100066, Uzbekistan; (A.K.) a.kutlimuratov@tsue.uz
* Correspondence: yich@gachon.ac.kr

Abstract: Emotion detection holds significant importance in facilitating human-computer interaction, enhancing the depth of engagement. By integrating this capability, we pave the way for forthcoming AI technologies to possess a blend of cognitive and emotional understanding, bridging the divide between machine functionality and human emotional complexity. This progress has the potential to reshape how machines perceive and respond to human emotions, ushering in an era of empathetic and intuitive artificial systems. This paper introduces a novel approach to multimodal emotion recognition, seamlessly integrating speech and text modalities to accurately infer emotional states. Employing CNNs, we meticulously analyze speech using Mel spectrograms, while a BERT-based model processes the textual component, leveraging its bidirectional layers for profound semantic comprehension. The outputs from both modalities are combined using an attention-based fusion mechanism that optimally weighs their contributions. The proposed method undergoes meticulous testing on two distinct datasets: Carnegie Mellon University's Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset and the Multimodal Emotion Lines Dataset (MELD). The results demonstrate superior efficacy compared to existing frameworks, achieving an accuracy of 88.4% and an F1-score of 87.9% on the CMU-MOSEI dataset, and a notable weighted accuracy (WA) of 67.81% and a weighted F1 (WF1) score of 66.32% on the MELD dataset. This comprehensive system offers precise emotion detection and introduces several significant advancements in the field.

Keywords: deep learning; CNN; BERT; emotion recognition

1. Introduction

In the contemporary era, the exponential proliferation of multimedia content has led to a precipitously rising interest in the domain of multimodal data interpretation among academic scholars. This discipline encompasses the evaluation of data through diverse channels, such as visual and textual streams [1-3]. A pivotal and emergent subfield in this realm is the study of multimodal emotional analysis, which involves deciphering human sentiments as manifested across diverse communication modalities. This avenue of innovation has witnessed escalating momentum within the scholarly fraternity over the past few years. Multimodal frameworks endorse an interactive paradigm that emulates human dialogue by permitting the concurrent utilization of multiple data ingestion and emission channels. The liberty to opt for an array of these modalities cultivates a more anthropomorphic dialogue experience, enhancing user immersion and communication efficacy. A significant body of scholarly work has been dedicated to exploring sentiment discernment in written content [4,5], facial affective interpretation [6,7], and auditory emotional cognition [8,9]. Nevertheless, scrutiny of these outcomes indicates that inquiries centered on singular modes have encountered a particular impasse, prompting amplified scholarly interest in the utilization of multimodal methodology. Amalgamating ancillary data from pictorial and textual sources can augment the fidelity of affective discernment, and thus foster the evolution of machines equipped with empathic faculties. The domain of multimodal affective identification encompasses an extensive array of

 doi:10.20944/preprints202404.1574.v1

implementations such as human-machine interfacing, automated detection systems, intelligent playthings, protective mechanisms, service dispensation, and linguistic transposition. Within the realm of multimodal emotion recognition, the paramount phases encompass the derivation of distinguishing attributes and the amalgamation of these elements across diverse modalities. The objective of attribute derivation is to discern the vital constituents of signals, formulate vector representations predicted on these constituents, and employ them to classify pertinent emotions. This process streamlines the subsequent phases of the emotion discernment task [10]. Prior investigations typically derived features intrinsic to each modality, such as lexical representations from textual data [11], temporal acoustic attributes from vocal utterances [12], and visual cues from facial imagery [13]. Existing extraction methodologies include techniques such as Fourier transformation, wavelet transformation, and high-order cross processes, which can be collectively categorized as low-level feature extraction methods. However, given the dichotomy between human affective states and low-level attributes, these extraction strategies may not be adequately robust for multimodal emotion identification. Conversely, deep-learning methodologies can detect the dispersion attributes of datasets by integrating basal characteristics, thereby constructing a more abstract, superior-level portrayal of information. Furthermore, multimodal amalgamation, which entails merging data obtained from various channels, profoundly affects the efficiency and outcomes of multimodal emotional discernment [14,15]. This amalgamation procedure provides supplementary data, consequently improving the fidelity of emotional identification. Therefore, conceptualizing a more advanced fusion strategy is imperative to yield a superior optimized discriminative feature representation for multimodal emotion recognition.

Emotional expressions often manifest as unique energy configurations when visualized in spectrograms. This means that when emotions are conveyed, especially in auditory formats such as speech, they can produce specific patterns or distributions of energy across various frequencies, making them distinguishable when analyzed through spectrograms. A convolutional neural network (CNN) exhibits a pronounced inclination towards emphasizing regions in data that exhibit high energy levels or sudden transitions in energy values. This intrinsic behavior of CNNs is especially advantageous when analyzing emotions, as they are adept at capturing emotions that manifest rapidly and with a heightened intensity. This focus enables the model to detect and differentiate between subtle and strong emotional cues, which is particularly beneficial in scenarios where emotions may surge abruptly or peak momentarily, thereby ensuring a comprehensive understanding of the emotional landscape.

However, text is an important medium for expressing and deciphering emotions. While spoken words and physical expressions often provide direct cues about feelings, written language carries its own set of nuanced emotional indicators. Through word choices, phrasing, and punctuation, authors can convey a vast spectrum of emotions, from joy and excitement to sadness and despair. One of the primary obstacles in constructing text-based models is the need for vast quantities of relevant data. Crafting efficient and accurate text models depends on the availability and accessibility of large datasets tailored for specific tasks. The larger and more diverse the dataset is, the better the model can generalize and adapt to real-world scenarios.

To address this challenge, the application of transfer learning and/or utilization of pre-trained models have emerged as promising solutions. These techniques leverage the knowledge gained from one task and apply it to a related task, potentially reducing the need for extensive datasets specific to each challenge.

Our study introduced a multimodal emotion-recognition system that combines speech and text data. First, a speech module was crafted using CNNs to extract patterns from the mel spectrograms, to produce a fully connected (FC) layer. This method excels in revealing intricate details in the time and frequency domains.

