FedMEKT: Split Multimodal Embedding Knowledge Transfer in Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

Federated Learning (FL) enables a decentralized machine-learning paradigm to collaboratively train a generalized global model without sharing users' private data. However, most existing FL approaches solely utilize single-modal data, thus limiting the systems for exploiting valuable multimodal data in future personalized applications. Furthermore, most FL methods still rely on the labeled data at the client side, which is limited in real-world applications due to the inability of data self-annotation from users. To leverage the representations from different modalities in FL, we propose a novel multimodal FL framework with a semi-supervised learning setting. Specifically, we develop the split multimodal embedding knowledge transfer mechanism in federated learning, namely, FedMEKT, which enables the personalized and generalized multimodal representations exchange between server and clients using a small multimodal proxy dataset. Hence, FedMEKT iteratively updates the generalized encoders from the collaborative embedding knowledge of each client, such as modality-averaging representations. Thereby, a generalized encoder could guide personalized encoders to enhance the generalization abilities of client models; afterward, personalized classifiers could be trained using the proxy labeled data to perform supervised tasks. Through the extensive experiments on three multimodal human activity recognition tasks, we demonstrate that FedMEKT achieves superior performance in both local and global encoder models on linear evaluation and guarantees user privacy for personal data and model parameters.

1 Introduction

With the tremendous emerging development of technologies, AI has gained many achievements with multiple remarkable applications such as virtual assistants, e-commerce, recommendation and healthcare Pawar et al. (2020); Lugano (2017). The rapid growth of AI technologies necessitates a massive amount of personalized data at the end-users. Consequently, data privacy concerns become a hindrance in centralized machine learning system where the server collects personal data and train the deep neural networks model to provide AI services. Due to privacy and security issues, federated learning (FL) has been proposed as a decentralized machine learning paradigm that aggregates the model parameters from multiple users without sharing private data, thus protecting user information McMahan et al. (2017); Kairouz et al. (2021). FL only requires the clients to transfer local model parameters to the server, therefore, guaranteeing data privacy of users.

Despite many advantages in avoiding privacy leakage, existing FL methods consider the scenario where clients own only single-modal data, restricting the utilization of multimodal data in various equipment. Until now, many recent works on deep multimodal learning illustrate that the complementary information from multimodal data provides more accuracy and robustness performance than single-modal data in many applications such as text and image in language translation Rajendran et al. (2015), different wearable sensors in healthcare Garcia-Ceja et al. (2018), audio and video for emotion recognitionLiang et al. (2018). Thus, the design for the FL framework using multimodal data becomes more practical where users own data generated from multiple data sources and devices such as smartphones and smartwatches. To leverage the benefits of multimodal data for decentralized machine learning systems, there have been some prior works in multimodal FL. One approach designs the co-attention layer to merge the representations from different modalities to obtain the fused features to train the personalized models Xiong et al. (2022). This method requires all users to have labeled data from all modalities, which means that users have to annotate the data. However, in the real world, this could be a cost hindrance for users to collect the labeled data from different modalities, such as various types of sensor data. One possible solution to save the annotation cost is to design the multimodal FL under the semi-supervised setting where clients own private unlabeled data, and the server holds labeled data for the supervised training task. To deal with the labeled constraint of local clients in multimodal FL, Zhao et al. (2022) proposed the framework that works under the semi-supervised setting using multiple autoencoders for different modalities. This work applies the mechanism of the traditional FL framework FedAvg McMahan et al. (2017) by aggregating model parameters from local clients to construct the global multimodal encoders from client autoencoder models that for supervised training tasks. Nonetheless, those above methods rely on the model parameters aggregation on the server from the skewed private data based on the average scheme, which can cause degradation in the generalization ability of personalized models and limit the personalization ability of the global model.

