GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

A Secure Federated Learning Algorithm for Emotion Recognition Towards Multimodal Speaker Signals on the Client Side

Xin Wang [1,*], Longlong Qiao [1], Guangxin Dai [1], Quanping Chen [2], Yan Zhang [3] and Wensong Li [4]

[1] School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, 710021, China
[2] School of Electrical and Control Engineering, Shaanxi University of Science and Technology, Xi'an, 710021, China
[3] China HuaYin Ordnance Center, Huayin, 714200, China
[4] CS&C Information System Engineering Co., Ltd, 102209, Beijing, China
* Correspondence: Xin Wang, School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, 710021, China

Abstract: With the widespread application of multimodal data in dialog emotion recognition, effectively integrating text, audio, and visual information while addressing data heterogeneity across multiple clients and ensuring user privacy has become a key research challenge. This paper integrates a Transformer self-distillation model with attention scores and federated learning algorithms to propose a multimodal emotion recognition framework. The framework employs intra-modal and inter-modal Transformers to capture multimodal interactions, enhances modality representations through attention weights, and incorporates a federated learning structure to safeguard data privacy. A global model distance-weighted aggregation strategy is introduced to mitigate model bias caused by heterogeneous data. Experimental results on the IEMOCAP dataset demonstrate that the proposed framework achieves superior emotion recognition accuracy and exhibits more stable model convergence compared to existing baseline models.

Keywords: multimodal emotion recognition; federated learning; transformer; attention scores; heterogeneity

Received: 01 November 2025
Revised: 20 November 2025
Accepted: 25 December 2025
Published: 05 January 2026

1. Introduction

Emotion Recognition in Conversations (ERC) aims to automatically identify the emotional labels of each utterance within a dialogue [1]. Due to its wide range of applications, such as opinion mining, healthcare, and the development of empathetic dialogue systems, this task has become a significant research focus in recent years. Unlike traditional context-independent sentence emotion recognition (ER), the core of ERC lies in modeling contextual dependencies and speaker-sensitive interactions. This process involves dynamically capturing the flow of dialogue and the evolution of emotional states throughout the entire conversation [2,3].

Emotion Recognition in Conversations (ERC) is crucial for constructing robust emotional analysis models. While existing research has primarily focused on capturing context and speaker-sensitive dependencies within the textual modality, it often overlooks the significance of multimodal information. Multimodal emotion recognition, by integrating textual, acoustic, and visual information, can effectively enhance the accuracy of emotional understanding in complex conversational environments.

However, two major challenges often arise in practical applications: first, the effective fusion and representation enhancement of multimodal data; second, data privacy protection and model training in heterogeneous environments. Traditional centralized

training methods carry the risk of privacy leakage, while Federated Learning (FL), although capable of protecting data privacy, often suffers from performance degradation due to heterogeneous data distributions and varying device capabilities. To address these issues, this paper proposes an innovative framework that integrates Transformer self-distillation with federated learning. The framework employs Transformer structures on the client side for multimodal feature fusion and enhancement, while leveraging a self-distillation mechanism to improve single-modal representation capabilities. On the server side, a federated learning approach is adopted, combined with distance correction and weighted aggregation strategies to mitigate model bias caused by heterogeneity, thereby achieving efficient and privacy-preserving multimodal emotion recognition.

2. Related Work

According to established psychological theories, individuals express emotions in various ways, including through linguistic content, vocal cues, and facial expressions [4]. Consequently, multimodal information is considered more valuable than unimodal data for achieving a comprehensive understanding of emotional states.

2.1. Multimodal Emotion Recognition

In recent years, Transformer-based multimodal fusion methods have demonstrated remarkable performance in the field of emotion recognition. Early research on multimodal fusion primarily focused on early fusion and late fusion strategies. Early fusion approaches integrate features from different modalities at the initial input stage [5,6]. In contrast, late fusion strategies construct separate models for each modality and subsequently integrate their outputs through methods such as majority voting or weighted averaging [7,8]. However, it has been observed that both of these fusion methods fail to effectively capture complex intra-modal and inter-modal interactions [9].

