Attention-based Multimodal Speech Emotion Recognition

Group 26

Luo Li
19002505

Yanning Chen
18011621

Jingyun Su
22058774

Shiyang Xing
19010935

Abstract

Humans express their emotions in various ways. When communicating face-to-face, people's emotions are easily discernible through the nuances of their facial expressions, the tones of their voices, or unconscious body movements. Modern human-machine interaction is moving in an important direction towards better emotion recognition. According to previous research, models based on mixed data from multiple modalities often yield higher accuracy than unimodal models due to the possibility of complementary information. Therefore, we utilised pre-trained models for feature extraction and compared the models' performance on bimodal data (text & audio) using two mechanisms: self-attention and cross-attention. Our models are evaluated on the IEMOCAP dataset for a 7-class emotion classification task to assess accuracy. The results show that the self-attention mechanism consistently outperforms cross-attention, achieving a weighted accuracy of 52.2% and a weighted F1-score of 51.9%.

1 Introduction

Speech emotion recognition (SER) plays a pivotal role in human-machine interaction, enabling machines to understand and respond to human emotions effectively. In face-to-face communication, emotions are conveyed through various channels, such as facial expressions, vocal intonations, and body language. As a result, multi-modal approaches that consider multiple sources of information have been proposed to improve emotion recognition accuracy in human-machine interaction scenarios.

Recent research in this domain has focused on exploring attention mechanisms, such as self-attention and cross-attention, to enhance the processing and fusion of multi-modal data (Tang et al., 2022; Priyasad et al., 2020; Vaswani et al., 2017; Yu and Kim, 2020; Yoon et al., 2020; Feng et al., 2020; Rajan et al., 2022; Xu et al., 2019; Wang et al., 2021; Tsai et al., 2019; Sun et al., 2021; Siriwardhana et al., 2020; Santoso et al., 2022; Yoon et al., 2019). Furthermore, the advent of powerful pre-trained models (Devlin et al., 2018; Schneider et al., 2019; Heusser et al., 2019; Pepino et al., 2021; Lu et al., 2019; Liu et al., 2019; Baevski et al., 2020) like BERT for text and Wav2Vec for audio has revolutionised feature extraction and representation, leading to improved performance in emotion recognition tasks.

This paper aims to build upon the existing literature by investigating the performance of self-attention and cross-attention mechanisms in a bi-modal emotion recognition system that leverages pre-trained models for feature extraction. Specifically, we extend the work of "Is Cross-Attention Preferable to Self-Attention for Multi-Modal Emotion Recognition" (Rajan et al., 2022), which compares the efficacy of cross-attention and self-attention for multi-modal emotion recognition, by using traditional features. Additionally, we draw inspiration from "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings"(Pepino et al., 2021), which demonstrates the effectiveness of Wav2Vec 2.0 embeddings in SER, and incorporate the prototype of pre-trained models, namely BERT (Devlin et al., 2018) for text and Wav2Vec (Schneider et al., 2019) for audio, into our proposed approach.

Our primary contributions in this paper include:

1. The investigation of the utilization of cross-attention and self-attention models for uni-modal and bi-modal emotion recognition using text and audio features extracted by pre-trained models.
2. The evaluation of different model variations and their impact on performance, including the removal of the statistical pooling layer and

using BERT's CLS token as the feature of the text.

3. The findings indicating that the self-attention model is more effective for multimodal emotion recognition, and that the text modality contains a significant amount of valuable emotion information.

The remainder of this paper is organised as follows: Section 2 presents the related works of relevant research in multi-modal emotion recognition, attention mechanisms, and pre-trained models. Section 3 outlines our proposed methodology, detailing the feature fusion methods including self-attention and cross-attention mechanisms, as well as pre-trained models for multi-modal emotion recognition. Section 4 provides a description of the experimental setup and implementation details. Section 5 reports the results of our experiments and offers an ablation analysis of the findings. Finally, Section 6 concludes the paper and suggests directions for future research in this domain.

2 Related Works

2.1 The Development of Multimodal SER

Yoon et al. (2018) propose a multimodal emotion recognition system that combines information from audio and text modalities. This research underscores the importance of integrating multiple sources of information to achieve superior emotion recognition accuracy in SER tasks.