Next, we employed a pre-trained BERT-based model for our text component. BERT's 12-layer bidirectional structure offers a profound understanding of textual semantics. Leveraging BERT's extensive training on large-scale datasets allows for a more enriched emotional representation. Post-BERT processing involved an FC layer for reducing dimensionality, a Bi-GRU for contextual

 doi:10.20944/preprints202404.1574.v1

understanding, and another FC layer to link features to specific emotions. These refined outputs were fed into our fusion module, thereby strengthening our emotion detection mechanism.

Finally, the fusion process involves an attention-based mechanism to balance the speech and text contributions. It allocates attention scores, prioritizing the most pertinent modality based on the context. We standardized the feature dimensions using an FC layer for unbiased attention allocation. The consolidated features were then passed through a softmax layer to classify the embedded emotion, ensuring comprehensive emotion analysis. This study introduced several vital contributions that deserve special emphasis. The research:

- unveiled an innovative multimodal emotion detection method that surpassed the precision of prevailing benchmark models. This groundbreaking technique will facilitate future studies on emotion detection.
- innovatively leveraged convolutional neural networks and a fully connected layer to extract mel-spectrogram features, offering enhanced accuracy and bridging the divide between raw speech and emotional nuances.
- employed a pre-trained BERT-based model to encode textual data effectively, capitalizing on its bidirectional transformer encoders to capture context-rich semantic information.
- employed an attention-based fusion mechanism for a bimodal system that emphasizes speech and text, dynamically prioritizes modalities, standardizes feature vectors, and utilizes multistage processing to enhance sentiment classification accuracy.
- conducted comprehensive tests on two benchmark datasets, CMU-MOSEI and MELD, and the outcomes clearly highlighted the superior performance of our approach in emotion recognition.

This article is arranged as follows. Section 2 provides a summary of pertinent research in the realm of multimodal emotion recognition. Section 3 describes the overall scheme of the proposed model. Section 4 presents an exhaustive account of the dataset used in the experiment, accompanied by the results, essential analyses, and the model's efficacy is discussed. Finally, the conclusions are presented.

2. Literature Review

The rapid increase in social media data surpasses the limitations of using only one type of content. The traditional method of depending on a network model that focuses on a single feature type has proven insufficient for accurately detecting speakers' emotions. Therefore, researchers have shifted their focus to improving recognition accuracy by incorporating additional modalities. Consequently, interest has increased in exploring multimodal emotion recognition, specifically incorporating data from various domains such as text and speech cues.

In recent years, the emotion-recognition domain has witnessed a significant increase in the successful formulation and utilization of pre-trained models. These models, trained through self-supervised learning on large quantities of unlabeled data, have exhibited notable accomplishments. Among these models, language architectures such as GPT and BERT have gained considerable recognition. The success of pre-training in the textual modality has inspired researchers dealing with other data types. In particular, there have been significant efforts towards formulating speech-language cross-modality pre-trained models. These latest models [16-19] were designed to comprehend and interpret information from both text and speech, enabling more intricate understanding and interpretation. The authors of [20] proposed an audio-text crossmodal transformer model built on the principles of RoBERTa and HuBERT, which are individually pre-trained unimodal models. Furthermore, they described a novel phase-oriented training methodology for their cross-modal transformer model, which included preliminary pre-training, task-specific adaptive pre-training, and eventual fine-tuning for particular downstream applications. The research in [21] was centered on speech emotion recognition, introducing an enhanced, emotion-focused pre-trained encoder known as Vesper. By utilizing a speech dataset with WavLM and adopting an emotion-led masking approach, Vesper used hierarchical and cross-layer self-supervision to capture effectively the acoustic and semantic representations that are crucial for emotion detection. Moreover,

 doi:10.20944/preprints202404.1574.v1

in [22] Hang et al. unveiled CTAL, a cross-modal transformer for audio and language, which aspires to decipher the intra- and inter-modality connections between auditory and linguistic elements. This objective was achieved by deploying two proxy tasks, masked language modeling and masked cross-modal acoustic modeling, on a large corpus of audio-language paired data.

Regarding multimodal speech emotion recognition, the authors in [23] introduced a novel strategy, referred to as a key-sparse Transformer, which evaluates the relevance of each word or speech frame within a sample, thereby allowing the model to focus more on emotion-related information. Leveraging the KS-Transformer, they further developed a cascaded cross-attention mechanism that enabled the highly efficient fusion of different modalities. The study in [24] introduced the LGCCT, a transformer engineered for multimodal speech emotion recognition that effectively blends information from different modalities. It employs CNN-BiLSTM to extract acoustic features and BiLSTM to gather textual features fused using a cross-attention module. A gate-control mechanism is implemented to balance the combination of the original and fused modality representations. Finally, they considered the level of attention focus by adopting a length-scaled dot product to calculate attention scores and ensure model adaptability across various testing sequence lengths. Furthermore, merging at the fusion stage, which is essential for multimodal emotion detection typically involves feature combinations to map distinct modalities into a unified domain, as indicated in several studies [25-27]. Otherwise expressed, a multilevel multimodal dynamic integration system [28] was introduced to create a cohesive representation centered on intermodal connections. This process begins by scrutinizing the latent yet significant relationships among disparate features, each individually harvested from several modalities, using a particular methodology. This investigation succeeded in the development of a multilevel integration network that subdivided the integration process into various phases based on previously discovered correlations. This arrangement facilitates the capture of more nuanced unimodal, bimodal, and trimodal linkages. By examining and analyzing various feature integration methods for text and speech separately, a multimodal feature fusion approach [29] was suggested for imbalanced sample data. The aim of this proposal is to implement multimodal emotion recognition effectively. In addition, [30] introduced a quantum neural network-driven multimodal integration framework for smart diagnostics capable of handling multimodal healthcare data relayed by the Internet of Medical Things devices. This system amalgamates data from diverse modalities, thereby enhancing the effectiveness of smart diagnostics. It leverages a quantum convolutional neural network for the efficient extraction of features from medical imagery.

