

# Gated Cross-modal Attention and Multimodal Homogeneous Feature Discrepancy Learning for Speech Emotion Recognition

Feng Li<sup>(✉)</sup> and Jiusong Luo

Department of Computer Science and Technology,  
Anhui University of Finance and Economics, Anhui, China  
lifeng@aufe.edu.cn

**Abstract.** Understanding human emotions from speech is crucial for computers to comprehend human intentions. Human emotions are expressed through a wide variety of forms, including speech, text, and facial expressions. However, most speech emotion recognition fails to consider the interactions between different information sources. Therefore, we propose a multimodal speech emotion recognition framework that integrates information from different modalities via a gated cross-modal attention mechanism and multimodal homogeneous feature discrepancy learning. Specifically, we firstly extract acoustic, visual and textual features using different pre-train model, respectively. Then, A-GRU-LVC (Auxiliary Gated Recurrent Unit with learnable Vision Center) and A-GRU (Auxiliary Gated Recurrent Unit) is used to further extract emotion-related information for visual and text features. Additionally, we design a gated cross-modal attention mechanism to dynamically fusion multimodal fusion features. Finally, we introduce multimodal homogeneous feature discrepancy learning to better capture differences among different emotion samples. Evaluation results show that our proposed model can achieve better recognition performance than the previous methods on the IEMOCAP dataset.

**Keywords:** Speech emotion recognition · Multimodal · Wav2vec 2.0 · Cross-modal attention mechanism.

## 1 Introduction

As human-computer interaction technology advances, there is a growing demand for machines to accurately recognize and understand human emotions. Speech emotion recognition (SER) plays a crucial role in establishing a foundation for emotional interaction in emotional intelligence systems, facilitating improved communication and emotion perception between machines and humans [1]. It aims to analyze human speech signals for recognizing and understanding the speaker's emotional state. Speech emotion recognition technology has found applications in various fields, including customer service and market research [2], learning and education [3], mental health [4], and social media analytics [5].

The most existing methods for speech emotion recognition can be generally divided into the following two branches. One line of work is machine learning methods that focus on using acoustic information to manually design a set of task-specific features and then applying traditional statistical learning methods to emotion prediction [6]. Another line of work revolves around various neural network models based on deep learning. It automatically learns to feature representations from the original speech signal by using neural network models and has strong expressive power in feature representation [7]. However, deep learning generally requires more data and computational resources compared to traditional machine learning methods. The adoption of transfer learning can effectively alleviate the overfitting problem caused by data scarcity. Speech-based pre-trained models, such as wav2vec [8], wav2vec 2.0 [9], WavLM [10], and VQwav2vec [11], have successfully leveraged large amounts of unlabeled data to learn robust representations of speech signals.

In real-life scenarios, individuals express their emotions through speech and other modalities, such as text and visuals [12,13]. Therefore, relying solely on speech for accurate emotion recognition is insufficient. The cross-modal attention mechanism enables the model to focus on relevant information from different modalities to capture interdependencies and adjust the cross-modal representation [14]. Tsai et al. [15] introduced multimodal Transformer (MuT) directed pairwise cross-modal attention to focus on the interaction between different time-invariant multimodal sequences. Despite notable advancements, certain challenges still persist in multimodal fusion. The important challenge is that most studies only consider differences between homogeneous features of the same modality but different emotions [16]. And it fails to consider the relationship between homogeneous features from different modalities with the same emotion. The latter can lead the model to make better use of the common information between different modalities [17].

In a word, we propose a multimodal speech emotion recognition method based on wav2vec 2.0, with audio as the major modality and text and vision as auxiliary modalities. First, to address the data sparsity problem, we extracted audio features (wav2vec 2.0), text features (Roberta), and visual features (efficientNet) using three different pre-trained models. Second, to capture the global and local information of visual features, we introduce an A-GRU-LVC module. Additionally, we utilize an A-GRU module to extract global information from text features. This enables us to capture important characteristics from each modality. To reduce redundant information during cross-modal fusion, we propose a gated cross-modal attention mechanism to fuse emotion-related information. Finally, we perform multimodal homogeneous feature discrepancy learning, aiming to minimize the distances between samples with the same emotion but different modalities, while maximizing the distances between samples with the same modality but different emotions to enhance the discriminative power of the model. Experimental results demonstrate that our model achieves state-of-the-art performance compared to other multimodal models.

