Multi-modal Speech Emotion Recognition: Improving Accuracy Through Fusion of VGGish and BERT Features with Multi-head Attention

Phuong-Nam Tran[1], Thuy-Duong Thi Vu[1], Duc Ngoc Minh Dang[1(✉)], Nhat Truong Pham[2], and Anh-Khoa Tran[3]

[1] Computing Fundamental Department, FPT University, Ho Chi Minh City, Vietnam
namtpse150004@fpt.edu.vn, {duongvtt9,ducdnm2}@fe.edu.vn

[2] Department of Integrative Biotechnology, Sungkyunkwan University, Suwon, Republic of Korea
truongpham96@skku.edu

[3] Modeling Evolutionary Algorithms Simulation and Artificial Intelligence, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam
trananhkhoa@tdtu.edu.vn

Abstract. Recent research has shown that multi-modal learning is a successful method for enhancing classification performance by mixing several forms of input, notably in speech-emotion recognition (SER) tasks. However, the difference between the modalities may affect SER performance. To overcome this problem, a novel approach for multi-modal SER called 3M-SER is proposed in this paper. The 3M-SER leverages multi-head attention to fuse information from multiple feature embeddings, including audio and text features. The 3M-SER approach is based on the SERVER approach but includes an additional fusion module that improves the integration of text and audio features, leading to improved classification performance. To further enhance the correlation between the modalities, a LayerNorm is applied to audio features prior to fusion. Our approach achieved an unweighted accuracy (UA) and weighted accuracy (WA) of 79.96% and 80.66%, respectively, on the IEMOCAP benchmark dataset. This indicates that the proposed approach is better than SERVER and recent methods with similar approaches. In addition, it highlights the effectiveness of incorporating an extra fusion module in multi-modal learning.

Keywords: 3M-SER · Multi-modal analysis · Speech Emotion Recognition · Multi-head Attention · Multi-feature Embeddings

1 Introduction

Speech emotion recognition (SER) is a rapidly growing field of research that focuses on the development of algorithms and systems capable of automati-

N.-S. Vo and H.-A. Tran (Eds.): INISCOM 2023, LNICST 531, pp. 148–158, 2023.
https://doi.org/10.1007/978-3-031-47359-3_11

cally detecting, analyzing, and interpreting emotions conveyed through speech. The capacity to detect emotional states from speech offers a broad scope of potential benefits and useful applications, encompassing healthcare, education, entertainment, and human-computer interaction. Traditionally SER focused on analyzing only the audio component of speech to detect and interpret emotions. Feature extraction from speech signals has been the mainstay of SER, where various acoustic features such as pitch, loudness, spectral [1], spectrograms [2] are extracted from the speech signal and used to recognize emotions. Deep learning (DL) models are often used as end-to-end models that can learn to extract relevant features from the audio and directly predict the emotion label. Pham *et al.* [2] developed a DL approach for speech emotion recognition that involved modifying an existing DL model and incorporating a novel loss function specifically designed for recognizing emotional states from speech signals. Bao *et al.* [3] took a different approach by proposing a DL-based model as a data augmentation method to improve the performance of SER systems. Specifically, they developed a generative adversarial network with emotional style transfer that can generate emotional data samples, thereby increasing the amount and diversity of training data and enhancing the performance of the SER system.

As stated in [4], most previous methods only focus on a single modality, either audio or text input, to recognize emotional states. In recent years, researchers have recognized that using multiple modalities, such as audio combined with text, can improve the accuracy of emotion recognition. A multi-modal SER system can capture complementary information from different modalities and combine them to provide a more complete understanding of the emotional state of the speaker. SERVER [5] is a recent example of a multi-modal SER system that combines information from both audio and text modalities. In the SERVER [5] system, audio features are extracted from the speech signal using Mel frequency cepstral coefficients (MFCCs) and are fed into the pre-trained VGGish [6]. To obtain the embedding of text input, the pre-trained BERT [7] model is utilized for text embedding. The embeddings are then combined using concatenation and fed to a fully connected layer for emotional state classification.