In this research, efforts have been made to suggest an enhanced multimodal emotion-recognition framework that leverages speech and text models. Fundamentally, the integration of various modules can boost the accuracy of emotion-detection models. This approach helps to diminish the influence of single-feature biases and augments the model's capacity to navigate different emotional states and speech contexts effectively.

3. Proposed Multimodal Emotion-Recognition Approach

Figure 1 illustrates how the proposed method harnesses the potential of both the speech and text modalities for emotion recognition. This approach acknowledges the rich and complex information conveyed through spoken words as well as the nuanced meanings captured in written text. By integrating these modalities, the method aims to achieve a more comprehensive and accurate understanding of emotional states. It leverages the strengths of both the speech and text analyses, potentially improving the efficacy of emotion-recognition systems. Specifically, emotional expressions extracted from both text and speech are introduced into a fusion module. The output derived from the fusion module is subsequently introduced into an FC layer. This step allows the consolidation of high-dimensional multimodal features. The final part of this process includes the application of a softmax layer, which computes the probabilities corresponding to different emotions. Each component of the emotion-recognition pipeline is explored in detail in the following subsections.

 doi:10.20944/preprints202404.1574.v1

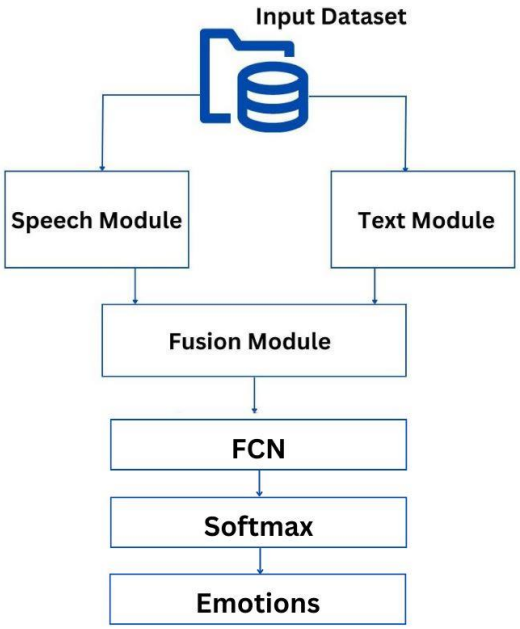

Figure 1. Proposed multimodal emotion-recognition approach.

3.1. Speech Module

To leverage the benefits of mel spectrogram features, our speech module includes a series of convolutional neural networks preceding a fully connected layer. This design captures the intricate patterns within the mel spectrogram and for enhanced performance and accuracy, ensures that the convolutional layers effectively extract the spatial characteristics, which are then synthesized in the fully connected layer. Specifically, we drew inspiration from the work outlined in [31], and adapted the architecture by implementing changes tailored to our objectives. This allowed us to build on established ideas while incorporating our unique modifications to better suit the needs of the proposed multimodal emotion-recognition approach. Figure 2 presents an in-depth depiction of the speech module architecture.

Initially, we utilized the 'Librosa' library [32] to procure the mel spectrogram features from the raw speech data. The extracted features were then channeled into two distinct convolutional layers that ran concurrently. These layers are specifically designed to capture and highlight patterns from both the time and frequency domains, ensuring a comprehensive analysis of the audio signal characteristics. We employed two simultaneous convolutional layers as our initial layers, with kernel sizes of (10,2) and (2,8). After padding, the output from each convolutional layer yielded eight channels. The outputs were merged, resulting in a combined 16-channel representation for further processing. Subsequently, four additional convolutional layers were employed to produce an 80-channel representation, which was forwarded to the fully connected layer for further analysis and processing. Following each convolutional layer, we incorporated a batch normalization (BN) layer coupled with a rectified linear unit (ReLU) activation function to enhance the stability and performance of the network. Moreover, convolutional layers 1 and 2 preceded a max-pooling operation with a kernel size of two that reduced data dimensionality and further streamlined processing.

Throughout these convolutional layers, the kernel size remained consistent at 3×3 with a stride of one. However, as we progressed from convolutional layers 1 and 4, the number of input channels started at 16 and doubled, subsequently increasing in increments of 16, ultimately reaching 80 output channels in convolutional layer 4. This structured increment in channels suggests a hierarchical feature extraction process wherein each subsequent layer aims to extract more complex and nuanced features from the input.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2024 doi:10.20944/preprints202404.1574.v1

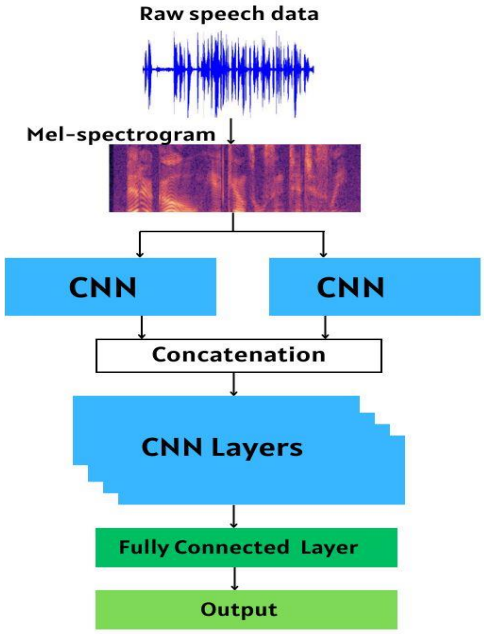

Figure 2. Speech module of the proposed multimodal emotion-recognition approach.

3.2. Text Module

The origin of textual modality can be traced back to the process of transcribing uttered speech. This involves converting spoken words into written forms, thereby creating a textual representation that can be easily analyzed and interpreted. This is a crucial step in the data-preparation phase for emotion recognition and other natural language processing tasks. Using a written transcript of spoken language, apply various text analysis techniques can be applied to extract meaningful features such as sentiment, tone, and other linguistic characteristics that could indicate emotional states. This extends our ability to understand and analyze the emotions expressed through the tone and inflection of speech and the choice of words and phrases, their arrangement, and other textual elements. Several strategies have been used to extract features from textual modalities. In the proposed approach, we employ a series of techniques, as depicted in Figure 3.