## 2 Proposed methodology

In this section, we focus on the overall structure of the model and then describe in detail the gated cross-modal attention mechanism and multimodal homogeneous feature discrepancy learning.

### 2.1 Model

In this work, the proposed model comprises the pre-train model feature encoder, gated cross-gated attention, and multimodal homogeneous feature discrepancy learning. First, we extract separate video and text features using the auxiliary modality encoder. For the videos, we extract fixed T-frames from each segment and use efficientNet, which is pre-trained on VGGface and AFEW dataset, as a feature extractor to obtain visual features  $\mathbf{e}_j^v$ . After obtaining video features, we feed them into the A-GRU-LVC module that consists of GRU, self-attention, and LVC block to capture global and local information  $X_j^v$ .  $j$  is the  $j_{th}$  video fragment.

Fig. 1 illustrates the structure of the A-GRU-LVC module, consisting of two parallel connected blocks. For the text, we use the roberta-base pre-trained model as a feature extractor to obtain  $\mathbf{e}_j^t$ , followed by using GRU and self-attention mechanism to obtain global features  $X_j^t$ . Second, we use wav2vec 2.0 as an encoder to learn the contextual information  $X_j^a$  of the audio sequences. Third, we design a gated cross-modal attention block to fuse information from different modality  $X_j^F$ . Finally, the shared encoder is responsible for integrating the features from both the auxiliary ( $X_j^v$  and  $X_j^t$ ) and major modalities ( $X_j^a$ ). It leverages multimodal feature discrepancy learning, which focuses on distinguishing representations of the same modality.

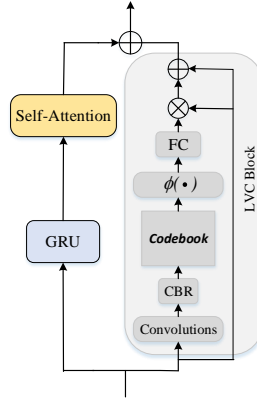


Fig. 1. The structure of the A-GRU-LVC module.

## 2.2 Gated cross-modal attention mechanism

The architecture of the Gated Cross-modal attention mechanism is shown in Fig. 2, which comprises two parallel cross-modal attention mechanisms and a gated filtering mechanism. First, we feed audio features  $X_j^a$ , text features  $X_j^t$  and visual features  $X_j^v$  into two cross-modal attention mechanisms to obtain inter-modal association information, respectively. Specifically, when performing attention computation, we take audio features as Query and textual features or visual features as Key and Value. It can enhance the representation of audio features by introducing information from the text or visual into the audio features.

$$X_j^{F1} = CM_{A-T}(X_j^a, X_j^t) \quad (1)$$

$$X_j^{F2} = CM_{A-V}(X_j^a, X_j^v) \quad (2)$$

Then, the augmented features  $X_j^{F1}$  and  $X_j^{F2}$  are processed through the following gated filtering mechanism.

$$P_* = \text{sigmoid}(FC(X_j^{F1} \oplus X_j^{F2})) \quad (3)$$

$$X_j^F = P_* \odot X_j^{F1} + (1 - P_*) \odot X_j^{F2} \quad (4)$$

The ratio of each channel can be dynamically defined by a learnable parameter that filters out misinformation generated during cross-modal interactions.

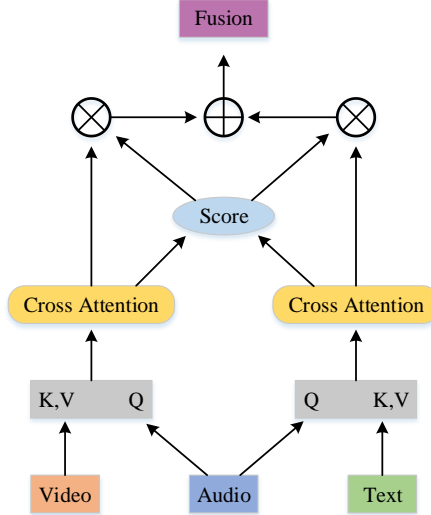


Fig. 2. The architecture of the Gated Cross-modal attention mechanism.

