Comparative analysis of multi-loss functions for enhanced multi-modal speech emotion recognition

Phuong-Nam Tran
Dept. of Computing Fundamental
FPT University
Ho Chi Minh City, Vietnam
namtpse150004@fpt.edu.vn

Thuy-Duong Thi Vu
Dept. of Computing Fundamental
FPT University
Ho Chi Minh City, Vietnam
duongvtt9@fe.edu.vn

Nhat Truong Pham
Dept. of Integrative Biotechnology
Sungkyunkwan University
Suwon, Republic of Korea
truongpham96@skku.edu

Hanh Dang-Ngoc
School of Electrical and Data Engineering
University of Technology Sydney
Sydney, NSW, Australia
hanh.n.dang@student.uts.edu.au

Duc Ngoc Minh Dang*
Dept. of Computing Fundamental
FPT University
Ho Chi Minh City, Vietnam
ducdnm2@fe.edu.vn

***Abstract*—In recent years, multi-modal analysis has gained significant prominence across domains such as audio/speech processing, natural language processing, and affective computing, with a particular focus on speech emotion recognition (SER). The integration of data from diverse sources, encompassing text, audio, and images, in conjunction with classifier algorithms has led to the realization of enhanced performance in SER tasks. Traditionally, the cross-entropy loss function has been employed for the classification problem. However, it is challenging to discriminate the feature representations among classes for multi-modal classification tasks. In this study, we focus on the impact of the loss functions on multi-modal SER rather than designing the model architecture. Mainly, we evaluate the performance of multi-modal SER with different loss functions, such as cross-entropy loss, center loss, contrastive-center loss, and their combinations. Based on extensive comparative analysis, it is proven that the combination of cross-entropy loss and contrastive-center loss achieves the best performance for multi-modal SER. This combination reaches the highest accuracy of 80.27% and the highest balanced accuracy of 81.44% on the IEMOCAP dataset.**

***Index Terms*—center loss, contrastive-center loss, cross-entropy loss, multi-modal analysis, multi-modal model, speech emotion recognition**

I. Introduction

Speech emotion recognition (SER) involves analyzing the tone and context of speech to predict the emotional states of the speakers. SER has a wide range of applications, including medical diagnosis, patient care, fraud detection, and lie detection. The rapid growth of artificial intelligence has enabled SER to achieve higher levels of accuracy, particularly with the emergence of multi-modal techniques that combine multiple forms of data. This progress has opened new opportunities for SER to be applied in various industries. Multi-modal techniques utilize multiple inputs from various sources to enhance the features and leverage their fusion to improve model performance. In the context of SER, multi-modal techniques involve combining audio, text, and image data. Recent research has demonstrated the effectiveness of multi-modal techniques in SER, such as the SERVER model [1], which uses BERT [2] for natural language processing and VGGish [3] for audio processing to achieve performance improvements. Another approach, 3M-SER [4], proposes a fusion module based on the multi-head attention mechanism in the Transformer model [5], effectively focusing on the useful features produced by the multiple inputs to further enhance performance.

While the fusion of multiple modalities can lead to diverse sizes, shapes, and value ranges in feature vectors, they all share the commonality of transforming multiple data into feature vectors within latent space. Subsequently, another algorithm is employed to generate distinct feature vectors, representing the fused features from the various inputs. In the training phase, the cross-entropy loss is often utilized to compute the cost function. However, relying solely on cross-entropy loss might not provide enough strength to effectively differentiate feature vectors, thus hindering the ability to specify unique feature vectors for each class within the dataset. Consequently, this limitation can potentially impact the performance of multi-modal techniques.

To address this challenge, we adopt an alternative approach to the loss functions aimed at enhancing the accuracy of multi-modal SER. Specifically, we evaluate the performance of multi-modal SER by comparing several loss functions, such as cross-entropy loss, center loss, contrastive-center loss [6], and their combinations. Subsequently, we integrate these loss functions with the

* *Corresponding author: Duc Ngoc Minh Dang (ducdnm2@fe.edu.vn)*

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

softmax function for the classification and adjustment of feature vectors, resulting in superior accuracy compared to employing only the cross-entropy loss. Our experimentation on the IEMOCAP [7] dataset has showcased a significant enhancement in the performance of the multi-modal SER using the combination of cross-entropy and contrastive-center losses with the softmax function, achieving the highest accuracy (ACC) of 80.27% and the highest balanced accuracy (BACC) of 81.44%.