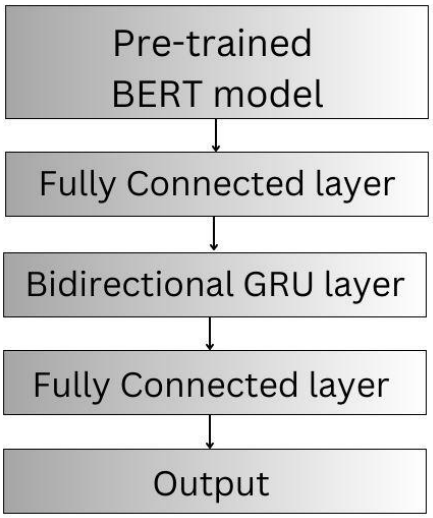

Figure 3. Text module of the proposed multimodal emotion-recognition approach.

Specifically, our methodology is initiated by utilizing a pre-trained BERT-based model [33] to encode textual data. The BERT-based model comprises 12 layers of transformer encoders. These transformer encoders are exceptional since they allow consideration of context from both directions,

 doi:10.20944/preprints202404.1574.v1

that is, left-to-right and right-to-left, for every layer of the model. The bidirectional nature of transformer encoders aids in capturing context more effectively, thereby providing a robust and nuanced understanding of a text's semantic meaning. It is worth noting that by leveraging a pre-trained BERT-based model, we gained from the model's extensive learning previously undergone on a massive corpus of text. This helps generate more meaningful and accurate representations of our textual data, which is especially beneficial when dealing with complex emotion-recognition tasks. In essence, using the pre-trained BERT-based model for text encoding sets a strong foundation for our proposed approach, offering the capacity to extract valuable insights from the text, thereby significantly improving the proposed emotion-recognition system's performance.

After leveraging the BERT-based model, an FC layer was used. This layer simplifies the BERT-generated high-dimensional vectors into 100 dimensions, creating compact yet effective representations termed the utterance features of the text modality. This process aids in managing computational complexity and overfitting while retaining essential textual information, serving as a robust basis for emotion-recognition analysis. After feature reduction, a bidirectional gated recurrent unit (Bi-GRU) was used to encode all utterances, capturing past and future dialogue information. Such bidirectional processing bolsters the model's understanding of the context within a conversation. The Bi-GRU hidden states were set to 100 dimensions in line with our prior feature reduction, ensuring consistency and capturing vital temporal and contextual dialogue information. After processing through the Bi-GRU layer, the resulting outputs were fed into the FC layer. This layer functioned as a translator, converting high-level features into dimensions representing different emotional categories. This was a critical bridge linking the complex features generated by the Bi-GRU to various emotion classes, thereby boosting the efficacy of our emotion-recognition system. The output generated from the FC layer was channeled into the fusion module.

To train the text module in our emotion-recognition system, we followed a robust approach. This involved running 300 epochs with 30 data batches. To prevent overfitting, we applied a dropout of 0.3 and L2 regularization with a weight of 0.0001. For efficient training, we used the Adam optimizer with an initial learning rate of 0.0005, which decayed at a rate of 0.0001.

3.3. Fusion Module and Classification

When classifying sentiments, not every modality contributes relevance or significance to the same degree. Different modalities such as text, speech, and facial expressions offer unique aspects of emotional insight. However, their contributions to sentiment classification vary significantly. Some may have a more pronounced influence, whereas others may make only subtle contributions. This discrepancy arises because of the inherent differences in the types of information that these modalities encapsulate as well as the varying capacities in which they express emotional cues. Therefore, when designing a multimodal system for sentiment classification, accounting for the diverse significance levels of the different modalities and strategically balancing their integration to achieve optimal performance, is crucial.

In the fusion module of our proposed multimodal emotion-recognition approach, we incorporated an attention-based fusion mechanism, as outlined in existing research [34]. This sophisticated approach allowed our system to assign varying degrees of importance to different modalities during the fusion process. The intention is to focus more heavily on the modalities that are considered most significant for a given context or dataset. This strategic prioritization helps refine the system output, enabling it to yield more accurate and contextually appropriate results in sentiment classification tasks. The research outlined in [34] incorporated three distinct modalities – audio, visual, and textual – to generate an attention score for each modality and determine its relevance to the final output. In contrast, our approach operates on a bimodal system, focusing solely on speech and text modalities. Despite this, our model effectively assigned attention scores to each of these two modalities, allowing for the dynamic allocation of significance based on their inherent contributions to the emotion-recognition task.

Before inputting the feature vectors from the speech and text modalities into the attention network, their dimensions were uniform. This was accomplished using an FC layer of size n. By

 doi:10.20944/preprints202404.1574.v1

implementing this layer, we could effectively transform feature vectors into equal dimensions, thereby ensuring that each modality was represented equally when processed in the attention network. This step was vital for maintaining fairness and balance in the allocation of attention weights between the two modalities.

Consider $F = [F_s, F_t]$as the standardized feature set, where 'F_s' represents the acoustic features, and 'F_t' denotes the textual features. Here, the dimensionality of each feature set has been equalized to size 'n,' resulting in 'F' belonging to the dimensional space $F \in \mathbb{R}^{n\times 2}$. Optimal performance was achieved when the value o of n was set to 250. The attention weight vector, denoted as v_{at_fusion} and the fused multimodal feature vector, referred to as M_f, were calculated using the following procedures:

$$L_{M_f} = tanh(P_{M_f} \cdot F) \tag{1}$$

$$v_{at\_fusion} = softmax(w_{M_f}^T \cdot L_{M_f} \tag{2}$$

$$M_f = F \cdot v_{at\_fusion}^T \tag{3}$$

where $P_{M_f} \in \mathbb{R}^{n\times n}$, $w_{M_f} \in \mathbb{R}^n$, $v_{at\_fusion}^T \in \mathbb{R}^2$, and $M_f \in \mathbb{R}^n$. The resulting output, denoted as M_f, signified the combined multimodal feature vector. This composite representation was then processed through a fully connected layer, which acted as an intermediary step to consolidate the high-dimensional multimodal features to be more manageable. Subsequently, a softmax layer was employed to finalize the speech emotion classification task. The softmax layer worked by outputting a probability distribution over predefined emotional categories, thus determining the most likely emotion to be present in the speech. This multistage process ensured a thorough evaluation of multimodal features and effectively identifies nuanced emotional content within a given speech input.