### 2.3 Multimodal homogeneous feature discrepancy learning

Multimodal homogeneous feature discrepancy learning has made significant progress in multimodal emotion recognition [16,17]. It can optimize the modal representation ability and extract richer and more accurate emotional information by learning the relationships and differences between homogeneous features. First, we feed unfused audio features  $X_j^a$ , text features  $X_j^t$  and visual features  $X_j^v$  into a shared encoder to obtain homogeneous features. It minimizes the feature gap from different modalities and contributes to multimodal alignment.

$$X_{com}^{m[i]} = SD(X_j^m), m \in (a, t, v) \quad (5)$$

where  $SD$  is the shared encoder function that consisting of a simple linear layer.

From [17], we argue that homogeneous representations from the same emotion but different modalities should be more similar than homogeneous representations from the same modality but different emotions. Therefore, in our work, we perform multimodal homogeneous feature discrepancy learning to enhance the interactions between the same emotions but different modalities, and amplify the differences between the same modalities but different emotions. We define this loss function as margin loss.

$$L_{mar} = \frac{1}{|M|} \sum_{(i,j,k) \in M} \max\left(0, \alpha - \cos\left(X_{com}^{m[i],c[i]}, X_{com}^{m[j],c[j]}\right) + \cos\left(X_{com}^{m[i],c[i]}, X_{com}^{m[k],c[k]}\right)\right) \quad (6)$$

where  $M = \{(i, j, k) / m[i] \neq m[j], m[i] = m[k], c[i] = c[j], c[i] \neq c[k]\}$  is a triple tuple set. The  $m[i]$  is the modality of sample  $i$ , and the  $c[i]$  is the class label of sample  $i$ .  $\cos$  denotes the cosine similarity between two feature vectors. By applying a distance margin  $\alpha$ , we ensure that the distance between positive samples is smaller than the distance between negative samples. Here, positive samples refer to the same emotion but different modalities, and negative samples refer to the same modality but different emotions.

Like many multimodal emotion recognition studies, cross-entropy is commonly used as a loss function to optimize model parameters and improve the accuracy for classification during training.

$$L_{task}^{emotion} = -\frac{1}{N_D} \sum_{j=0}^{N_D} y_j \cdot \log y_j \quad (7)$$

where,  $y_j$  is the true label of the sample, and  $\hat{y}_j$  is the prediction of the sample.  $N_D$  is the number of samples in the dataset  $D$ .

Finally, we combine the constraints to form the whole emotion recognition loss.

$$L_{total}^{emotion} = L_{task}^{emotion} + \lambda L_{mar} \quad (8)$$

where  $\lambda$  is the balance factor.

### 3 Experiment evaluation

#### 3.1 Dataset

The IEMOCAP dataset is a widely used benchmark in emotion recognition research. It consists of 12 hours of improvised and scripted audio-visual data from 10 UC theatre actors (five males and five females). The dataset is divided into five binary sessions, and each conversation is annotated with emotional information in four modalities: video, audio, transcription, and motion capture of facial movements. We evaluate our model using audio, transcribed, and video data, focusing on six emotions: happy, neutral, angry, excited, sad and frustrated. The dataset contains a total of 7380 data samples. For evaluation, we employ a five-fold cross-validation approach. The dataset is divided into five equal parts (80% training and 20% testing), with each session serving as the test set once.

#### 3.2 Setting

Our experiments are performed on a single NVIDIA GeForce RTX 3060 card using the PyTorch. The epochs are set to 50, and the batch size is set to 2 during training. For calculating the loss in the training phase, we employ CrossEntropy loss and margin loss. The model parameters are updated using the Adam optimizer, with a learning rate of  $1e-5$ . To control the intensity of the margin loss, we introduce a hyperparameter denoted as  $\lambda$ . The  $\lambda$  allows us to adjust the impact of the margin loss on the overall training process. The number of gated cross-modal attention mechanism is 3.