In this paper, we design a novel multimodal FL framework under a semi-supervised setting to tackle the limitations of existing multimodal FL works Zhao et al. (2022); Xiong et al. (2022) and resolve the labeled data constraint in users. We thus propose FedMEKT, a novel split multimodal embedding knowledge transfer-based semi-supervised learning technique that adopts the federated learning systems to achieve the generalized encoder for participating users. To this end, we formulate the embedding knowledge transfer mechanism in multimodal FL utilizing the split multimodal autoencoders Ngiam et al. (2011), which enable communication between the server and clients by leveraging the small multimodal proxy dataset. Specifically, instead of updating the global model by aggregating local model parameters, FedMEKT updates generalized autoeconder model via the collaborative embedding knowledge from all clients such as modality-averaging representations. We illustrate that our proposed method can achieve significant speedups and outperform the multimodal version of FedAvg Zhao et al. (2022) in both global and personalized encoders on supervised tasks. Our main contributions are:

- For the first time, we propose the knowledge transfer mechanism in multimodal FL. We design a novel multimodal FL framework FedMEKT that allows the embedding knowledge exchange between the server and clients.
- We design the embedding knowledge transfer with four general problems: 1) generalized multimodal autoencoder construction to transfer the knowledge from client encoders to global encoders, 2) personalized multimodal autoencoder learning to transfer knowledge from the global encoder to client encoders, 3) generalized classifier learning to train the classifier for global supervised tasks, 4) personalized classifier learning to train the personalized classifier for personalized supervised tasks.
- We deploy the personalized classifier for each client to improve the personalized performance of the client encoders. We validate the proposed method by conducting extensive experiments over three multimodal activity recognition datasets and achieve superior performance in both global and local classification tasks compared to multimodal FedAvg.

2 Related Work

Semi-supervised Learning Semi-supervised learning (SSL) has been applied in various machine learning tasks to leverage the unlabeled data to solve the labeling cost issue. The semi-supervised setting considers the scenario where the system holds both unlabeled data and labeled data Zhu & Goldberg (2009). Pseudo labeling Lee et al. (2013) has become one of the popular trends in SSL, which generates the pseudo label for unlabeled dataset and compute the loss function based on the loss of original labels and pseudo labels. The combination of pseudo-label and consistency regularization has been widely applied in SSL with many state-of-the-art methods, such as UDA Xie et al. (2020), FixMatchSohn et al. (2020), and MixMatchBerthelot et al. (2019). In our work, we consider the semi-supervised setting scenario in the decentralized setting where private unlabeled data are provided by users, and labeled data is utilized in training classifiers to perform supervised tasks.

Federated Learning based Knowledge Distillation Knowledge distillation (KD) Buciluǎ et al. (2006); Ba & Caruana (2014) has become a promising technique that enables FL to solve heterogeneity issues. KD generally provides communication methods between global and local models

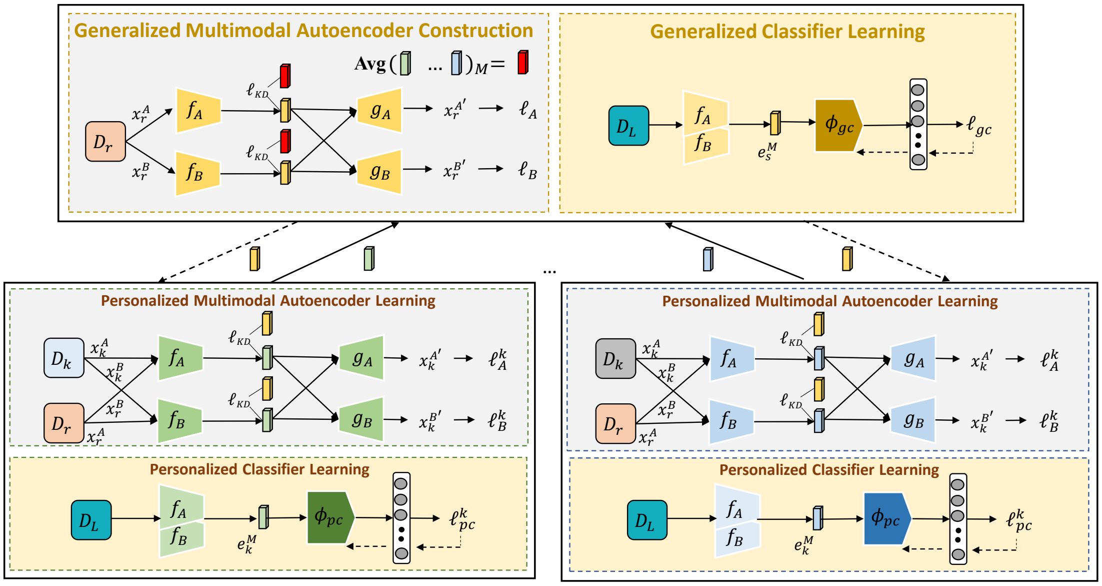

Figure 1: The Embedding Knowledge Transfer in Multimodal Federated Learning.