4. Results and Discussion

4.1. Datasets

To validate the efficacy of the proposed model, two well-established datasets, namely MELD MELD [35] and CMU-MOSEI [36], were employed. These datasets, which are rich in multimodal emotional content, provide a comprehensive foundation for evaluating a model's proficiency in multimodal emotion-recognition tasks. Further details regarding the characteristics of these datasets and their contributions to the assessment of the model are discussed below.

4.1.1. MELD

MELD is a large-scale multimodal dataset specifically designed to enhance emotion-recognition research. Developed by refining and augmenting the EmotionLines dataset, MELD includes data from various modalities such as text, audio, and visual cues, which are further associated with six different emotional categories.

The MELD dataset is unique since it is derived from the popular television series, "Friends." This source provides real-life conversations among multiple speakers, which are inherently dynamic and rich in emotional expressions. The dataset includes a large number of dialogue instances, capturing over 1400 dialogues and 13000 utterances from the series.

Each utterance in the MELD dataset is annotated using emotion and sentiment labels. The six emotions are anger, disgust, joy, sadness, surprise, and neutrality. The dataset also contains sentiment labels, such as positive, negative, and neutral.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2024 doi:10.20944/preprints202404.1574.v1

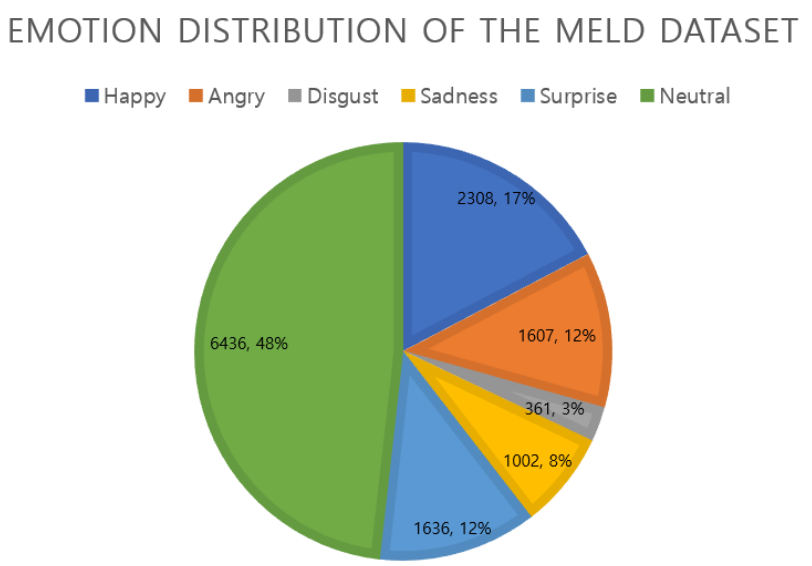

Figure 4. Emotion data distribution of the MELD dataset.

4.1.2. CMU-MOSEI

The CMU-MOSEI dataset is an extensive multimodal compilation of conversational video data dedicated to emotion recognition. This dataset includes more than 23,000 video fragments extracted from 1,000 distinct sessions involving over 1,200 contributors. The video entries are accompanied by speech transcripts, auditory and visual characteristics, and labels signifying varying degrees of valence and arousal. The CMU-MOSEI dataset classifies emotions into six categories: anger, happiness, sadness, disgust, fear, and surprise. The dataset encompasses a wide variety of emotional samples: anger (4600), sadness (5601), disgust (3755), surprise (2055), happiness (10752), and fear (1803). This variety ensures a comprehensive representation of emotional states, facilitating a more robust and accurate analysis in subsequent emotion-recognition studies.

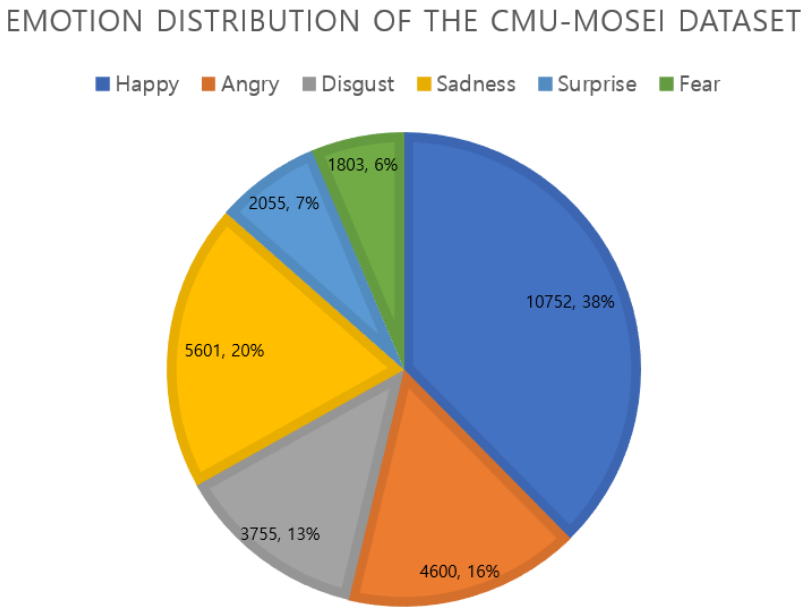

Figure 5. Emotion data distribution of the CMU-MOSEI dataset.

 doi:10.20944/preprints202404.1574.v1

4.2. Implementation Configuration

Our model outcomes are represented using the Accuracy and F1-score metrics on the CMU-MOSEI dataset and the WA and WF1-score metrics on the MELD dataset, owing to the inherent disparity among different emotions [35].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{4}$$

$$F_1 = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \tag{5}$$

$$\text{WA} = \frac{\text{TP} \times \text{N/P} + \text{TN}}{2\text{N}} \tag{6}$$

The weighted accuracy, which is identical to the mean recall across all emotion categories, provides an understanding of the model's performance considering the imbalance between different classes.