The rest of the paper is organized as follows. Section II presents a literature review and related studies. Section III provides a detailed explanation of the methodology itself. We also discuss the advantages of our approach over the previous methods. Section IV and Section V present the experiment setup, the dataset used, and the results. Finally, Section VI concludes the study and discusses potential avenues for future research.

II. Related work

A. Multi-modal SER

The multi-modal analysis aims to fuse diverse modalities or sources of information to take advantage of the combined knowledge and enhance the performance of various tasks. These modalities can range from text, images, audio, and videos to sensor data and other machine-readable data. By integrating multiple modalities, multi-modal analysis can extract more comprehensive and accurate insights than single-modal models. Multi-modal approaches have been increasingly used for SER, as they have shown promising results in improving the accuracy and robustness of emotion recognition systems. By combining speech signals with other modalities such as facial expressions, audio, and texts, the multi-modal models can capture a complete picture of the users' emotional states, reducing ambiguity and enhancing the model's ability to recognize and respond to different emotions accurately.

Many studies have applied multi-modal approaches to SER before. A common approach is the fusion of audio and textual information. Textual features, such as transcriptions of spoken words or text-based sentiment analysis, can provide semantic information that complements the acoustic cues captured by audio features. By combining audio and textual modalities, researchers aim to capture both the acoustic and semantic aspects of emotions, improving emotion recognition accuracy.

SERVER [1] proposed a deep learning-based approach that combines audio and textual features for emotion recognition using two well-known architectures, BERT [2] and VGGish [3]. In SERVER, audio features are extracted from the raw audio signals using spectrograms, and textual features are obtained from transcriptions of the corresponding audio segments. SERVER [1] showed a significant improvement in SER accuracy by incorporating both audio and textual features using the BERT and VGGish architectures.

Moreover, the subsequent research conducted in the 3M-SER [4] study further improved the performance by incorporating fusion modules. In the 3M-SER, an attention mechanism was employed to enhance the fusion of textual and audio features, enabling the multi-modal models to gain a deeper understanding of the emotional content within the speech. By leveraging the attention mechanism, the fusion module facilitated improved integration of the two modalities, leading to a more meaningful representation of emotions.

B. Loss functions for classification

Cross-entropy loss with softmax function is a widely used and effective method for solving classification problems. It has proven to be successful in numerous applications. Additionally, this versatile loss function can be employed to train feature-extracted models, enabling them to convert inputs into feature vector representations that capture the underlying meaning of the inputs. By leveraging the power of cross-entropy loss with the softmax function, we can enhance the performance of classification models, leading to better outcomes in various domains. While cross-entropy loss function is a commonly used method for feature separation, it may not always capture sufficient discriminative information. Wen *et al.* [8] highlights the limitations of cross-entropy loss in terms of capturing intra-class variations, which can reduce the model's ability to distinguish among similar samples within the same class.

To address this issue, various techniques have been proposed to enhance the discriminative power of feature extraction models. For instance, feature learning approaches like contrastive-center loss [6] or center loss [8] can be incorporated alongside cross-entropy loss to encourage greater inter-class separation and intra-class compactness. By incorporating these techniques, the model can better differentiate similar samples within the same class while maintaining a clear separation among different classes. Center loss is a powerful technique that enhances the quality of results by minimizing intra-class variations based on Euclidean distances while simultaneously preserving inter-class distinctions using cross-entropy loss.

Despite its effectiveness, the center loss function has a limitation in that it overlooks inter-class separability, as highlighted in [6]. To address this limitation, the contrastive-center loss function [6] is proposed to consider both intra-class compactness and inter-class separability, thereby reducing the weakness observed in the center loss function. Indeed, the contrastive-center loss function has demonstrated remarkable success in numerous applications, particularly in the field of single-modal SER [9], [10]. The utilization of the contrastive-center loss function in these studies has yielded important performance improvements.