The SERVER [5] has shown an increase in performance in the system by using multi-modal. However, the difference between the emotions represented in text and in audio may affect the performance of the model. SERVER [5] only uses the concatenation of the feature of text and audio which may create a huge impact on the classifier head if the model relies on the text feature too much to classify the emotion rather than rely on the audio feature. For instance, if the text contains the word "cry", but the audio does not reflect sadness, the system may bias to recognize the emotions of the audio as sadness rather than neutral or the label emotion.

To overcome the aforementioned problem, we propose a novel method that fusion the feature embeddings of text and audio through an attention [8] mechanism. Our approach uses multi-head attention, which is a type of attention mechanism that allows the model to attend to different parts of the input simultaneously and learn which parts are most relevant for predicting the emotion.

Experimental results of our method IEMOCAP [9] have shown improved performance by integrating the attention fusion module into the SERVER model. Our method achieved a new highest score on IEMOCAP [9] with a UA of 79.96% and a WA of 80.66%, respectively.

The structure of this paper is organized as follows. Section 1 provides an introduction to the study. Section 2 presents a summary of the literature review and related studies. In Sect. 3, we elaborate on the motivation behind the proposed methodology and provide a detailed explanation of the methodology itself. The employed dataset, experimental setup, preliminary results, and discussions are presented in Sect. 4. Finally, Sect. 5 concludes the study and outlines potential future work.

2 Related Work

The emotions in human speech are complicated and are not easy to recognize even if the listener is a human. Numerous research efforts have focused on analyzing speech features and accurately classifying them to enhance speech emotion recognition.

Particularly, Google researchers [6] have recently proposed a method that applies CNN architectures to convert audio into a latent space dimension known as audio embedding. A feature extraction model, namely VGGish is applied to the log Mel-Spectrogram, which is transformed from the audio input, to retrieve an audio embedding. The design of VGGish is influenced by the popular VGG networks used in image classification and can function either as a feature extractor or as a downstream classification model. VGGish has shown a high performance on a large-scale audio dataset (AudioSet) [10].

BERT (Bidirectional Encoder Representations from Transformers) [7] is a powerful language model that has been widely used in natural language processing (NLP) tasks such as sentiment analysis and emotion detection. In addition to its applications in sentiment analysis and emotion detection, BERT has also been combined with other modalities such as text and audio to create multimodal models. Multimodal models combine different types of data to gain a more comprehensive understanding of the input and improve the accuracy of the models.

For instance, Lee *et al.* [11] took multimodal modeling further by combining BERT with heterogeneous features extracted from multi-modal inputs, including textual, visual, and acoustic information, to enhance the ability of the BERT model to recognize emotional states. By incorporating these additional features, the model achieved improved accuracy and outperformed previous state-of-the-art models on several benchmark datasets. This demonstrates the potential of combining BERT with other modalities to create more effective multimodal models for emotion recognition.

Recent studies have investigated the use of both audio and text inputs in DL models for SER. Lee *et al.* [12] proposed a cross-attention network that aligns audio and text embeddings for multimodal SER. By employing bidirectional

LSTM, the audio embedding was created by processing the Mel-frequency cepstral coefficients (MFCCs) that were extracted from the audio waveform. Similarly, the text embedding was generated by using bidirectional LSTM to process the extracted GloVe embeddings. These embeddings were then fed into the cross-attention network for final emotion classification. Yoon *et al.* [13] proposed an audio recurrent encoder and a text recurrent encoder for multi-modal SER based on MFCC features and word embedding that was extracted from audio and text, respectively. Pham *et al.* [5] proposed a multi-modal speech emotion recognition using BERT and VGGish (SERVER). SERVER is very competitive and better than most of the latest and state-of-the-art methods using multi-modal analysis for SER. It achieves 63.10% unweighted accuracy and 63.00% weighted accuracy on the IEMOCAP [9] dataset.