Subsequently, model-based fusion gained popularity, leading to the development of various specialized models. For instance, specific frameworks explicitly model unimodal, bimodal, and trimodal interactions by computing the Cartesian product of features [10]. Other approaches utilize low-rank weight tensors for multimodal fusion to reduce the complexity of the interaction modeling [9]. Furthermore, attention mechanisms have been employed to learn cross-modal interactions while storing information over time through multi-view gated memory structures [11]. Cross-modal transformers have also been utilized to model long-range dependencies across different modalities [12]. Recent advancements include the fine-tuning of large pre-trained transformer models for multimodal language by designing a Multimodal Adaptive Gate (MAG) [13]. Other methods employ a unimodal label generation strategy to obtain independent unimodal supervision and then jointly learn multimodal and unimodal tasks [14]. Additionally, transformer encoders have been adopted to model intra-modal and inter-modal interactions within sequences of modalities [15].

However, analysis reveals that while these methods consider both intra-modal and inter-modal information, they often fail to account for the varying degrees of contribution from different modalities during the model fusion process. In practical applications, the roles played by each modality in the fusion model are not identical. Specifically, contextually linked textual and auditory information often plays a more critical role than visual imagery in emotional expression.

2.2. Privacy-Preserving Federated Learning

Federated Learning (FL) achieves effective protection of user data privacy by training models on distributed devices and aggregating their parameters. Consequently, it is increasingly being adopted by research institutions in fields such as healthcare and finance. FL is a distributed machine learning training framework. During the FL training process, clients distributed across different geographical locations can train models based

on their local datasets. These locally trained models are then sent to a central server, where they are averaged and aggregated into a global model. Subsequently, the server distributes the updated global model to a selected set of clients for the next round of FL training [16]. In 2018, specific large-scale keyboard systems integrated multi-party user data through Federated Learning, significantly enhancing prediction accuracy for users with diverse input habits [17]. This distributed framework offers a feasible approach to breaking down data silos while ensuring user privacy and security.

To address the issue of data heterogeneity in FL, research has suggested that reducing the number of local training epochs and increasing the communication frequency between clients and the server can partially mitigate the deviation between local and global models. It has been pointed out that such deviation consists of two components: accumulated historical gradient errors and data distribution errors from the current iteration [18]. To tackle problems caused by data heterogeneity and system heterogeneity in the FedAvg algorithm, the FedProx algorithm was proposed [19]. This algorithm introduces a proximal term to the local objective and computes an inexact solution to the local objective function, enabling heterogeneous devices to achieve convergence more quickly and proceed to the global aggregation stage. However, it has been noted that the naive averaging aggregation methods used in both FedAvg and FedProx can indirectly lead to inconsistencies between local and global objectives, thereby reducing the test accuracy of the model [20]. Furthermore, some argue that most federated learning methods handling heterogeneous data rely on the transmission and aggregation of gradient parameters, which can incur substantial communication overhead and risk gradient leakage [21].

Alternative strategies have proposed the FedProto algorithm, which employs the prototype concept to update local models, offering a new perspective for handling heterogeneous data in federated learning [22,23]. However, the FedProto algorithm requires a fixed workload for local training even with small-sample data, and heterogeneity is more likely to lead to overfitting or underfitting of the model [24]. Moreover, the weighted aggregation method in FedProto treats all local prototypes equally, without considering that some prototypes may have a greater influence on the global prototype, resulting in local deviations from the global prototype. Additionally, the feature extraction method of the prototype network embedding layer can easily cause class embeddings to be compressed to a single point when the number of input training samples is limited [25]. To address the issue of the collapse of different sample embedding vectors into a single point, triplet loss contrastive learning was proposed, which aims to bring features with the same label closer in spatial position while pushing those with different labels farther apart [26]. However, since triplet loss essentially involves pairwise similarity comparisons between samples, it is susceptible to the influence of sample size and the selection of positive and negative samples [27]. If triplet loss alone is used to train a network on a small-sample dataset, the model may converge to a suboptimal solution.

In summary, while existing FL algorithms possess certain strengths, they have not adequately addressed the negative impacts of data sample and training heterogeneity on model performance. Effectively handling heterogeneous data, utilizing prototypes to transmit heterogeneous model parameters, and efficiently aggregating local models remain ongoing challenges in the field of Federated Learning. Therefore, considering specific contexts, especially the privacy protection requirements in multimodal emotion recognition within healthcare settings—such as handling sensitive medical data—this paper addresses the issue of data heterogeneity across clients. Building upon the FedMPD algorithm, we propose a strategy based on distance optimization and aggregation to achieve stable convergence during client-side multimodal training in heterogeneous environments.