III. Methodology

Feature vectors are commonly used to represent the features of input contexts in classification and verification tasks. Various methods have been employed to create a margin among these feature vectors to improve classification accuracy. In multi-modal techniques, multiple features from various sources are combined to achieve efficient feature vectors that represent the closest meaning to the inputs. To create such a feature, the contrastive-center loss is an effective approach, as it creates a high margin among feature vectors. In this study, we compare the effectiveness of the contrastive-center loss function on the fusion feature vectors of two high-performance multi-modal models: SERVER [1] and 3M-SER [4]. Both models utilize text to enhance audio features and improve model accuracy. They use BERT [2] to extract features from the text, which are then concatenated with the audio features extracted from VGGish [3]. Additionally, 3M-SER incorporates a fusion module after concatenating these features to create better fusion feature vectors and further improve the performance of multi-modal techniques.

The fusion feature vectors play a crucial role in the effectiveness of multi-modal models as they capture the primary characteristics of multiple inputs. The accuracy of multi-modal models improves when the feature vectors can better differentiate among unrelated features. However, achieving better intra-class compactness and inter-class separability in the training of fusion feature vectors is challenging when using cross-entropy loss. To address this, we employ center loss and contrastive-center loss [6], which is well-suited for this task and achieves high separability in inter-class and intra-class features. Center loss is a strategy for constructing widely separated classes. It adds a penalty term to the standard supervised loss based on the distance of each data point to its class center. The formula for the center loss function ($\mathcal{L}_c$) is given as follows:

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1}^{n} ||x_i - c_{y_i}||_2^2 \tag{1}$$

where n represents the total number of samples, $x_i \in \mathcal{R}^d$ corresponds to the i^{th} training sample, y_i denotes the label associated with the i^{th} training sample, and a trainable parameter $c \in \mathbb{R}^d$, which represents a center feature vector for each class.

Contrastive-center loss is an extension of center loss that addresses its weakness in inter-class separation. It achieves this by introducing a penalty loss among each class in the dataset, which creates a large distance among inter-class samples in the latent dimension. The formula for the contrastive-center loss function ($\mathcal{L}_{ct-c}$) is given as follows:

$$\mathcal{L}_{ct-c} = \frac{1}{2}\sum_{i=1}^{n} \frac{||x_i - c_{y_i}||_2^2}{(\sum_{j=1,j\neq y_i}^{m} ||x_i - c_j||_2^2) + \theta} \tag{2}$$

where n represents the number of samples in a mini-batch. $x_i \in \mathcal{R}^d$ denotes the i^{th} training sample in the mini-batch, where d is the dimension of the fusion feature vectors. y_i is the label of the i^{th} training sample. m represents the number of classes in the dataset. The parameter θ is added to the inter-class distance to prevent division by zero and to create a margin for easier separation of features. In our experiments, we adopt the default value of $\theta = 1.0$, which is recommended in [6].

We apply the contrastive-center loss function to the fusion feature vectors, which is the concatenation of the text feature vector and audio feature vector in SERVER [1], and the output of the fusion module in 3M-SER [4]. The fully connected layer still employs the cross-entropy loss ($\mathcal{L}_{ce}$) and is given by the following formula:

$$\mathcal{L}_{ce} = -\frac{1}{n}\sum_{i=1}^{n} t_i \log(p_i) \tag{3}$$

where n represents the number of samples in a mini-batch, t_i denotes the ground truth label of the i^{th} training sample, and p_i refers to the predicted probability of the i^{th} training sample using softmax function.

Based on Equations 1, 2 and 3, we have the combination of the contrastive-center loss function and cross-entropy loss function as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{ct-c} \tag{4}$$

And the combination of center loss and cross-entropy loss is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{c} \tag{5}$$

IV. Implementation Detail

A. Hyper-parameters

We utilize the same settings for both SERVER [1] and 3M-SER [4], except for the inclusion of the contrastive-center loss function. All models are trained on a Linux machine (Debian Bookworm) with an Intel(R) Core(TM) i9-12900K processor, 64GB RAM, and 1 Nvidia GeForce RTX 3090 graphics Card. Both SERVER and 3M-SER are trained for 250 epochs with an initial learning rate of 0.0001, which is divided by 10 every 30 epochs. The remaining parameters remain unchanged, following the configuration utilized in SERVER and 3M-SER.