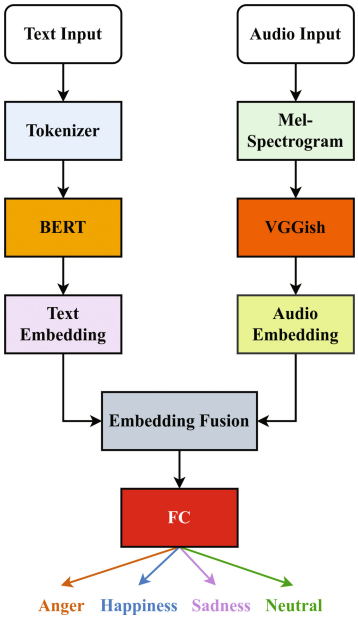


Fig. 1. The flowchart of the SERVER [5].

3 Methodology

3M-SER is an improved version of SERVER [5] by extending the architecture with the addition of multi-head attention over the fusion module and the text module. As shown in Fig. 1, the audio input is transformed to log Mel-Spectrogram of 96 × 64 bins and fed into the pre-train VGGish [6] model to extract audio features. The text features are extracted using the pre-trained

BERT [7] model. Both audio and text are transformed to the latent spaces dimension which can represent their features in a fixed size. These features can be called text-embedding $v_t \in \mathcal{R}^{d_t}$ and audio-embedding $v_a \in \mathcal{R}^{d_a}$ The different sizes of the latent spaces in each feature require designing a module to combine these features. SERVER [5] proposes to transform the text-embedding v_t spaces to the v_a by simply adding a linear layer after the output of the BERT [7] model. After obtaining the text and audio embeddings, they are combined through concatenation to create a fusion feature, which is employed in the classification of emotional states. This method shows an improvement in the performance model, however, we can further improve this result by designing an attention fusion module rather than the simple linear. The details of our method are shown in Fig. 2.

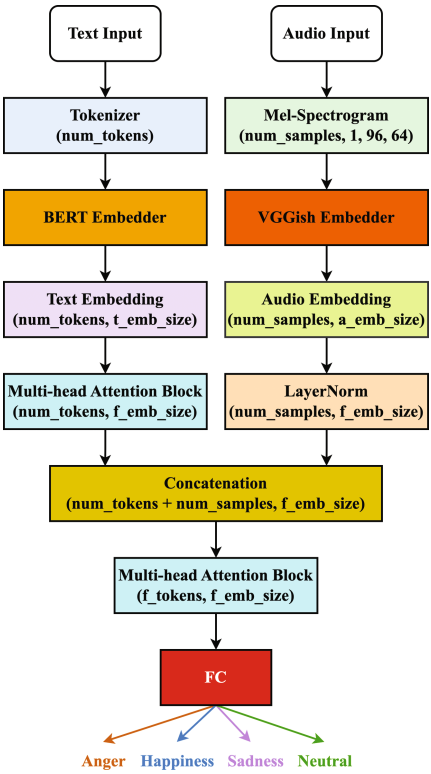


Fig. 2. The flowchart of the proposed 3M-SER.

A multi-head attention [8] block is applied to the text embeddings to figure out which feature is useful to the classification process. This block also converts the text embeddings from v_t spaces to the v_a by using a linear and LayerNorm [14] after the multi-head attention.