2.2 From Traditional Feature Fusion to Attention Mechanisms

In "Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Feature" (Pepino et al., 2020), the authors explore various fusion techniques for combining acoustic and text-based features in SER. They demonstrate that integrating these features leads to enhanced recognition accuracy compared to unimodal approaches, emphasizing the value of fusing information from different sources in SER tasks.

A number of recent studies (Yu and Kim, 2020; Yoon et al., 2020; Sun et al., 2021) have utilised attention mechanisms for feature fusion in SER, demonstrating their effectiveness in capturing salient features and improving performance. For instance, Sun et al. (2021) combines cross-attention and self-attention mechanisms to create a network for SER. The authors demonstrate that the proposed architecture effectively fuses information from both modalities, leading to improved performance in emotion recognition tasks. These studies collectively emphasize the value of attention mechanisms in handling feature fusion for SER tasks, harnessing the complementary information across different modalities to achieve better recognition accuracy.

In "Multimodal Emotion Recognition with Transformer-Based Self-Supervised Feature Fusion", Siriwardhana et al. (2020) present a pioneering approach that combines text, audio, and visual data for multimodal emotion recognition while employing attention mechanisms. This innovative study is the first to integrate all three modalities using a transformer-based self-supervised learning approach, resulting in a more effective feature fusion and enhanced performance in SER tasks.

2.3 Feature Extraction Based on Pre-trained Models

Pepino et al. (2021) in their paper "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings" demonstrate the effectiveness of using pre-trained models, specifically Wav2Vec 2.0 embeddings, for emotion recognition from speech. The authors show that Wav2Vec 2.0 embeddings provide valuable information for SER tasks, outperforming traditional feature extraction methods. This study serves as a strong motivation for our proposed approach, which also employs pre-trained models, namely RoBERTa for text modality (Liu et al., 2019) and Wav2vec 2.0 for audio modality (Baevski et al., 2020), for feature extraction in bi-modal emotion recognition tasks.

Our proposed approach for multi-modal emotion recognition builds upon the findings of the above-mentioned paper and "Is Cross-Attention Preferable to Self-Attention for Multi-Modal Emotion Recognition"(Rajan et al., 2022; Pepino et al., 2021). We intend to explore the use of pre-trained models for feature extraction and compare the performance of self-attention and cross-attention mechanisms on bi-modality data for text and audio. This integrated approach seeks to advance the understanding of multi-modal emotion recognition by identifying the most effective attention mechanism for different modalities while leveraging the power of pre-trained models for feature extraction.

3 Methodology

In this section, we present our attention-based feature fusion method for multi-modal speech emotion recognition. We begin by describing the feature extraction process using pre-trained models, followed by an explanation of the attention mechanism for capturing inner- and inter-modality information. Lastly, we discuss the statistical pooling layer used for feature fusion.

3.1 Feature Extraction with Pre-trained Models

Our first step was to extract high-level features from the raw text and audio data using pre-trained models. Training a large model from scratch is computationally expensive; therefore, we used publicly available pre-trained models. RoBERTa (Liu et al., 2019) and Wav2vec 2.0 (Baevski et al., 2020) were employed for text and audio feature extraction, respectively. Both RoBERTa and Wav2Vec 2.0 are self-supervised models; thus, the features extracted using these models are referred to as self-supervised embeddings. We downloaded checkpoints of the models through the Hugging Face's Transformers library, which provides an easy-to-use interface for preprocessing raw data and extracting features using pre-trained models. No fine-tuning was performed due to limited computational resources.

3.1.1 Text Feature Extractor: RoBERTa

The robustly Optimized BERT approach (RoBERTa) is an improvement over the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) model. BERT is a state-of-the-art pre-trained language model based on the Transformer (Vaswani et al., 2017) architecture. BERT uses two pre-training objectives: masked language modeling (MLM) and next sentence prediction (NSP), to learn contextual representations of words using a large corpus of text. RoBERTa improves the pre-training process of BERT by using only the MLM objective with a dynamic masking technique. It has 24 layers of transformer encoders and consists of 355M parameters. RoBERTa takes tokenized text data (IDs) as input, and the last hidden states are commonly used as features to represent the contextualized information of the input tokens. We used the default embeddings size of 1024.