To explore the impact of different loss functions on the multi-modal model's performance, we conducted experiments with five different loss functions: cross-entropy loss function, center loss function, contrastive-center loss function, and the combinations of center loss function with cross-entropy loss function and contrastive-center loss function with cross-entropy loss function. The results of these experiments are presented in Section V.

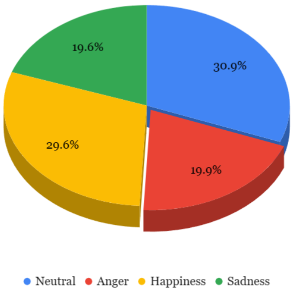

Fig. 1. The IEMOCAP dataset comprises a diverse distribution of emotional samples.

B. Dataset

We utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [7] which is an acted, multimodal and multispeaker database for our experiments. The IEMOCAP dataset contains audiovisual data, such as video recordings, speech recordings, motion capture of facial expressions, and text transcriptions. To assess the effectiveness of SERVER [1] and 3M-SER [4] in conjunction with the contrastive-center loss function, we investigated the same text and audio samples in this study. We consider four major classes, namely anger, happiness, sadness, and neutral, and the distribution of each class is given in Figure 1.

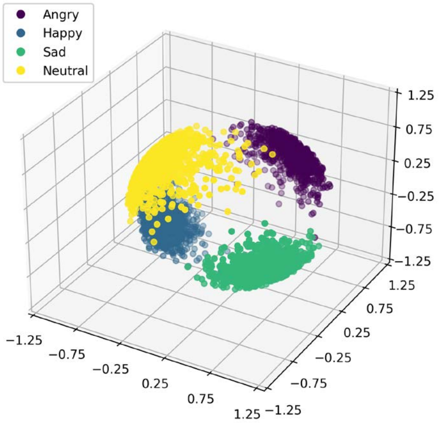

Fig. 2. Visualization of feature representation learning of cross-entropy loss function.

V. Experiments

Figures 2, 3 and 4 depict the learned feature representations using different loss functions. The features obtained from the cross-entropy loss function in Figure 2 are insufficiently discriminative. There is a high closeness and overlap among different classes. Conversely, combining the cross-entropy loss function with the center loss function in Figure 3 yields more discriminative features. The different classes are better separated, and the intra-class variance is reduced. However, even with this combination, there are still instances of overlapping features among different classes.

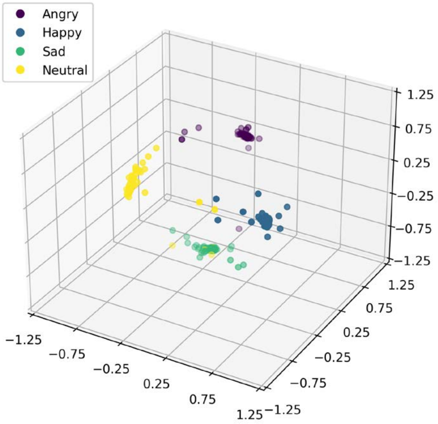

Fig. 3. Visualization of feature representation learning of cross-entropy loss function combined with center loss function.

In contrast, the combination of cross-entropy and contrastive-center loss functions in Figure 4 results in a significant improvement by further reducing intra-class variance and increasing inter-class variance. The features learned from the contrastive-center loss function exhibit enhanced discriminative qualities, with greater separation among different classes. The intra-class variance is minimized, while the inter-class variance is amplified. This is achieved by considering both the similarity among different classes and the dissimilarity within the same class. Consequently, the contrastive-center loss function effectively reduces intra-class variance and amplifies inter-class variance, leading to highly discriminative features and improved class separation.

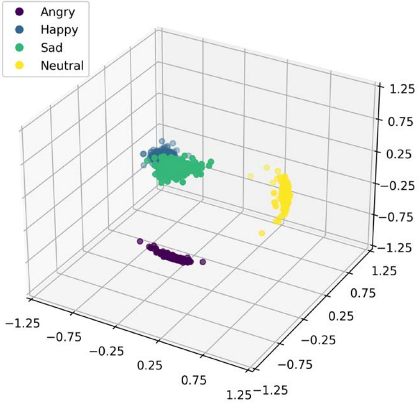

Fig. 4. Visualization of feature representation learning of cross-entropy loss function combined with contrastive-center loss function.