Would you like me to move on to the Methodology section or help you with the Abstract and Conclusion using these same constraints?

3. Method

The multimodal federated learning emotion recognition algorithm proposed in this paper, designed for client-side speaker heterogeneous data, is broadly divided into two phases: client-side multimodal modeling and server-side federated aggregation.

3.1. Client-Side Multimodal Intra-Modal and Inter-Modal Modeling

This section sequentially presents the definition of multimodal, the design of modality encoders, multimodal fusion, and an introduction to the emotion classifier.

3.1.1. Client-Side Multimodal Definition

Let the conversation consist of N consecutive utterances $\{u_1, u_2, \cdots, u_N\}$ and M clients, assuming that each client represents a speaker $\{s_1, s_2, \cdots, s_N\}$. Each utterance u_i is spoken by a speaker $S_{\varphi(u_i)}$, where φ is a mapping between an utterance and its corresponding speaker index. Each utterance$u_i$involves textual (t), acoustic (a), and visual (v) modalities. We denote the sequences of textual, acoustic, and visual modalities for all utterances in the conversation as $U_t = [u_1^t; u_2^t; \cdots; u_N^t] \in R^{N\times dt}$, $U_a = [u_1^a; u_2^a; \cdots; u_N^a] \in R^{N\times da}$, and $U_v = [u_1^v; u_2^v; \cdots; u_N^n] \in R^{N\times dv}$, respectively. The ERC task aims to predict the emotional label DDD u_i for each utterance from a predefined set of emotion categories. Each client performs local training using a Transformer-based self-distillation model [28].

3.1.2. Modality Encoders

The primary function of the modality encoder is to learn intra-modal and inter-modal interactions among conversational utterances by obtaining enhanced modality sequence representations. To describe the intra-modal and inter-modal encoders more clearly, let two modalities be denoted as m and n, with H_m and H_n representing the multimodal utterance sequences corresponding to modalities m and n, respectively, where $m \in \{t, a, v\}$, $n \in \{t, a, v\} - \{m\}$, and t, a, v stand for text, audio, and video, respectively.

-- For the intra-modal encoder, a Transformer-based model is employed, where the query, key, and value are set to the same H_m. The extraction of intra-modal information can thus be expressed as Equation (1):

$$H_{m\to m} = \text{Transformer}(H_m, H_m, H_m) \in R^{N\times d} \quad (1)$$

This intra-modal transformer enhances the representation of the m-modal sequence for capturing intra-modal interactions within the utterance sequence.

-- For the cross-modal transformer, with H_m as the query and H_n as the key and value, the inter-modal interaction information can be represented by Equation (2):

$$H_{n\to m} = \text{Transformer}(H_m, H_n, H_n) \in R^{N\times d} \quad (2)$$

This Transformer model enables the transfer of information from the m-modal to the n-modal, capturing inter-modal interactions within the utterance sequence.

3.1.3. Multimodal Fusion Based on Attention Weight Scores

The main function of this module is to dynamically learn the weights of different modalities and enhance the fused representation between them.

Let the output of the m-modal corresponding to text, audio, and visual be a $d \times 1$ dimensional sequence vector $X \in R^{d\times 1}$, where d is an integer. The normalized Pearson correlation coefficient between the previous modality and the current m-modal is denoted as $\rho \in \{0, \cdots, 1\}$. This coefficient is used as the initialized attention score, and the modality vector X is then optimized by readjusting it based on the attention score.

Let $W \in R^{F\times U}$ be the trainable weight matrix corresponding to the three modalities at the current moment, where U is the number of units in the intermediate layer (typically a hyperparameter). tanh denotes the hyperbolic tangent activation function. According to Equation (3), the nonlinear relationship of the extracted feature u_i is:

$$u_i = \tanh(X \cdot W) \quad (3)$$

Let $V \in R^{U\times 1}$ be the training matrix from the previous modality. Using Equation (4), the feature relationship can be transformed into a scalar score $e_i \in R^{N\times 1}$:

$$e_i = u_i \cdot V \tag{4}$$

To prevent the attention scores e_i from having an excessive range of values, appropriate scaling is required. Let U denote the dimension of matrix V. The attention weights a_i are calculated by normalization using the softmax function:

$$a_i = \text{softmax}\left(\frac{e_i}{\sqrt{U}}\right) \tag{5}$$

Then, adjust the weight a_i based on Equation (5) and the initial attention score ρ. First, an anomaly factor β is introduced. To prevent the weight from becoming negative, let the range of β be $\{-1,+\infty]$. When $\beta = 0$, it indicates no additional weight is added to the data; when $-1 < \beta < 0$it indicates a reduction in the weight assigned to the data; and when $\beta > 0$, it indicates an additional weight is added to the data. Using the correlation coefficient ρ and the attention weight a_i, the adjusted attention weight a_i' for the current modality encoding X' is calculated by the formula:

$$a_i' = \rho \cdot (1 + \beta \cdot a_i) \tag{6}$$

Each feature in the three-modality vector X is scaled using the adjusted attention weight a_i' described above, resulting in:

$$X' = a_i' \cdot X \tag{7}$$

Thus, the feature data X of modality m has been scaled based on the adjusted weights.

3.1.4. Emotion Classifier Module for Predicting Emotional Labels

To compute the probability of identifying emotion categories, the multimodal sequence of conversational utterances obtained in the previous step can be fed into a softmax classifier. The procedure can be referred to in reference [29].

Here, the Transformer self-distillation mechanism is employed to transfer knowledge from the global model's soft labels and hard labels to each modality, thereby enhancing the representation capability of individual modalities.

3.2. Distance-Optimized Server-Side Federated Aggregation

The server receives the locally computed feature centroids (i.e., the feature centers for each category) uploaded by each client and performs distance-weighted aggregation.

(1) Feature Extraction： The client extracts features via an embedding network and performs clustering locally to obtain local feature centers.

The core idea of the client-side embedding network is to transform high-dimensional heterogeneous image inputs into low-dimensional embedded vector outputs through multi-layer nonlinear transformations. Taking the image modality in the m-modality as an example, the client speaker inputs the local heterogeneous image data x_i into the embedding network. After a series of transformations by convolutional layers and fully connected layers, the embedded vector $g_\phi(x_i)$ is obtained. This embedded vector is mapped to a space of OUTDIM dimensions, where OUTDIM represents the dimensionality of the embedded vector. In this space, feature vectors of samples from the same category are tightly clustered together, while feature vectors of samples from different categories are pushed apart to maintain sufficient distance, thereby enhancing the discriminative capability of the features. Without loss of generality, each dimension of the three modalities is processed in this manner.

Distance Correction： Distance correction is primarily reflected in the design of the client-side loss function. By introducing a contrastive loss term, inter-class discriminability is enhanced. In the FedMPD algorithm, the client-side loss function $\mathcal{L}(\cdot)$ is designed as shown in Equation (8):

$$\mathcal{L}(\omega_i) = \sum_{i=1}^{n} \frac{|D_i|}{N} \mathcal{L}_S(f_i(\omega_i;x_i),y_i) + \lambda \sum_{i=}^{n} \sum_{k=1}^{K} \frac{|D_{i,k}|}{N_k} \mathcal{L}_{reg}^{k}(c_i^{(t,k)},c^{(t,k)}) \tag{8}$$

In Equation (8), ω_i represents the local model of the client, x_i denotes the data samples used by the client for model training, while $f_i(\omega_i;x_i)$ and y_i correspond to the predicted value and the true label of the sample, respectively. The second term on the right-hand side of the equation is the regularization term of the original loss function in Equation (8), which aims to mitigate the deviation between the local and global label class prototypes. Here, $c_i^{(t,k)}$,$c^{(t,k)}$ represent the local label class prototype and the global prototype, respectively, and λ is the regularization hyperparameter for this term, which is set to a fixed value of 0.1 in this paper.

To ensure that prototypes that are close in distance but belong to different label classes remain distinguishable, the FedMPD algorithm introduces a contrastive loss regularization term $\mathcal{L}_{reg}^{k}(c_i^{(t,k)},c^{(t,k)})$, formulated as shown in Equation (9), in addition to the supervised learning loss function $\mathcal{L}_S(f_i(\omega_i;x_i),y_i)$. This term encourages local prototypes of the same label class to converge toward the global prototype, while local prototypes of different label classes are pushed away from it, with a minimum separation distance of r. As shown in Equation (9), the distance r is defined as the distance between the k-th label class prototype and its nearest $(k+1-th)$ global prototype, i.e., $r = \min d(c^{(t,k)},c^{(t,k+1)})$, as expressed in Equation (9).