To evaluate our model fairly and objectively using the MELD and CMU-MOSEI datasets, we employed a robust training procedure as explained in [37]. This involved restructuring the original datasets and subsequently dividing the data into training and testing subsets (Table 1), allocating 80% of the data for training and 20% for testing.

Table 1. Statistics of the MELD and CMU-MOSEI datasets.

Datasets	Categories	Emotions							Percentage
		Happy	Angry	Disgust	Sadness	Surprise	Neutral	Fear	
MELD	Train	1846	1285	288	801	1308	5148	–	80%
	Test	462	322	73	201	328	1288	–	20%
CMU-MOSEI	Train	8601	3680	3004	4480	1644	–	1442	80%
	Test	2151	920	751	1121	411	–	361	20%

This methodology ensured a comprehensive and detailed evaluation of the performance of the proposed model. This procedure yielded 10676 training samples and 2674 testing samples from the MELD dataset. Similarly, for the CMU-MOSEI dataset, 22,851 training and 5,715 testing samples were obtained. Unlike the methodology outlined in [38], we did not implement 10-fold cross-validation in our study. This choice was driven by the practical complications associated with deploying cross-validation on deep-learning models, considering the extensive time and computational resources required.

The proposed model was subjected to an extensive training and testing regimen of over 300 epochs using batches of 32. To execute these deep learning tasks efficiently we used an Nvidia GeForce RTX 3090 24 GB graphics processing unit coupled with an Intel Core i7-13700K 10-Core Processor. The Ubuntu platform was complemented by 128 GB of RAM, and this provided a robust and high-performance computational environment for model training and evaluation.

4.3. Recognition Performances

In multimodal emotion recognition, several studies have used various datasets and modalities to evaluate the effectiveness of their methodologies. A crucial aspect of these evaluations lies in performance metrics such as accuracy and F1-Score, which provide insights into their efficacy and reliability.

Table 2 presents a comparative analysis of emotion-recognition studies using multimodal data. Focusing on the performance metrics of the Accuracy and F1-Score, we observed a range of results. Li et al. [39] reported a commendable accuracy of 81.57% and an F1-Score of 81.16%. Close on its heels is Delbrouck et al [40] with an accuracy of 81.52%, though its F1-Score is not provided. A slight dip

 doi:10.20944/preprints202404.1574.v1

is seen in Bi et al [41] who recorded an accuracy of 75.0% and an F1-Score of 74.5%. Lio et al [42] only offer an F1-Score of 81.0%, omitting accuracy data. Remarkably, our method, denoted as Ours, outperforms the referenced studies with an accuracy of 83.2% and an F1-Score of 82.9%. This suggests that our methodology is competitive and also potentially sets a new benchmark in the realm of emotion recognition using the specified modality.

Table 2. Comparison of the recognition performances on the CMU-MOSEI dataset.

Research	Modality	Metrics	
		Accuracy	F1
Li et al [39]	Speech + Text	81.57	81.16
Delbrouck et al [40]	Speech + Text	81.52	–
Bi et al [41]	Speech + Text	75.0	74.5
Lio et al [42]	Speech + Text	–	81.0
Ours	Speech + Text	88.4	87.9

Following our earlier evaluation of the CMU-MOSEI dataset, we further expanded our comparative analysis to include the results of the MELD dataset, another key benchmark in the domain of emotion recognition. This additional comparison is vital to ensure the robustness and adaptability of the proposed system across diverse datasets.

Table 3 presents the performance metrics of various studies on the MELD dataset. Each of these studies consistently employs the "Speech + Text" modality. Guo et al. [43] developed a list with a WA of 54.79% and WF1 of 48.96%. Interestingly, while Soumya et al. [44] and Sharma et al. [45] only disclosed their WF1 scores at 65.8% and 53.42%, respectively, Lian et al. [46] presented a more comprehensive result with a WA of 65.59% and a WF1 of 64.50%. Significantly, our approach surpasses these metrics with a leading WA of 66.81% and WF1 of 66.12%. This further consolidation of the MELD dataset underscores the efficiency and superiority of the proposed methodology for multimodal emotion recognition.

Table 3. Comparison of the recognition performances on the MELD dataset.

Research	Modality	Metrics	
		WA	WF1
Guo et al [43]	Speech + Text	54.79	48.96
Soumya et al [44]	Speech + Text	–	65.8
Sharma et al [45]	Speech + Text	–	53.42
Lian et al [46]	Speech + Text	65.59	64.50
Ours	Speech + Text	66.81	66.12

4.4. Discussion

In this study, we introduced a multimodal emotion recognition approach that leverages both the speech and text modalities to better understand emotional states. From the experiments and results presented, several key insights emerged that provided a comprehensive understanding of the potential and efficacy of the proposed model.

The proposed system, which harnesses the strengths of the speech and text modalities, displays a marked improvement in emotion-recognition accuracy. The fusion of these modalities, coupled with the attention mechanism, enables the model to focus on the most pertinent features, thereby enhancing the emotion detection process.

 doi:10.20944/preprints202404.1574.v1

Contrasting our results with the existing literature reveals that our model offers an advanced degree of precision. The nuances of emotional states, which are often overlooked by unimodal systems, were captured using a multimodal approach. Our model's ability to prioritize modalities depending on the context is particularly noteworthy and could be the foundation for its superior performance.

Although our findings are promising, there were some limitations. The current model primarily focuses on the speech and text modalities, omitting visual cues that are crucial for emotion recognition. Furthermore, the efficiency and scalability of the model in real-world scenarios and on external controlled datasets must be explored in depth.

There are numerous potential applications of such comprehensive emotion-recognition systems. From customer service chatbots' understanding of user sentiments to mental health applications detecting potential distress signals in speech, the real-world utility of the proposed model cannot be underestimated.