To preprocess the raw text data, we used RoBERTa's default tokenizer from the Transformers library. The tokenizer takes sentences as input and first performs text normalization by lowercasing all the characters and adding spaces around punctuations. It then carries out Byte-Pair Encoding (BPE) to encode words as a sequence of subword units, maximizing the coverage of the constructed vocabulary. Next, it truncates and pads the input sequences to a fixed length. To reduce computation time, we set the maximum sequence length to 128 instead of the default value of 512. Finally, the tokenizer adds special tokens to the sequences and maps each token into an integer ID that can be processed by the RoBERTa model.

3.1.2 Audio Feature Extractor: Wav2Vec 2.0

Wav2Vec 2.0 is a state-of-the-art speech recognition model pre-trained on 960 hours of audio recordings of people speaking in various languages. The Wav2Vec (Schneider et al., 2019) model uses a self-supervised pre-training approach called contrastive predictive coding (CPC) to capture both local and global structures of audio signals. Wav2Vec 2.0 improves upon Wav2Vec with a more advanced architecture called Time-Contrastive Networks (TCNs). TCNs consist of a stack of convolutional neural network (CNN) layers designed to effectively capture long-term dependencies in input audio signals. The output of the final CNN layer (last hidden state) is a sequence of fixed-length embeddings that capture the acoustic and phonetic properties of the input audio signals. We use this output as our audio feature with an embedding size of 768.

Before extracting features using the model, we preprocessed the raw audio data using Wav2Vec 2.0's default processor from the Transformers library. The processor takes WAV files as inputs and first resamples the audio signals to 16kHz, which is the default sampling rate of the Wav2Vec 2.0 model. Then, the processor pads the audio signals and splits them into overlapping windows of 960 samples with a hop size of 320 samples (20ms). By doing this, each segment of the input signals is represented in multiple windows, helping the model capture fine-grained temporal patterns in the audio signals. Finally, the processor normalizes the input signals using statistics computed on a large dataset of speech. This step effectively reduces the impact of speaker-specific and channel-specific characteristics of the audio signals.

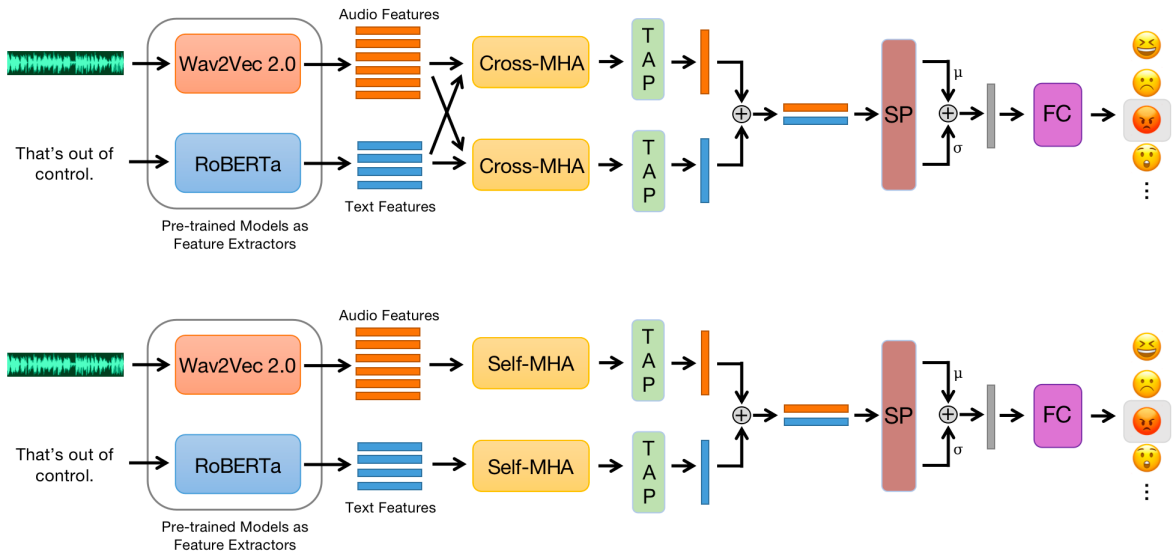

Figure 1: The Structure of the Proposed Models

3.2 Feature Fusion Methods

We employed the attention mechanism (Vaswani et al., 2017) to capture emotional information within the extracted multi-modal (text & audio) features. More specifically, self- and cross-attention were applied to exploit inner-modality and inter-modality information. Between the attention layer and the final prediction layer, we applied statistical pooling to combine the exploited features.