Although the dimension of the audio embeddings stays the same, the combination of the audio embeddings with the text embeddings feature creates a significant disparity in the feature values of v_t and v_a that are derived from their corresponding pre-trained feature extraction model. The BERT model tries to transform a text to $v_t \in \mathcal{R}^{d_t}$ which R^{d_t} mostly in $[-2.0, 2.0]$ while the VGGish [6] transforms the log Mel-Spectrogram to $v_a \in \mathcal{R}^{d_a}$ which R^{d_a} mostly in $[0.0, 255.0]$. This imbalance may lead to audio-based judgment much more than the text feature and may overwhelm the text feature if we simply concatenate two features without performing any linear layer. To make the fairness between each feature, we apply the LayerNorm [14] to both the audio embedding and text embedding. LayerNorm [14] will make both the v_a and v_t have a closer value which makes the 3M-SER slightly study better.

In the concatenate layer, rather than the fusion of two features based on the dimension space like SERVER [5], our 3M-SER fuses two features based on the tokens and samples axis of text embeddings and audio embedding, respectively. This technique will help 3M-SER view the entire sentence and Mel-Spectrogram samples to assess the emotion in the sound through another multi-head attention [8] block.

4 Preliminary Results and Discussion

4.1 Dataset

In our experiments, we use the same dataset which is used to evaluate the SERVER [5] and other single-modal approaches. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, as described in [9], is a multimodal and multi-speaker database containing the acted audiovisual data. The dataset comprises around 12 h of content, including video, speech, motion capture of facial expressions, and text transcriptions. To validate the effectiveness of the 3M-SER along with SERVER [5], the same text, audio, and number of samples are investigated in this study. The same dataset contains only four major classes such as anger (1,103 samples), happiness (1,635 samples), sadness (1,084 samples), and neutral (1,708 samples). The distribution of each class used in this study is shown in Fig. 3.

4.2 Experimental Setup

The 3M-SER is implemented using the PyTorch [15] DL framework and trained on a Linux machine (Debian Bookworm) with Intel(R) Core(TM) i9-12900K, 64 GB RAM, and 1 Nvidia GeForce RTX 3090 Graphics Card. We follow the settings in SERVER [5] to set our optimizer, the learning rate decay, and the dataset. The multi-head attention component, which is composed of 8 heads, is succeeded by a linear layer and a LayerNorm layer [14]. The multi-head attention block after the text embedding has the linear layer with 768 in-feature and 128 out-feature to convert text latent space dimension from $v_t \in \mathcal{R}^{768}$ to $v_t \in \mathcal{R}^{128}$

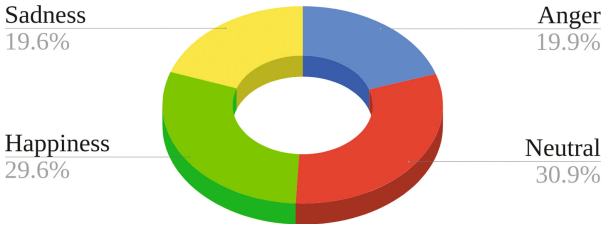


Fig. 3. The distribution of the employed emotions in the IEMOCAP dataset used for training 3M-SER.

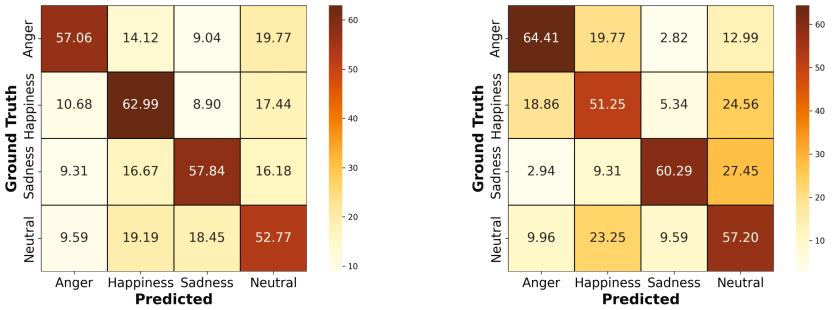
for the concatenating process in fusion module. The same multi-head attention block is applied after the fusion module, however, the in-feature and out-feature in the linear layer are set to 128. Two linear layers with 64 and 4 units in the fully connected (FC) layers are added after embedding fusion. The 4 units are the classification head with softmax activation which is used to calculate the category cross-entropy loss.