In continuing this work, integrating visual cues would be fruitful to further refine the emotion-recognition process. Investigating the adaptability of the model across different languages and cultural contexts would also be beneficial. In addition, exploring methods to streamline the model for more efficient real-time processing can create opportunities for more immediate applications.

5. Conclusions

In this study, we delved deeply into a multimodal approach, consolidating speech and text, to develop a potent and efficient emotion-recognition system. The robust architecture, from convolutional neural networks for speech to the pre-trained BERT-based model for text, ensures a comprehensive analysis of inputs. Our attention-based fusion mechanism stands out, demonstrating its ability to discern and weight the contributions from each modality. As emotion recognition evolves, methodologies, such as ours, will facilitate future innovations, showcasing the potential for multimodal integration. Although our system has garnered commendable outcomes on the CMU-MOSEI and MELD datasets, it has certain limitations. First, although using a pre-trained model provides the advantage of leveraging extensive training on a vast corpus of text, it may not always perfectly align with specific domain requirements. Certain nuances or domain-specific emotional cues in a dataset may not be captured efficiently. Second, the system design incorporates several layers, including convolutional layers, transformer encoders, FC layers, and attention mechanisms. This complexity can introduce significant computational demands that may not be feasible for real-time applications or systems with limited computational resources. Therefore, this study paves the way for numerous potential avenues for future research. Thus, future studies should design lightweight architectures or employ model quantization and pruning techniques to render the system feasible for real-time applications. Moreover, we aim to develop models that recognize cultural nuances in emotional expressions to ensure that they are universally applicable.

Author Contributions: This manuscript was designed and written by F.M. and A.K. and Y.-I.C. supervised the study and contributed to the analysis and discussion of the algorithm and experimental results. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Republic of Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2017-0-01630) supervised by the IITP (Institute for Information & communications Technology Promotion) and by the Gachon University research fund of 2020 (GCU-202004350001).

Acknowledgments: The authors F.M., A.K., would like to express their sincere gratitude and appreciation to the supervisor, Young Im Cho (Gachon University) for her support, comments, remarks, and engagement over the period in which this manuscript was written. Moreover, the authors would like to thank the editor and anonymous referees for the constructive comments in improving the contents and presentation of this paper.

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

 doi:10.20944/preprints202404.1574.v1

References

1. Sun, Z., Sarma, P., Sethares, W., and Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proc. AAAI Conf. Artif. Intell. 34, 8992–8999. doi: 10.1609/aaai.v34i05.6431
2. Pepino, P. Riera, L. Ferrer, A. Gravano, Fusion approaches for emotion recognition from speech using acoustic and text-based features, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6484–6488.
3. A. Abdusalomov, A. Kutlimuratov, R. Nasimov and T. K. Whangbo, "Improved speech emotion recognition focusing on high-level data representations and swift feature extraction calculation," *Computers, Materials & Continua*, vol. 77, no.3, pp. 2915–2933, 2023.
4. Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, **Simon Karanja Hingaa**, and Amena Mahmoud. Text-Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience, 2022. 1687-5265. https://doi.org/10.1155/2022/2645381
5. Zygadło, A.; Kozłowski, M.; Janicki, A. Text-Based Emotion Recognition in English and Polish for Therapeutic Chatbot. *Appl. Sci.* **2021**, *11*, 10146. https://doi.org/10.3390/app112110146
6. Deepak Kumar Jain, Pourya Shamsolmoali, Paramjit Sehdev. Extended deep neural network for facial emotion recognition. Pattern Recognition Letters, Volume 120, 2019. https://doi.org/10.1016/j.patrec.2019.01.008.
7. Khattak, A., Asghar, M.Z., Ali, M. *et al.* An efficient deep learning technique for facial emotion recognition. *Multimed Tools Appl* **81**, 1649–1683 (2022). https://doi.org/10.1007/s11042-021-11298-w
8. Makhmudov, F.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047. https://doi.org/10.3390/electronics11234047
9. S. Akinpelu, S. Viriri and A. Adegun, "Lightweight Deep Learning Framework for Speech Emotion Recognition," in *IEEE Access*, doi: 10.1109/ACCESS.2023.3297269.
10. Zhang, Y., Cheng, C. & Zhang, Y. Multimodal emotion recognition based on manifold learning and convolution neural network. *Multimed Tools Appl* **81**, 33253–33268 (2022). https://doi.org/10.1007/s11042-022-13149-8
11. L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu and S. Ding, "Emotion Recognition With Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information," in *IEEE MultiMedia*, vol. 29, no. 2, pp. 94-103, 1 April-June 2022, doi: 10.1109/MMUL.2022.3161411.
12. Ravi Raj Choudhary and Gaurav Meena and Krishna Kumar Mohbey. Speech Emotion Based Sentiment Recognition using Deep Neural Networks. 2022. https://dx.doi.org/10.1088/1742-6596/2236/1/012003
13. Martin Maier, Florian Blume, Pia Bideau, Olaf Hellwich, Rasha Abdel Rahman. Knowledge-augmented face perception: Prospects for the Bayesian brain-framework to align AI and human vision. Consciousness and Cognition. Volume 101,2022. https://doi.org/10.1016/j.concog.2022.103301.
14. J. Heredia *et al*., "Adaptive Multimodal Emotion Detection Architecture for Social Robots," in *IEEE Access*, vol. 10, pp. 20727-20744, 2022, doi: 10.1109/ACCESS.2022.3149214.
15. Tzirakis, Panagiotis & Chen, Jiaxin & Zafeiriou, Stefanos & Schuller, Björn. (2021). End-to-end multimodal affect recognition in real-world environments. Information Fusion. 68. 46-53. 10.1016/j.inffus.2020.10.011.
16. Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu and Yongbin Li. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. 2022. https://doi.org/10.48550/arXiv.2211.11256
17. Jacob Devlin, Ming-Wei Chang, Kenton Lee et al., "Bert: Pre-training of deep bidirectional transformers for language understanding", *NAACL-HLT (1)*, 2019.
18. Zhang Hua, Gou Ruoyun, Shang Jili, Shen Fangyao, Wu Yifan, Dai Guojun. Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition. Frontiers in Physiology. 12, 2021. https://www.frontiersin.org/articles/10.3389/fphys.2021.643202
19. Ilyosov, A.; Kutlimuratov, A.; Whangbo, T.-K. Deep-Sequence–Aware Candidate Generation for e-Learning System. *Processes* **2021**, *9*, 1454. https://doi.org/10.3390/pr9081454
20. Iek-Heng Chu, Ziyi Chen, Xinlu Yu, Mei Han, Jing Xiao, and Peng Chang. 2022. Self-supervised Cross-modal Pretraining for Speech Emotion Recognition and Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5105–5114, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics\
21. Weidong Chen, Xiaofen Xing, Peihao Chen and Xiangmin Xu. Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition. 2023. https://doi.org/10.48550/arXiv.2307.10757
22. Hang Li, Yu Kang, Tianqiao Liu, Wenbiao Ding, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. arXiv preprint arXiv:2109.00181.