3.2.1 Inner-Modality: Self-Attention

Self-attention was applied to capture complex relationships between different parts of the input text and audio embedding sequences. In emotion recognition, the emotional content of a sentence or audio clip may be spread across multiple words or segments. Self-attention can weigh the importance of each element in a sequence based on its relevance to the overall emotional tone, selectively attending to parts that are more informative for making predictions. For example, words like "happy" or "sad" may be given higher attention weights than more neutral words. For audio signals, self-attention highlights features corresponding to changes in pitch, tone, or volume, which may be more indicative of emotional expression. Therefore, we can effectively capture intra-modality emotional information by applying self-attention.

To go a step further, we used multi-head self-attention, which allows us to attend to multiple aspects of the input sequences in parallel. In multi-head self-attention, the input sequences are first transformed into several distinct representations through linear projections. Self-attention is then applied to these representations separately. The outputs are concatenated and passed through another linear projection layer to generate the final output.

We applied multi-head self-attention separately to the embedding sequences of the text and audio data using the nn.MultiheadAttention (MHA) module, where the self-attention computation is depicted in Equation (1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

The Query (Q), Key (K), and Value (V) are the three inputs to the MHA module. In the case of self-attention, all three inputs are the same (either the text sequence or the audio sequence).

3.2.2 Inter-Modality: Cross-Attention

Cross-attention serves the purpose of sharing and fusing emotionally related information across different modalities. It encodes information from the representations of one modality into the representations of another modality. Given a text-audio pair, cross-attention can align the elements in the text sequence with the corresponding elements in the audio signal at the level of semantic content. Through cross-attention, the model can selectively attend to and exploit parts of the input that are emotionally prominent in both the text and audio sequence. For example, the model can learn to associate specific emotional stimuli found in the text with particular emotions expressed in speech. Furthermore, cross-

attention prevents the problem of modality bias by balancing the contributions of the text and audio modalities. By attending to both modalities simultaneously, the model can also learn to recognize emotions based on the underlying semantic content rather than relying on modality-specific features. This improves the generalization of a multimodal emotion recognition model.

Similar to self-attention, we used the nn.MultiheadAttention (MHA) module to perform multi-head cross-attention. The computation of cross-attention is the same as shown in Equation (1), except it creates the Query (Q) from the sequence of the major modality and the Key (K) and Value (V) from the minor modality. In our case, we performed two different multi-head cross-attention, each using text and audio as the major modality. The attention mechanism requires the three inputs (QKV) to have the same embedding size. Therefore, we reduced the embedding size of the text features from 1024 to 768 before cross-attention.

3.2.3 Fusion and Classification Layers

With two different modalities, we always end up with two output vectors from the self- or cross-attention module. Therefore, we applied temporal average pooling, concatenation, statistical pooling, and linear projection to fuse the two output features before the prediction layer.

Temporal average pooling is average pooling applied over the temporal dimension, which outputs a single embedding vector as a global representation for an input sequence. We stacked the two temporally averaged global representations along the temporal dimension. Then, we applied statistical pooling over the stacked vector to produce the mean and standard deviation vectors. Next, we concatenated these two vectors along the feature dimension. Finally, the concatenated vector was passed through a fully connected layer to reduce the dimension before making predictions as shown in Figure 1.

We defined the emotion recognition task as classifying the input text-audio pair to a specific emotion class. Therefore, the final layer of our model is a fully connected layer with the same number of output features as the number of emotion classes.