Table I demonstrates a noteworthy improvement compared to the previous method SERVER [1]. The combination of the cross-entropy loss function and center loss function achieves the highest ACC of 65.81% and

TABLE I
Performance comparison of different loss functions in SERVER on IEMOCAP

Method	Accuracy (%)	
	ACC	BACC
Center loss	57.34	58.09
Contrastive-center loss	60.25	60.88
Cross-entropy loss	63.00	63.10
Cross-entropy loss + Center loss	65.81	66.58
Cross-entropy loss + Contrastive-center loss	64.82	66.49

TABLE II
Performance comparison of different loss functions in 3M-SER on IEMOCAP

Method	Accuracy (%)	
	ACC	BACC
Center loss	76.10	78.01
Contrastive-center loss	78.24	79.52
Cross-entropy loss	79.96	80.66
Cross-entropy loss + Center loss	78.56	79.35
Cross-entropy loss + Contrastive-center loss	80.27	81.44

the highest BACC of 66.58%, surpassing the single loss function method. Contrary to expectations, replacing cross-entropy loss does not lead to any improvement in the model's performance. The performance decreases when the cross-entropy loss function is replaced with these alternative loss functions. Notably, the 3M-SER [4] model in Table II exhibits considerable performance enhancement on the IEMOCAP dataset. By leveraging the combined power of the cross-entropy loss function and contrastive-center loss function, the model achieves the highest ACC of 80.27% and the highest BACC of 81.44%. These aforementioned results underscore the importance of adapting fusion features to align with the SER task. By carefully adjusting and fine-tuning the fusion features, we can effectively enhance the model's capabilities and achieve superior results.

Nevertheless, Table II does not show any improvement by combining the cross-entropy loss function with the center loss function. Surprisingly, the performance of the model trained on combined loss functions is even worse than that of the model using only the cross-entropy loss function for training 3M-SER. In contrast, Table I shows enhanced performance in both combination approaches. This discrepancy can be attributed to the fact that 3M-SER employs attention fusion modules to selectively choose and fuse relevant features, while SERVER simply concatenates audio and textual features. The attention fusion modules in 3M-SER aim to ensure that the fusion feature leverages only the useful portion of inputs. Consequently, when an additional loss function is applied to the fusion module, it may not significantly impact the other feature if it has small weights in the attention layer. As a result, the adjustment of feature distances using the additional loss function has a low impact on the overall performance of the model.

VI. Conclusion

To sum up, this paper presents a comprehensive investigation into the impact of different loss functions on the performance of multi-modal SER. Three distinct loss functions and two combinations of them are considered. The experimental results demonstrate that the combination of loss functions can effectively enhance the performance of multi-modal SER. Specifically, by employing the combination of contrastive-center and cross-entropy loss functions on the IEMOCAP dataset, the study achieves an impressive ACC of 80.27% and a BACC of 81.44%. In the future, research endeavors will focus on exploring alternative combinations of loss functions as well as pre-trained audio/speech/text models to further enhance the performance of multi-modal SER.

References

[1] N. T. Pham, D. N. M. Dang, B. N. H. Pham, and S. D. Nguyen, "SERVER: Multi-modal speech emotion recognition using transformer-based and vision-based embeddings," in *Proceedings of the 2023 8th International Conference on Intelligent Information Technology*, 2023, pp. 234–238.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[3] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 131–135.

[4] P.-N. Tran, T.-D. T. Vu, D. N. M. Dang, N. T. Pham, and A.-K. Tran, "Multi-modal speech emotion recognition: Improving accuracy through fusion of vggish and bert features with multi-head attention," in *International Conference on Industrial Networks and Intelligent Systems*. Springer, 2023.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[6] C. Qi and F. Su, "Contrastive-center loss for deep neural networks," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2851–2855.

[7] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision – ECCV 2016*, vol. 9911, 10 2016, pp. 499–515.

[9] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, "A method upon deep learning for speech emotion recognition," *Journal of Advanced Engineering and Computation*, vol. 4, no. 4, pp. 273–285, 2020.

[10] N. T. Pham, D. N. M. Dang, N. D. Nguyen, T. T. Nguyen, H. Nguyen, B. Manavalan, C. P. Lim, and S. D. Nguyen, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *Expert Systems with Applications*, p. 120608, 2023.