4.3 Results and Discussion

Figures 4a, 4b, and 5 display the confusion matrices of the models using only text embedding, only audio embedding, and both embeddings, respectively, which were reported in SERVER [5]. Figures 6a and 6b display our 3M-SER confusion matrices which show the impact of the fusion module on the SERVER [5] model and the effect of LayerNorm [14]. It is observed that adding an attention mechanism can improve the performance of the model through the meaning of text and audio. The accuracy of 3M-SER helps improve the model recognition of "anger", "happiness", and "sadness", however, the model is still confused about "neutral" emotion and seems to fail to recognize it.

Table 1. Performance comparison of the different multi-modal SER methods on the IEMOCAP dataset.

Method	Accuracy (%)	
	UA	WA
Ref. [16]	51.70	–
Ref. [17]	56.00	61.20
Ref. [12]	48.70	57.90
SERVER [5]	63.00	63.10
3M-SER	**75.35**	**76.81**
3M-SER with LayerNorm	**79.96**	**80.66**



(a) Text Only [5]. (b) Audio Only [5].

Fig. 4. The confusion matrix of the SERVER [5] using single data.

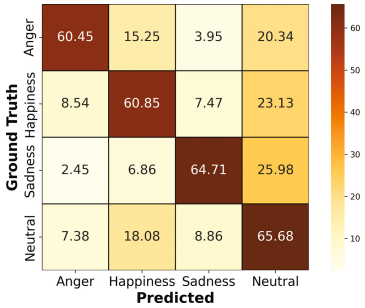
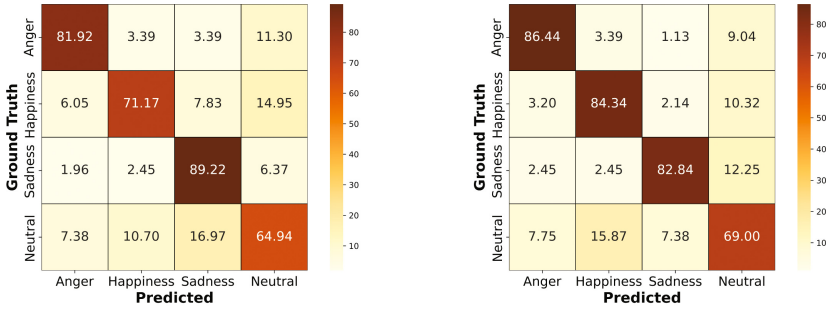


Fig. 5. The confusion matrix of the SERVER [5] using both text and audio embeddings.



(a) Without LayerNorm. (b) With LayerNorm.

Fig. 6. The confusion matrix of the proposed 3M-SER using both text and audio embeddings with attention fusion module.

As shown in Table 1, the proposed method has the best accuracy in terms of unweighted accuracy and weighted accuracy. Table 1 demonstrates that the proposed method outperforms other methods in both unweighted accuracy and weighted accuracy. Compared to notable references such as [12,16,17] and [5], the proposed method achieves the improvements of 10.88%, 15.18%, 18.18% and 3.88%, respectively, in terms of UA. Similarly, the proposed method surpasses [12,17], and [5] by 6.01%, 9.31% and 4.11%, respectively, in terms of WA. Moreover, Table 2 presents a comparison of different methods in terms of model complexity and performance, including the number of parameters (Params), FLoating point OPerations (FLOPs), and accuracy. Based on Tables 1 and 2, although the proposed 3M-SER method has the highest complexity, its performance shows a significant improvement.

Table 2. Comparison of model complexity and performance for different multi-modal SER methods on the IEMOCAP dataset.