4 Experiments

4.1 Dataset and Problem Formulation

To obtain comparable results, we utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008) dataset, which is among the most widely used datasets for multimodal SER (Yoon et al., 2018, 2020, 2019; Siriwardhana et al., 2020). It comprises approximately 12 hours of audio-visual emotional interactions, along with text transcriptions. The dataset is gender-balanced, containing improvised and scripted dialogues between 5 male and 5 female actors. Each utterance within the dataset was annotated by three distinct annotators using categorical labels (e.g., happy, neutral, and sad), with the majority vote determining the final emotion labels. In line with previous studies, we removed classes with fewer than 100 utterances (fear, disgust, and other) and formulated the problem as a 7-class classification task. The final dataset consists of 7,487 utterances with the following distribution: 1,103 angry, 1,041 excited, 595 happy, 1,084 sad, 1,849 frustrated, 108 surprised, and 1,708 neutral.

We employed 5-fold cross-validation to evaluate the model's performance. To ensure consistent comparisons with prior work, we followed (Yoon et al., 2020) to shuffle and split the data in each fold into training (80%), validation (10%), and test (10%) sets. For each fold, we utilized the validation set to determine the best model and evaluated it using four classification metrics: weighted accuracy (WA), unweighted accuracy (UWA), weighted F1-score (WF1), and unweighted F1-score (UWF1). This approach allowed us to thoroughly assess the performance of our model across different aspects of classification while maintaining consistency with established methodologies in the field. By adhering to these evaluation methods, we ensured the reliability and comparability of our results with those of previous studies that have employed the IEMOCAP dataset for multimodal SER.

4.2 Implementation Details

We utilized only the text and audio modalities of the dataset, with each example containing a brief audio clip stored as a WAV file and the corresponding text transcription. We loaded the audio and text data separately and employed pre-trained models to extract features. For the audio modality, the raw audio data were first preprocessed using the default processor of the Wav2Vec 2.0 model, and

features with a sequence length of 499 (31ms) and an embedding size of 768 were extracted. For the text data, we first truncated and padded all text sequences to a length of 128. Then, we used the default tokenizer of the RoBERTa model to perform preprocessing. Finally, features with a sequence length of 129 (including an additional <CLS> token from the model) and an embedding size of 1024 were extracted.

The entire framework (feature extraction, model, training, and inference) was implemented in PyTorch. After extensive hyperparameter tuning, we finalized our choices of hyperparameters, as shown in Table 1. We trained the models on each fold of data for a maximum of 50 epochs. To avoid being trapped in local optima during training, we adopted a strategy called learning rate scheduling, where we reduced the learning rate by a factor of 0.1 when there was no improvement in validation loss for 5 consecutive epochs. Moreover, we used early stopping to prevent the models from overfitting when the validation loss did not decrease for 10 consecutive epochs. These two training strategies proved to be effective through experimentation. During training, we evaluated the models using the four metrics after each epoch on both the validation and test sets. We selected and saved the best-performing model based on the weighted accuracy score computed on the validation set.

Number of Attention Heads	6
Dropout-out Rate	0.1
Batch Size	32
Learning Rate	0.0001

Table 1: Hyperparameters Used in Training

4.3 Comparison Against Baseline Models

To assess the effectiveness of the proposed self- and cross-attention models with self-supervised embeddings, we implemented and evaluated the following baseline models on the IEMOCAP dataset for comparison:

- Random Guesser: A random model that predicts labels with equal probabilities for each class
- Text Model: A single modality model based solely on the text data and self-attention mechanism
- Audio Model: A single modality model based solely on the audio data and self-attention mechanism
- MDRE (Yoon et al., 2018): A bimodal (text and audio) model featuring a dual recurrent neural network encoder and a simple concatenation fusion layer

5 Results and Discussion

In this section, we present the results of our study on unimodal and bimodal emotion recognition using cross-attention and self-attention models with text and audio features extracted by pre-trained models. We employed the IEMOCAP dataset, which comprises approximately 12 hours of emotional interactions in both scripted and unscripted settings (Busso et al., 2008).

5.1 Experiment Results and Error Analysis

Table 2 presents the performance of the self-attention and cross-attention models for the 7-class unimodal and bimodal emotion recognition task, alongside four baseline models. The findings indicate that, by using text and audio features, the self-attention model outperforms the cross-attention model, which is consistent with previous studies (Rajan et al., 2022).