$$\mathcal{L}_{reg}^{k}(c_i^{(t,k)},c^{(t,k)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} d(c_i^{(t,k)},c^{(t,k)}) + \sum_{l\neq k}\left(\max\left\{0,r - d(c_i^{(t,k)},c^{(t,l)})\right\}\right) \quad (9)$$

Weighted Aggregation： Weights are dynamically assigned based on the distance between local and global, where closer distances correspond to larger weights, as shown in Equation (10):

$$c^{(t+1,k)} = \text{ContrScore}\cdot\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla c_i^{(t,k)}。 \quad (10)$$

3.3. Overall Training Procedure

Based on Sections 3.1 and 3.2 above, the overall workflow of the multimodal federated learning emotion recognition algorithm for client-side speaker heterogeneous data can be divided into four steps:

（1）Clients train the SDT model using local multimodal data;

（2）Clients upload the trained model parameters to the server;

（3）The server performs distance-weighted aggregation to update the global model;

（4）The global model is distributed to clients for the next round of training.

The detailed multimodal federated learning emotion recognition algorithmic process is outlined in Table 1.

Table 1. Multimodal Federated Learning Emotion Recognition Algorithm.

For client i in round t of FL:
Input: Client's local multimodal dataset $\mathcal{D}_i$, global model $c^{(t,k)}$, number of clients $i = 1,\dots,n$, hyperparameter for the contrastive loss regularization term λ, local model from the previous round w_i^t
Output: The client trains the Self-Distillation Transformer (SDT) model using local multimodal data, producing the local model w_i^{t+1} for round t
1 n client speakers participate in the training
2 The server sends the global model $c^{(t,k)}$ to the client
3 The client receives the global model
4 The client performs the following operations:
5 Update the local model $c_i^{(t,k)} = \frac{1}{\lvert S_{i,k}\rvert}\sum_{(x_i,y_i)\in S_{i,k}} g_\phi(x_i)$
6 IF $d(g_\phi(x_i),c_i^{(t,k)}) + \alpha < d(c_i^{(t,k)},c_i^{(t,k+1)})$ THEN
7 Compute the loss
$\mathcal{L}(\omega_i) = \sum_{i=1}^{n}\frac{\lvert\mathcal{D}_i\rvert}{N}\mathcal{L}_S(\mathcal{F}_i(\omega_i;x_i),y_i) + \lambda\sum_{i=}^{n}\sum_{k=1}^{K}\frac{\lvert\mathcal{D}_{i,k}\rvert}{N_k}\mathcal{L}_{reg}^{k}(c_i^{(t,k)},c^{(t,k)})$

8 Update the local model using gradient descent

$$w_i^{t+1} \leftarrow w_i^t - \eta \frac{\partial \mathcal{L}(\omega_i; x_i)}{\partial \omega_i}$$

9 The local model is retained on the client side w_i^{t+1}

10 Upload the local model $c_i^{(t,k)}$ to the server

11 The server performs distance-weighted aggregation to update the global prototype $c^{(t+1,k)}$

$$c^{(t+1,k)} = \sum_{i=1}^{n} \sum_{k=1}^{K} (r - d(c_i^{(t,k)}, c^{(t,k)})) c_i^{(t,k)}$$

12 The server distributes the local model for round $t+1$ denoted as $c^{(t+1,k)}$。

4. Experiments

This paper implements the proposed model using PyTorch and compares it with the following baseline models.

CMN: It uses two GRUs and memory networks to model contextual information for two speakers but is only suitable for dyadic conversations [30].

ICON: An extension of CMN, it uses another GRU to capture emotional influence between speakers. Similar to CMN, this model is designed for dyadic conversations. It employs three different GRUs to track the speaker, context, and emotional state in the conversation, respectively. The aforementioned models concatenate textual, acoustic, and visual features to obtain multimodal utterance representations [31].

MMGCN: It constructs a conversational graph based on all three modalities and designs a multimodal fusion graph convolutional network to model contextual dependencies across multiple modalities [32].

dialogueTRM: It uses a hierarchical transformer to handle differentiated contextual preferences within each modality and designs a multi-grained interactive fusion to learn the varying contributions of an utterance across modalities [33].

MM-DFN: It designs a graph-based dynamic fusion module to integrate multimodal contextual features, reduce redundancy, and enhance complementarity between modalities [34].