instead of exchanging model parameters. The authors in Jeong et al. (2018) applied KD in FL to minimize the communication overhead by using the distillation regularizer between student and teacher logits. The clients update the local model to achieve the averaged global predictions on the server. Another approach FedDF Lin et al. (2020), first uses the model parameters aggregation to obtain a global model and then updates the averaged global model again by performing ensemble distillation from all student client models. Unlike FedDF, which still uses the model parameters exchange between the clients and the server, KT-pFL Zhang et al. (2021) formulates the personalized knowledge transfer for personalized FL leveraging the proxy data to update the local soft predictions. Most of these schemes apply KD in traditional FL with single-modal labeled data. In our framework, we first propose the knowledge transfer scheme in multimodal FL.

Multimodal Learning Multimodal learning has attracted lots of attention in recent years. The multimodal deep learning systems enable leveraging data from multiple modalities such as image, video, sensors, etc., hence provide better performance than unimodal data. There have been many emerging techniques to study multimodal learning implementation. One of the first designs in multimodal deep learning was fusion Baltrušaitis et al. (2018) that fuses representations of different layers from multiple modalities using various methods such as concatenation, multiplication, or weighted sum. However, this method still faces misalignment in different fusion levels. In later years, researchers propose different model architectures for other multimodal applications such as co-attention Lu et al. (2016) for VQA tasks Kumar et al. (2020), various types of transformers for language-video tasksSun et al. (2019). In terms of the encoder-decoder framework, Ngiam et al. (2011) proposed the autoencoders for audio and visual data. Another popular approach is CCA Andrew et al. (2013) combination of canonical correlation analysis and autoencoders to fuse multimodal representations in the feature subspace.

Emerging multimodal FL works Xiong et al. (2022); Zhao et al. (2022) motivate us to initiate the novel design of a knowledge transfer scheme in multimodal FL. In this work, the generalized encoders could alternatively build from personalized encoders and drive the personalized encoders to improve both personalized and generalized encoder performance. As a result, the better encoder helps to strengthen the supervised tasks' performance. In the following section, we presented the proposed Split Multimodal Embedding Knowledge Transfer scheme and algorithm.

3 Split Multimodal Embedding Knowledge Transfer

Recent multimodal FedAvg scheme Zhao et al. (2022) enables exploiting multimodal data with FL framework using the averaging parameter aggregation mechanism. The global autoencoder model is aggregated based on the model parameters from unimodal and multimodal client models after the local training process with a private unlabeled dataset $D_k = \{\mathbf{x}_k^M\}$ in client k. The multimodal client models are given more weights than unimodal client models to align the representation from different modalities. Regarding two modalities such as A and B, the goal of MM-FedAvg is to learn the split global autoencoder model that minimizes the total loss function over the entire unlabeled dataset from K number of clients. Using the similar setting in this work, we denote $n_A = \sum_{k \in m_A, m_{AB}} n_k$, $n_B = \sum_{k \in m_B, m_{AB}} n_k$ are the total number data samples of the modality A and B, respectively and m denotes the set of clients in each modality (i.e., client with modality m_{AB} holds both data from modality A and B). In this work, we consider all clients have multimodal data such as clients own different types of sensory data (e.g., accelerometer data, gyroscope data). We first define the split autoencoder model with the embedding knowledge (i.e., $e^M = f(x^M)$), where f and g are the encoder and decoder for the modality M of the autoencoder model, respectively. The embedding knowledge could be extracted from different hidden layers of the encoder (i.e., e^{M_h}), where h is the number of hidden layers of the encoder for the modality M. In particular, the learning objective functions for the global autoencoders A and B are denoted as:

$$\min_{f_A, g_A} \ell_s(f_A, g_A) = \sum_{k \in m_A} \frac{n_k}{n_A} \ell_k(f_A, g_A) + \alpha \sum_{k \in m_{AB}} \frac{n_k}{n_A} \ell_k(f_A, g_A) \tag{1}$$

$$\min_{f_B, g_B} \ell_s(f_B, g_B) = \sum_{k \in m_B} \frac{n_k}{n_B} \ell_k(f_B, g_B) + \alpha \sum_{k \in m_{AB}} \frac{n_k}{n_B} \ell_k(f_B, g_B) \tag{2}$$

where $\ell_k(f_A, g_A)$, $\ell_k(f_B, g_B)$ are the loss functions for the split autoencoder for each modality A and B. Specifically, the loss functions at client k with two modalities can be defined as:

$$\ell_k(f_A, g_A) = \min_{f_A, g_A} \ell_A(x^A, x^{A'}) + \ell_B(x^B, x^{B'}) \tag{3}$$

$$\ell_k(f_B, g_B) = \min_{f_B, g_B} \ell_A(x^A, x^{A'}) + \ell_B(x^B, x^{B'}), \tag{4}$$

where $x^{A'}$ and $x^{B'}$ are the reconstructed outputs of two modalities, ℓ_A and ℓ_B are reconstruction loss functions Zhao et al. (2022) (e.g., MSE loss) for modalities A and B, respectively. Different from existing multimodal FL methods, instead of aggregating the parameters from clients in each global round to train the global model, we deploy the multimodal embedding knowledge transfer mechanism to transfer the local embedding knowledge from all multimodal clients to collectively build the generalized global encoder model which can generate more generalized representations. Hence, we leverage the small unlabeled proxy data $D_r = \{\mathbf{x}_r^A, \mathbf{x}_r^B\}$ that all clients can access and provide their embedding knowledge, as shown in Fig. 1. Moreover, the global encoder model can transfer back the generalized multimodal embedding knowledge to all clients. Thereby, the generalization capability of personalized autoencoder models could be enhanced by mimicking the received global representations of proxy data. In this design, the FL framework could guarantee the model parameters and data privacy compared to parameters exchange between the server and clients.

Turn this multimodal FL scheme into reality, we develop the FedMEKT algorithm (Alg. 1) to perform the embedding knowledge transfer mechanism between multimodal clients and the server. At the beginning of each communication round, the server randomly selects a subset of clients from the total of K multimodal clients to participate in the local training, and the global autoencoder model broadcasts the generalized embedding knowledge to all selected clients. Each client performs P local training steps with its private multimodal data and proxy dataset in the personalized knowledge transfer problem (11), (12) and outputs the local embedding knowledge using proxy data D_r, then sends it to the server. At the same time, we attach a local classifier to each client and perform the personalized classifier training on labeled dataset $D_L = \{\mathbf{x}_L^A, \mathbf{x}_L^B, \mathbf{y}_L\}$ to solve the personalized classifier learning (14) problem. The generalized autoencoder model is constructed on the server side by solving the generalized multimodal learning problem (5), (6). Then the server uses the updated global encoder to extract the multimodal representations of input data in the labeled dataset D_L to train the classifier for the supervised learning task (13). Primarily, we design the embedding knowledge transfer in multimodal FL for FedMEKT with four general problems: *generalized multimodal autoencoder construction*, *personalized multimodal autoencoder learning*, *generalized classifier learning*, and *personalized classifier learning* in the following subsections.