	WA	UWA	WF1	UWF1
Random	0.138± 0.002	0.135± 0.004	0.151± 0.002	0.126± 0.002
Audio	0.398± 0.018	0.342± 0.006	0.367± 0.016	0.335± 0.007
Text	0.506± 0.027	**0.492± 0.029**	0.498± 0.025	0.475± 0.022
MDRE(2018)	0.491± 0.039	0.466± 0.056	0.482± 0.035	0.470± 0.041
Self-MHA	**0.522± 0.014**	0.486± 0.020	**0.519± 0.016**	**0.488± 0.014**
Cross-MHA	0.509± 0.009	0.477± 0.027	0.509± 0.009	0.478± 0.024

Table 2: Performance of Self- and Cross-Attention Models for the 7-class Unimodal/Bimodal SER Task

The results also show that our self-attention model surpasses all four baseline models in terms of WA, WF1, and UWF1. However, the unimodal text-only model achieves the highest UWA, suggesting that the text modality indeed contains loads of valuable emotion information. It is also evident that the text modality outperforms the audio modality. This observation can be attributed to the complexity and dynamic nature of speech signals,

which are influenced by various factors such as background noise, speaker accent, and speech rate, among others. These factors render the audio signal more challenging to analyze and classify accurately compared to the text modality, which offers a more controlled and structured input. Moreover, text data is less susceptible to variations caused by the speaker's emotions or other external factors, unlike the audio signal, which is directly influenced by the speaker's emotions and tone of voice.

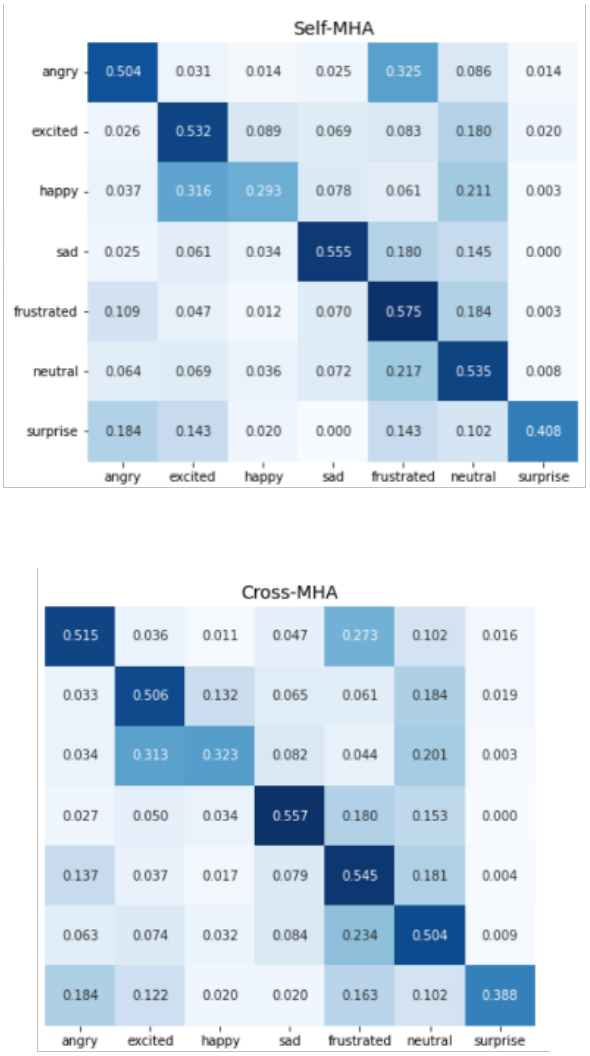

Figure 2: Confusion Matrices for the Self- and Cross-Attention Models

Figure 2 displays the confusion matrices for the self- and cross-attention models in the 7-class bimodal emotion recognition tasks. In the bimodal text + audio model, we observed frequent confusion between the classes "angry" and "frustrated", as well as between "happy" and "excited". These findings align with observations from previous studies (Rajan et al., 2022).

5.2 Ablation Study

We further evaluated some variations of the proposed model. First, we removed the statistical pooling (SP) layer from both the self- and cross-attention models to assess its significance. The results are presented as Self-noSP and Cross-noSP in Table 4. Even after removing the SP layer, the self-attention model performed better than the cross-attention model, but both models experienced decreased performance without it. The SP layer plays a crucial role in reducing feature space dimensionality and extracting relevant information from the input data. Without it, the model's ability to extract relevant information is hindered, resulting in a significant impact on its performance.