On the IEMOCAP datasets, we adopt the Accuracy and Weighted F1 as evaluation metrics. The calculation of these metrics is represented by Equation 9. These metrics provide insights into the overall accuracy and the weighted performance across different emotion classes.

$$\begin{aligned}
 Accuracy &= \frac{\sum_1^k n_i}{\sum_1^k N_i} \\
 precision &= \frac{n_i}{M_i}, recall = \frac{n_i}{N_i} \\
 F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
 Weighted\ F1 &= \sum_1^k \frac{N_i \times F1_i}{N}
 \end{aligned} \tag{9}$$

where the  $N_i$  means the number of utterances in  $i_{th}$  class, the  $n_i$  means the number of correctly recognized utterances in  $i_{th}$  class and  $k$  means the number of classes. The  $M_i$  represents the number of all emotions identified as  $i_{th}$  class.

## 4 Experiment results

### 4.1 Ablation Study

In the ablation studies, we train the proposed network in several scenarios: Audio, Text, Visual, Audio and Text, Audio and Visual, and the whole modalities. Specifically, when using a single modality, we excluded the gated cross-modal attention and multimodal homogeneous feature discrepancy learning. When using both modalities, we removed the gate mechanism within gated cross-modal attention and multimodal homogeneous feature discrepancy learning.

The results presented in Table 1 demonstrate that the highest accuracy and weighted average F1 scores were achieved when using all three modalities. This outperforms other scenarios that utilized only one or two modalities. In the unimodal setting, audio features performed the best, followed by text, and then vision, which had the lowest performance. Combining audio with text yielded superior results compared to combining audio with visual, indicating a higher correlation between audio and text modalities. However, combining audio with video resulted in a reduced WF1 value compared to audio unimodal, suggesting that the fusion of audio and video information might introduce some redundancy or noise.

**Table 1.** Ablation studies on each modality.

Modality	ACC(%)	WF1(%)
A	66.06	65.59
T	58.74	58.63
V	29.88	26.31
A+T	67.75	67.45
A+V	66.33	64.14
<b>A+V+T</b>	<b>70.26</b>	<b>70.29</b>

We also investigate the impact of multimodal homogeneous feature discrepancy learning in our framework. During these experiments, we assigned weights to the balance factor ( $\lambda$ ) for margin loss and observed its effects under various weight values. The corresponding results are presented in Table 2. The results indicate that the best performance on the IEMOCAP dataset is achieved when  $\lambda = 1$ .

In this study, the model shows a significant improvement in accuracy (ACC) by 2.98% and weighted F1 score (WF1) by 2.94% compared to without incorporating margin loss ( $\lambda = 0$ ). This demonstrates the effectiveness of multimodal homogeneous feature difference learning in enhancing the model's ability to distinguish emotions across different modalities. However, we also noticed that the performance degrades when the balance factor is too large ( $\lambda = 10$ ). This suggests that excessively emphasizing the margin loss might negatively impact the original classification task.

**Table 2.** Ablation study on balance factor  $\lambda$  of margin loss.

$\lambda$	ACC	WF1
0	67.41	67.31
0.01	68.23	67.94
0.1	69.34	69.12
1	70.26	70.29
10	65.06	64.49

## 4.2 Comparative Analysis

To validate the robustness of the proposed network, we conducted a thorough comparison of our proposed model with several state-of-the-art techniques. The results are presented in Table 3, which show the average performance on the IEMOCAP. Our method outperformed the previous state-of-the-art by 0.57% in ACC and 0.43% in WF1.

**Table 3.** Quantitative comparison with multimodal methods on IEMOCAP dataset.

Methods	ACC(%)	WF1(%)	Year
DialogueTRM [19]	68.92	69.23	2020
MMGCN [20]	-	66.22	2021
COGMRN [21]	68.20	67.63	2022
MM-DFN [13]	68.21	68.18	2022
M2FNet [12]	69.69	69.86	2022
HAAN-ERC[22]	69.48	69.47	2023
<b>Ours</b>	<b>70.26</b>	<b>70.29</b>	<b>2024</b>

## 5 Conclusion

In this paper, we propose a new approach for speech emotion recognition by leveraging pre-trained models and cross-modal attention. In contrast to previous work using pre-trained models and cross-modal attention mechanisms, we designed gated cross-modal attention to dynamically fuse features from different modalities. In addition, we introduce the concept of multimodal homogeneous feature discrepancy learning, which helps the model to effectively learn and distinguish representations of the same modalities but different emotions. Experiments are performed on the IEMOCAP dataset and have achieved state-of-the-art results. In future work, we will explore more efficient optimization algorithms to construct novel architectures for SER.