3.1 Generalized multimodal autoencoder construction

To solve the generalized multimodal autoencoder construction problem, we utilize the proxy data D_r to collect the embedding knowledge from all multimodal devices to build the global autoencoder model. On the server, we conduct the modality-averaging mechanism of the local embedding knowledge based on the modality label. Accordingly, we gather the embedding knowledge from same modality of all clients and then perform the averaging operation to obtain the collective knowledge of each modality for the global model. Subsequently, we design the generalized multimodal autoencoder construction problem for the split autoencoder with two modalities using embedding knowledge distillation (EKD) method as follow:

$$\ell_s(f_A, g_A|D_r) = \ell_A(x_r^A, x_r^{A'}|D_r) + \ell_B(x_r^B, x_r^{B'}|D_r) + \beta\, \ell_{EKD}\bigg(e_s^{A_h}, \sum_{k \in K} \frac{1}{K} e_k^{A_h}|D_r\bigg) \tag{5}$$

$$\ell_s(f_B, g_B|D_r) = \ell_A(x_r^A, x_r^{A'}|D_r) + \ell_B(x_r^B, x_r^{B'}|D_r) + \beta\, \ell_{EKD}\bigg(e_s^{B_h}, \sum_{k \in K} \frac{1}{K} e_k^{B_h}|D_r\bigg), \tag{6}$$

where β is the parameter to control the trade-off between reconstruction loss of proxy data and EKD regularizer. The personalized knowledge is obtained from each layer h in the personalized encoders. The collaborative embedding knowledge from clients is aggregated sequentially for modalities A and B by averaging representations according to their modality. The EKD regularizer attempts to close the gap between the embedding knowledge of the server and collaborative embedding knowledge from multimodal clients.

3.2 Personalized Multimodal Autoencoder Learning

At the client side, we design the similar mechanism for personalized multimodal autoencoder learning that encourage the local models to improve the generalization capabilities by mimicking the generalized embedding knowledge from the global encoder model. In the personalized learning problem, the client models update depends on private unlabeled data D_k and proxy data D_r, which is accessible for all clients. In particular, we propose the personalized multimodal loss function for each device k on both modalities A and B as follows:

$$\ell_c^k(f_A, g_A|D_k, D_r) = \ell_A^k(x_k^A, x_k^{A'}|D_k) + \ell_B^k(x_k^B, x_k^{B'}|D_k) + \alpha\, \ell_{EKD}(e_k^{A_h}, e_s^{A_h}|D_r) \tag{11}$$

$$\ell_c^k(f_B, g_B|D_k, D_r) = \ell_A^k(x_k^A, x_k^{A'}|D_k) + \ell_B^k(x_k^B, x_k^{B'}|D_k) + \alpha\, \ell_{EKD}(e_k^{B_h}, e_s^{B_h}|D_r) \tag{12}$$

where α is the parameter to manipulate the trade-off between the local reconstruction loss and the embedding knowledge transfer regularizer. The local regularization term helps to enhance the generalization of local multimodal encoder models and avoid the biasness issue when training on the skewed private dataset.

3.3 Generalized classifier learning

On the server, we attach the global classifier ϕ_{gc} to generalized encoder part of the global autoencoder model to perform the supervised learning task by using multimodal spare labeled dataset

Algorithm 1 FedMEKT Algorithm

```
Input: $T, K, R, N, L, P, \eta_1, \eta_2, \eta_3, \eta_4$
for $t = 0, \dots, T-1$ do
    – Client Execution –
    Personalized Multimodal Autoencoder Learning: Each device $k$ receives the global embedding knowledge $\boldsymbol{e}_s^{M_h}$ from the server and updates local autoencoder model for each modality $M$ sequentially with $D_k$, $D_r$;
    for $n = 0, \dots, N-1$ do
        We loop for each batch of private and proxy data (i.e., $S_k$ and $S_r$):
            $w_k^{t,M} = w_k^{t,M} - \eta_1 \nabla \ell_c^k(f_M, g_M | S_k, S_r),\ \forall M \in \{A, B\};$    (7)
    end for
    Personalized Classifier Learning: Each device $k$ updates their private classifier with $D_L$;
    for $p = 0, \dots, P-1$ do
        We loop through batches of labeled data (i.e., $S_L$):
            $w_{\phi k}^{t,M} = w_{\phi k}^{t,M} - \eta_2 \nabla \ell_{pc}^k(z_k^M | S_L),\ \forall M \in \{A, B\};$    (8)
    end for
    Device $k$ generates the local embedding knowledge from multiple hidden layers $e_k^{M_h}$ of all modalities by using proxy data $D_r$ and send to the server;
    – Server Execution –
    Generalized Multimodal Autoencoder Construction: Server updates the global model with the collaborative embedding knowledge from the selected clients for each modality $M$ sequentially
    for $r = 0, \dots, R-1$ do
        We loop through batches of proxy data (i.e., $S_r$):
            $w_g^{t,M} = w_g^{t,M} - \eta_3 \nabla \ell_s(f_M, g_M | S_r),\ \forall M \in \{A, B\};$    (9)
    end for
    Generalized Classifier Learning: Server updates the generalized classifier with the representation from the aggregated generalized encoder using spare multimodal labeled dataset $D_L$
    for $l = 0, \dots, L-1$ do
        We loop through batches of labeled data (i.e., $S_L$):
            $w_{\phi g}^{t,M} = w_{\phi g}^{t,M} - \eta_4 \nabla \ell_{gc}(z_s^M | S_L),\ \forall M \in \{A, B\};$    (10)
    end for
end for
```