	WA	UWA	WF1	UWF1
Self-noSP	0.514± 0.022	0.464± 0.042	0.507± 0.027	0.473± 0.046
Self-CLS	0.505± 0.008	0.461± 0.019	0.504± 0.009	0.468± 0.021
Self-MHA	**0.522± 0.014**	**0.486± 0.020**	**0.519± 0.016**	**0.488± 0.014**
Cross-noSP	0.498± 0.005	0.442± 0.021	0.493± 0.005	0.448± 0.024
Cross-CLS	**0.518± 0.011**	0.451± 0.020	**0.512± 0.010**	0.456± 0.022
Cross-MHA	0.509± 0.009	**0.477± 0.027**	0.509± 0.009	**0.478± 0.024**

Table 3: Results after Variations of the Proposed Model

We also implemented a variation of the proposed model by replacing each input text sequence with only the embedding of the first element. This element corresponds to the special <CLS> token produced by the RoBERTa model, which aggregates the information embedded in the entire sequence. We aimed to examine the effectiveness of using the entire text sequence as input compared to using just the <CLS> token. The results are reported as Self-CLS and Cross-CLS in Table 3 as well. For the self-attention model, using BERT's <CLS> token as the feature of the texts reduces performance. However, for the cross-attention model, using BERT's <CLS> token increases the model's performance in terms of WA and WF1, but decreases its performance in terms of UWA and UWF1. This suggests that the cross-attention model using BERT's <CLS> token performs better in the majority classes but struggles with the minority classes.

6 Conclusion and Future work

We evaluated the performance of self-attention and cross-attention models with text and audio features extracted by pre-trained models on the IEMOCAP dataset for unimodal and bimodal 7-class classification. Results indicate that self-attention models outperform cross-attention models, leading us to conclude that, based on the dataset and architecture employed in our study, self-attention is more effective for multimodal emotion recognition.

Having said that, our study has a few limitations. Despite being a popular and varied dataset, the IEMOCAP dataset's limited duration of approximately 12 hours may not provide an exhaustive portrayal of emotional states across various speakers and scenarios. Additionally, the lack of computing power may have resulted in suboptimal hyperparameter tuning, which could have constrained the models' performance. For future research, we suggest exploring the integration of video components in multimodal models and improving the model's ability to distinguish between confused emotion classes.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Han Feng, Sei Ueno, and Tatsuya Kawahara. 2020. End-to-end speech emotion recognition combined with acoustic-to-word asr model. In *INTERSPEECH*, pages 501–505.

Verena Heusser, Niklas Freymuth, Stefan Constantin, and Alex Waibel. 2019. Bimodal speech emotion recognition using pre-trained language models. *arXiv preprint arXiv:1912.02610*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustín Gravano. 2020. 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)*, pages 6484–6488. IEEE.

Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2020. Attention driven fusion for multi-modal emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3227–3231. IEEE.

Vandana Rajan, Alessio Brutti, and Andrea Cavallaro. 2022. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4693–4697. IEEE.

Jennifer Santoso, Takeshi Yamada, Kenkichi Ishizuka, Taiichi Hashimoto, and Shoji Makino. 2022. Speech emotion recognition based on self-attention weight correction for acoustic and text features. *IEEE Access*, 10:115732–115743.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. 2020. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285.

Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279. IEEE.

Yuwu Tang, Ying Hu, Liang He, and Hao Huang. 2022. A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition. *Speech Communication*, 143:21–32.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yuhua Wang, Guang Shen, Yuezhu Xu, Jiahang Li, and Zhengdao Zhao. 2021. Learning mutual correlation in multimodal transformer for speech emotion recognition. In *Interspeech*, pages 4518–4522.

Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. 2019. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826. IEEE.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE.

Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung. 2020. Attentive modality hopping mechanism for speech emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3362–3366. IEEE.

Yeonguk Yu and Yoon-Joong Kim. 2020. Attention-lstm-attention model for speech emotion recognition and analysis of iemocap database. *Electronics*, 9(5):713.