MMTr: It utilizes different bidirectional long short-term memory networks (Bi-LSTMs) to learn contextual representations at both the speaker's self-context level and the conversational context level, and designs a cross-modal fusion module to enhance representations of weaker modalities [35].

For a fair comparison, we re-ran all baseline models.

4.1. Dataset and Settings

The multimodal emotion recognition training model for clients is primarily evaluated on the IEMOCAP dataset. By simulating the heterogeneity of client data in federated learning, the performance of the proposed model is assessed.

The federated learning framework follows the experimental approach outlined in reference, with the local training epochs fixed at 50. The data is distributed across 20 node devices, and the number of sample categories per device is generated using a random number generator. During the training phase, the server prioritizes selecting "online" clients and those with data volumes in the top 50% to participate in federated modeling [36].

For client-side multimodal emotion recognition, Adam is used as the optimizer. The initial learning rate for IEMOCAP is set to $1.0e^{-4}$, with a batch size of 16 and a temperature τ set to 1. For the 1D convolutional layers, the number of input channels for the text, acoustic, and visual modalities (i.e., their corresponding feature dimensions) on IEMOCAP is set to 1024, 1582, and 342, respectively. For all three modalities on the dataset, the number of output channels and kernel size are set to 1024. For the Transformer

encoder, its hidden size, number of attention heads, feedforward size, and number of layers are set to 1024, 8, 1024, and 1, respectively. To prevent overfitting, the L2 weight decay is set to $1.0e^{-5}$, and dropout is applied with a rate of 0.5. All results are averaged over 10 runs [28].

4.2. Results and Analysis

The experiments show that Table 2 presents the performance of the baselines and the proposed method on the IEMOCAP dataset. On the IEMOCAP dataset, the proposed method outperforms all baselines, achieving improvements of 1.76% in overall accuracy and 2.35% in weighted F1-score compared to MMTr. Additionally, significant improvements are observed in the F1-scores for most emotion categories.

Table 2. Performance of Various Models on the IEMOCAP Dataset.

Modcls	IEMOCAP													
	happy		sad	neutral		angry		CXcited		fnustrated				
	ACC	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	F1
CMN	23.31	20.30	56.33	61.52	52.34	51.31	60.76	60.17	57.19	61.75	72.46	62.27	54.87	56.33
ICON	26.00	32.30	66.35	72.17	56.99	58.50	68.41	66.29	70,90	68.01	75.92	65.08	62,85	62.25
MMGCN	32.64	38.63	71.65	73.89	65.10	62.81	73.53	72.43	77.93	74.42	65.19	63.63	66.61	66.25
DialogueTRM	60.21	56.85	85,40	80.45	66.27	68.56	76.47	65.99	75.15	76.16	51.39	58.09	68.52	68.20
MM-DFN	34.44	45.41	76.55	77.10	72.15	66.99	75.88	70.88	74.65	76.42	58.27	61.57	67.84	67.85
MMTr	67.64	54.67	84.49	87.77	76.14	71.66	71.59	65.04	75.24	76.26	55.91	62.29	69.77	69.91
Our method	73.81	67.12	78.81	82.94	78.63	76.22	72.98	69.96	77.79	81.18	66.14	70.68	74.95	75.08

Moreover, in heterogeneous environments, the model demonstrates more stable convergence with smaller fluctuations. The proposed model does not require uploading raw data, meaning the data never leaves the local devices—only model training parameters are exchanged. This satisfies the privacy protection requirements of each client.

5. Conclusion

This paper proposes a federated learning-based multimodal emotion recognition model that balances the efficient fusion of multimodal information with data privacy protection. By employing intra-modal and inter-modal encoders, the model captures interactions both within and across modalities in conversational utterances. To dynamically learn the weights between different modalities, we designed an attention mechanism-based fusion strategy. This allows the model to adjust the weights of different modalities according to the context when uploading to the global model, further enhancing overall performance. Experiments were conducted on the IEMOCAP dataset, and the results demonstrate the high effectiveness, superiority, and privacy security of the proposed method.

Through the experiments in this paper, we observed that, beyond optimizing the model structure and adapting to the computational constraints of edge devices, dynamically adjusting modality weights, enhancing the discriminative capability of multimodal representations, and addressing low-correlation emotions under privacy protection remain further challenges in emotion recognition. These aspects will be

explored in future work. The framework presented in this paper offers a new perspective at the intersection of multimodal emotion recognition and federated learning.