D_L. The classifier training process helps to update solely the generalized classifier for classification downstream tasks. Hence, we use cross-entropy loss to learn the generalized multimodal classifier:

$$\ell_{gc}(D_s) = \ell_{CE}(z_s^M | D_L),\ \forall M \in \{A, B\}; \tag{13}$$

where $e_s^M = f_M(x_L^M)$, $z_s^M = \phi_{gc}(e_s^M)$ are the representations from the global encoder and the outcome of the global classifier, respectively.

3.4 Personalized classifier learning

To solve the personalized supervised problem, we design the private classifier ϕ_{pc}^k for each client k from personalized encoder on different modalities. By using labeled data D_L, we freeze the parameters in the updated personalized encoder and train the private classifier using the cross-entropy loss:

$$\ell_{pc}^k(D_L) = \ell_{CE}^k(z_k^M | D_L),\ \forall M \in \{A, B\}; \tag{14}$$

where $e_k^M = f_M^k(x_L^M)$, $z_k^M = \phi_{pc}^k(e_k^M)$ are the private representations of each client and the outcome of the personalized classifier, respectively.

4 Experimental Results

4.1 Experimental Setup

Datasets In this section, we evaluate the efficiency of FedMEKT algorithm with Opportunity (Opp) Chavarriaga et al. (2013), mHealth Banos et al. (2014), UR Fall Detection Kwolek & Kepski (2014) datasets which are three multimodal human activity recognition (HAR) tasks. We conducted the experiments with 10 selected clients in each round from 30 clients. For the data generation in three multimodal datasets, we follow the experimental setup from Zhao et al. (2022) to generate training and testing data for federated systems. In terms of the proxy dataset in each dataset, we generate from the subset of data that are separated from training data and testing data, and the size is approximate to the testing data. For the labeled data for supervised training, we randomly sampled from the training dataset. We provide the dataset details in the Appendix.

Baselines For comparison, we evaluate different settings of our FedMEKT algorithm and compared them with MM-FedAvg Zhao et al. (2022) algorithm. Regarding the global performance in our method, the setting $h1$, and $h2$ stand for using the knowledge from one hidden layer, and two hidden layers, respectively. For the local evaluation, we compare our method with a personalized classifier for each client and the global classifier the same as MM-FedAvg. In particular, pc and gc stand for personalized classifier and global classifier, respectively.

Implementation Details We develop our FedMEKT algorithm on Pytorch library Paszke et al. (2019). We simulate the experiments on our server with one NVIDIA GeForce GTX-1080 Ti GPU using CUDA version 11.2 and Intel Core i7-7700K 4.20GHz CPU with sufficient memory for model training. We use the LSTM Hochreiter & Schmidhuber (1997) autoencoders with 2 LSTM layers, and extract the knowledge from 2 hidden layers for the knowledge transfer, scheme. Both global and local classifiers are implemented as two-layer perceptrons for supervised tasks using ReLU activation function.

Evaluation metrics We evaluate the global and local encoders by extracting the representations for the supervised training tasks. We train a linear classifier on the frozen representations from global and local encoders and report the F_1 score as the results. For the global performance, we report the F_1 score from the global classifier, and for the local models' performance, we report the mean of all local classifiers' F_1 results.