**Acknowledgments.** This work was supported in part by the Innovation Support Program for Returned Overseas Students in Anhui Province under Grant No. 2021LCX032, Natural Science Foundation of the Higher Education Institutions of Anhui Province under Grant No. KJ2021A0486, and Excellent Research and Innovation Team of Universities at Anhui Province under Grant No. 2023AH010008.

## Reference

1. Xu, M., Zhang, F., Cui, X. and Zhang, W.: Speech emotion recognition with multiscale area attention and data augmentation. In ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp. 6319-6323 (2021)
2. Li, X. and Lin, R.: Speech emotion recognition for power customer service. In 2021 7th International Conference on Computer and Communications (ICCC), pp. 514-518 (2021)
3. Li, W., Zhang, Y. and Fu, Y.: Speech emotion recognition in e-learning system based on affective computing. In Third International Conference on Natural Computation (ICNC), Vol. 5, pp. 809-813 (2007)
4. Elsayed, N., ElSayed, Z., Asadizanjani, N., Ozer, M., Abdelgawad, A. and Bayoumi, M.: Speech emotion recognition using supervised deep recurrent system for mental health monitoring. In 2022 IEEE 8th World Forum on Internet of Things (WF-IoT), pp. 1-6 (2022)
5. Ahire, V. and Borse, S.: Emotion detection from social media using machine learning techniques: a survey. In Applied Information Processing Systems: Proceedings of ICCET, pp. 83-92 (2022)
6. Schuller, B., Rigoll, G. and Lang, M.: Hidden Markov model-based speech emotion recognition. In 2003 IEEE International conference on acoustics, speech and signal processing (ICASSP) (2003)
7. Fayek, H.M., Lech, M. and Cavedon, L.: Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92, pp.60-68 (2017)
8. Schneider, S., Baevski, A., Collobert, R. and Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)
9. Baevski, A., Zhou, Y., Mohamed, A. and Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, pp.12449-12460 (2020)
10. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X. and Wu, J.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), pp.1505-1518 (2022)
11. Baevski, A., Schneider, S. and Auli, M.: vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv preprint arXiv:1910.05453 (2019)
12. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P. and Onoe, N.: M2fnet: Multi-modal fusion network for emotion recognition in conversation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4652-4661 (2022)

13. Hu, D., Hou, X., Wei, L., Jiang, L. and Mo, Y.: MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7037-7041 (2022)
14. Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F. and Shen, H.T.: Cross-modal attention with semantic consistence for image–text matching. *IEEE transactions on neural networks and learning systems*, 31(12), pp.5412-5425 (2020)
15. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P. and Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the conference. Association for computational linguistics. Meeting (Vol. 2019, p. 6558). NIH Public Access (2019)
16. Hazarika, D., Zimmermann, R. and Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on multimedia, pp. 1122-1131 (2020)
17. Li, Y., Wang, Y. and Cui, Z.: Decoupled multimodal distilling for emotion recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6631-6640 (2023)
18. Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017)
19. Mao, Y., Sun, Q., Liu, G., Wang, X., Gao, W., Li, X. and Shen, J.: Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637* (2020)
20. Hu, J., Liu, Y., Zhao, J. and Jin, Q.: Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779* (2021)
21. Joshi, A., Bhat, A., Jain, A., Singh, A. and Modi, A.: COGMEN: COntextualized GNN based multimodal emotion recognitioN. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4148-4164) (2020)
22. Zhang, T., Tan, Z. and Wu, X.: HAAN-ERC: hierarchical adaptive attention network for multimodal emotion recognition in conversation. *Neural Computing and Applications*, 35(24), pp.17619-17632 (2023)