Funding: This work was supported by Shaanxi University of Science & Technology Teaching Reform Research Project No. 25Z014 and Shaanxi University of Science & Technology Graduate Course Development Project No. KC2024Y09.

References

1. A. Kumar, P. Dogra, and V. Dabas, "Emotion analysis of Twitter using opinion mining," In *2015 Eighth International Conference on Contemporary Computing (IC3)*, August, 2015, pp. 285-290. doi: 10.1109/ic3.2015.7346694
2. B. C. Gül, S. Nadig, S. Tziampazis, N. Jazdi, and M. Weyrich, "FedMultiEmo: Real-Time Emotion Recognition via Multimodal Federated Learning," *arXiv preprint arXiv:2507.15470*, 2025. doi: 10.1109/iceccme64568.2025.11277502
3. L. Zhou, J. Gao, D. Li, and H. Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53-93, 2020.
4. A. Mehrabian, "Silent messages (Vol. 8, No. 152, p. 30)," *Belmont, CA: Wadsworth*, 1971.
5. M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46-53, 2013.
6. S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," In *2016 IEEE 16th international conference on data mining (ICDM)*, December, 2016, pp. 439-448. doi: 10.1109/icdm.2016.0055
7. B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L. P. Morency, "Deep multimodal fusion for persuasiveness prediction," In *Proceedings of the 18th ACM international conference on multimodal interaction*, October, 2016, pp. 284-288.
8. O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," *arXiv preprint arXiv:1805.00705*, 2018.
9. Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July, 2018, pp. 2247-2256.
10. A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.
11. A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," In *Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1)*., April, 2018.
12. Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," In *Proceedings of the conference. Association for computational linguistics. Meeting*, July, 2019, p. 6558.
13. W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L. P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," In *Proceedings of the 58th annual meeting of the association for computational linguistics*, July, 2020, pp. 2359-2369.
14. W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," In *Proceedings of the AAAI conference on artificial intelligence*, May, 2021, pp. 10790-10797. doi: 10.1609/aaai.v35i12.17289
15. Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," In *Proceedings of the 29th ACM international conference on multimedia*, October, 2021, pp. 4400-4407. doi: 10.1145/3474085.3475585
16. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," In *Artificial intelligence and statistics*, April, 2017, pp. 1273-1282.
17. A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.
18. Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
19. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429-450, 2020.
20. J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611-7623, 2020.
21. A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," In *International conference on machine learning*, May, 2019, pp. 634-643.
22. Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," In *Proceedings of the AAAI conference on artificial intelligence*, June, 2022, pp. 8432-8440.

23. C. Chen, J. Zhang, A. K. Tung, M. Kankanhalli, and G. Chen, "Robust federated recommendation system," *arXiv preprint arXiv:2006.08259*, 2020.
24. V. Smith, C. K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.
25. O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.
26. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815-823. doi: 10.1109/cvpr.2015.7298682
27. A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
28. H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Transactions on Multimedia*, vol. 26, pp. 776-788, 2023. doi: 10.1109/tmm.2023.3271019
29. J. C. D. SILVA, "MFFER: Multimodal Federated-Learning based Facial Emotion Recognition," 2025.
30. D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L. P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, June, 2018, p. 2122.
31. D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2594-2604. doi: 10.18653/v1/d18-1280
32. J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," *arXiv preprint arXiv:2107.06779*, 2021. doi: 10.18653/v1/2021.acl-long.440
33. Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li, "DialogueTRM: Exploring multi-modal emotional dynamics in a conversation," In *Findings of the association for computational linguistics: EMNLP 2021*, November, 2021, pp. 2694-2704. doi: 10.18653/v1/2021.findings-emnlp.229
34. D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "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)*, May, 2022, pp. 7037-7041. doi: 10.1109/icassp43922.2022.9747397
35. S. Zou, X. Huang, X. Shen, and H. Liu, "Improving multimodal fusion with main modal transformer for emotion recognition in conversation," *Knowledge-Based Systems*, vol. 258, p. 109978, 2022. doi: 10.1016/j.knosys.2022.109978
36. A. Nandi, and F. Xhafa, "A federated learning method for real-time emotion state classification from multi-modal streaming," *Methods*, vol. 204, pp. 340-347, 2022. doi: 10.1016/j.ymeth.2022.03.005