23. W. Chen, X. Xing, X. Xu, J. Yang and J. Pang, "Key-Sparse Transformer for Multimodal Speech Emotion Recognition," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 6897-6901, doi: 10.1109/ICASSP43922.2022.9746598.
24. Liu, F.; Shen, S.-Y.; Fu, Z.-W.; Wang, H.-Y.; Zhou, A.-M.; Qi, J.-Y. LGCCT: A Light Gated and Crossed Complementation Transformer for Multimodal Speech Emotion Recognition. *Entropy* **2022**, *24*, 1010. https://doi.org/10.3390/e24071010
25. Jiang Li, Xiaoping Wang, Guoqing Lv, Zhigang Zeng. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation, Neurocomputing, Volume 550, 2023. https://doi.org/10.1016/j.neucom.2023.126427.
26. J. Pan, W. Fang, Z. Zhang, B. Chen, Z. Zhang and S. Wang, "Multimodal Emotion Recognition based on Facial Expressions, Speech, and EEG," in *IEEE Open Journal of Engineering in Medicine and Biology*, doi: 10.1109/OJEMB.2023.3240280.
27. D. Peña, A. Aguilera, I. Dongo, J. Heredia and Y. Cardinale, "A Framework to Evaluate Fusion Methods for Multimodal Emotion Recognition," in *IEEE Access*, vol. 11, pp. 10218-10237, 2023, doi: 10.1109/ACCESS.2023.3240420.
28. Chen, S., Tang, J., Zhu, L. *et al.* A multi-stage dynamical fusion network for multimodal emotion recognition. *Cogn Neurodyn* **17**, 671–680 (2023). https://doi.org/10.1007/s11571-022-09851-w
29. Zhao, J.; Dong, W.; Shi, L.; Qiang, W.; Kuang, Z.; Xu, D.; An,T. Multimodal Feature Fusion Method for Unbalanced Sample Data in Social Network Public Opinion. Sensors 2022, 22, 5528. https://doi.org/10.3390/s22155528
30. Zhiguo Qu, Yang Li, Prayag Tiwari. QNMF: A quantum neural network based multimodal fusion system for intelligent diagnosis. Information Fusion, Volume 100, 2023, 101913, ISSN 1566-255. https://doi.org/10.1016/j.inffus.2023.101913.
31. Xu, Mingke et al. "Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion." *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)* (2020): 1058-1064.
32. Brian Mcfee, Colin Raffel, Dawen Liang, Daniel Ellis, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in Python in Science Conference, 2015.
33. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
34. Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in ICDM, 2017, pp. 1033–1038.
35. S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: a multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, January 2018.
36. A. A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246, Melbourne, Australia, January 2018.
37. Dai, W., Cahyawijaya, S., Liu, Z., and Fung, P. (2021). "Multimodal end-to-end sparse model for emotion recognition," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Mexico City: Association for Computational Linguistics), 5305–5316.
38. Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intelligent Systems, vol. 33, no. 6, pp. 17–25, 2018.
39. *Li Zuhe, Guo Qingbing, Feng Chengyao, Deng Lujuan, Zhang Qiuwen, Zhang Jianwei, Wang Fengqin and Sun Qian. Multimodal Sentiment Analysis Based on Interactive Transformer and Soft Mapping. 2022.* https://doi.org/10.1155/2022/6243347
40. *Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. In Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), pages 1–7, Seattle, USA. Association for Computational Linguistics.*
41. Bi W, Xie Y, Dong Z, Li H. Enterprise Strategic Management From the Perspective of Business Ecosystem Construction Based on Multimodal Emotion Recognition. Front Psychol. 2022 Mar 3;13:857891. doi: 10.3389/fpsyg.2022.857891. PMID: 35310264; PMCID: PMC8927019.
42. *Liu, F.; Shen, S.-Y.; Fu, Z.-W.; Wang, H.-Y.; Zhou, A.-M.; Qi, J.-Y. LGCCT: A Light Gated and Crossed Complementation Transformer for Multimodal Speech Emotion Recognition. Entropy 2022, 24, 1010. https://doi.org/10.3390/e24071010*

 doi:10.20944/preprints202404.1574.v1

43. *L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu and S. Ding, "Emotion Recognition With Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information," in IEEE MultiMedia, vol. 29, no. 2, pp. 94-103, 1 April-June 2022, doi: 10.1109/MMUL.2022.3161411.*
44. *Soumya Dutta and Sriram Ganapathy. HCAM -- Hierarchical Cross Attention Model for Multi-modal Emotion Recognition. 2023. https://doi.org/10.48550/arXiv.2304.06910*
45. *Sharma, A., Sharma, K. & Kumar, A. Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. Neural Comput & Applic (2022). https://doi.org/10.1007/s00521-022-06913-2*
46. *Z. Lian, B. Liu and J. Tao, "SMIN: Semi-supervised Multi-modal Interaction Network for Conversational Emotion Recognition," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2022.3141237.*