Method	Params	FLOPs	Accuracy (%)	
			UA	WA
Text Only	109M	0.00683G	57.67	57.77
Audio Only	72M	1.73G	57.56	58.41
SERVER [5]	181M	1.74 G	63.00	63.10
Ours	**203M**	**1.74 G**	**79.96**	**80.66**

5 Conclusion and Future Work

In this paper, a novel multi-head attention fusion mechanism has been proposed to improve the accuracy of multi-modal speech emotion recognition. Learning from the text embeddings and audio embeddings which are extracted from the BERT and VGGish respectively using the attention mechanism helps model learning better on understanding the meaning of text along with the audio. The experimental results have shown that our proposed method improved the performance of the previous multi-modal. The proposed method achieves the highest UA of 79.96% and WA of 80.66%, respectively, on the IEMOCAP dataset.

In future work, we plan to investigate multi-lingual and multi-task learning approaches to further extend this study. This extension aims to enhance the generalization and robustness of multi-modal speech emotion recognition (SER) systems. Additionally, the exploration of new architectures, along with the utilization of data collection and augmentation techniques such as hybrid data augmentation (HDA) [18], will be considered. These efforts aim to further improve the model performance, reduce the bias among features, and generate additional data for the task of speech emotion recognition. Moreover, in this preliminary study, we have observed the significance of correlation or similarity

between different modalities. Therefore, in the extended version of this study, we will place significant emphasis on exploring the similarity/correlation between text embeddings and audio embeddings to improve the overall performance of the multi-modal SER system.

References

1. Liu, D., Chen, L., Wang, Z., Diao, G.: Speech expression multimodal emotion recognition based on deep belief network. J. Grid Comput. **19**(2), 22 (2021)
2. Pham, N.T., Dang, D.N.M., Nguyen, S.D.: A method upon deep learning for speech emotion recognition. J. Adv. Eng. Comput. **4**(4), 273–285 (2020)
3. Bao, F., Neumann, M., Vu, N.T.: Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In: Kubin, G., Kacic, Z. (eds.) Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15–19 September 2019, pp. 2828–2832. ISCA (2019)
4. Pham, N.T., et al.: Speech emotion recognition: a brief review of multi-modal multi-task learning approaches. In: AETA 2022-Recent Advances in Electrical Engineering and Related Sciences: Theory and Application. Springer, Cham (2022)
5. Pham, N.T., Dang, D.N.M., Pham, B.N.H., Nguyen, S.D.: SERVER: multi-modal speech emotion recognition using transformer-based and vision-based embeddings. In: ICIIT 2023: 8th International Conference on Intelligent Information Technology, Da Nang, Vietnam, 24–26 February 2023. ACM (2023)
6. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 March 2017, pp. 131–135. IEEE (2017)
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019)
8. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, Red Hook, NY, USA, pp. 6000–6010. Curran Associates Inc. (2017)
9. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)
10. Gemmeke, J.F., et al.: Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, 5–9 March 2017, pp. 776–780. IEEE (2017)
11. Lee, S., Han, D.K., Ko, H.: Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. IEEE Access **9**, 94557–94572 (2021)
12. Lee, Y., Yoon, S., Jung, K.: Multimodal speech emotion recognition using cross attention with aligned audio and text. In: Meng, H., Xu, B., Zheng, T.F. (eds.) Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25–29 October 2020, pp. 2717–2721. ISCA (2020)

13. Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text. In: 2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, 18–21 December 2018, pp. 112–118. IEEE (2018)
14. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
15. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library (2019). https://pytorch.org/
16. Tseng, S.-Y., Narayanan, S., Georgiou, P.G.: Multimodal embeddings from language models for emotion recognition in the wild. IEEE Signal Process. Lett. **28**, 608–612 (2021)
17. Sun, L., Liu, B., Tao, J., Lian, Z.: Multimodal cross- and self-attention network for speech emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021, pp. 4275–4279. IEEE (2021)
18. Pham, N.T., et al.: Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. Expert Syst. Appl. 120608 (2023)