4.2 Experimental Results

4.2.1 Performance Comparison

In summary, we summarize the experimental results of global and local performance on three multimodal human activity recognition datasets with the mean accuracy from round 90 to 100 in Table 1, Table 2, and Table 3. In the case of global performance, FedMEKT obtains the comparable or better accuracy performance MM-FedAvg method. Compared to the Mm-FedAvg algorithm, FedMEKT outperforms in most modalities combination cases in mHealth and UR Fall Detection datasets and achieves a slightly better performance in the Opp dataset. The results illustrate the effectiveness when transferring embedding knowledge from more hidden layers. In terms of local linear evaluation, we obtain comparable performance in most cases in all datasets. For some combination cases, although our method cannot outperform the MM-FedAvg, we still obtain competitive results. Moreover, FedMEKT can improve personalized performance by utilizing the personalized classifier for each client as depicted in problem 14. We also show the curve of global performance on the UR Fall Detection dataset in Fig. 2 . These figures demonstrate that our proposed scheme achieves a more stable and better convergence in terms of global performance in most scenarios. For other datasets, we provide the figures in the Appendix.

4.2.2 Ablation Studies

Effect of Embedding Knowledge Transfer Steps R In this experiment, we compare the performance of FedMEKT under the different settings of embedding knowledge transfer (EKT) steps on the UR Fall Detection dataset. As shown in Table 4, in most scenarios, the value $R = 2$ achieves the highest performance compared to other values. For the Rgb modality in the Rgb-Depth combina-

Table 1: The comparison of global peformance on mHealth and UR Fall Detection datasets from 90 to 100 rounds

Methods	**mHealth**						**UR Fall Detection**					
	Acce-Gyro		Acce-Mage		Gyro-Mage		Acce-Rgb		Acce-Depth		Rgb-Depth	
	Acce	Gyro	Acce	Mage	Gyro	Mage	Acce	Rgb	Acce	Depth	Rgb	Depth
MM-FedAvg	64.53	62.41	68.84	**71.82**	61.24	66.76	61.70	57.88	60.63	60.76	69.88	67.61
FedMEKT($h1$)	66.70	61.28	70.28	70.36	64.34	**67.96**	65.81	59.02	66.28	61.00	71.93	68.93
FedMEKT($h2$)	**68.44**	**65.04**	**70.62**	71.48	**65.14**	67.14	**69.32**	**60.21**	**69.25**	**65.85**	**73.81**	**70.33**

Table 2: The comparison of local peformance on mHealth and UR Fall Detection datasets from 90 to 100 rounds

Methods	**mHealth**						**UR Fall Detection**					
	Acce-Gyro		Acce-Mage		Gyro-Mage		Acce-Rgb		Acce-Depth		Rgb-Depth	
	Acce	Gyro	Acce	Mage	Gyro	Mage	Acce	Rgb	Acce	Depth	Rgb	Depth
MM-FedAvg	**64.24**	63.02	**68.77**	**71.39**	61.22	66.78	60.75	57.93	60.83	60.04	68.52	**67.20**
FedMEKT(gc)	58.80	61.63	64.34	65.51	60.35	65.32	61.36	57.28	53.10	60.42	65.61	63.68
FedMEKT(pc)	63.57	**63.04**	67.26	67.28	**63.48**	**67.62**	**61.85**	**61.52**	**62.27**	**63.06**	**69.11**	66.36

Table 3: The comparison of performance on Opp dataset from 90 to 100 rounds

Table 3a:Global performance

Methods	**Opp**	
	Acce-Gyro	
	Acce	Gyro
MM-FedAvg	71.51	72.12
FedMEKT($h1$)	71.47	71.58
FedMEKT($h2$)	**72.12**	**72.20**

Table 3b:Local Performance

Methods	**Opp**	
	Acce-Gyro	
	Acce	Gyro
MM-FedAvg	71.40	72.11
FedMEKT(gc)	70.24	70.89
FedMEKT(pc)	**71.83**	**72.17**

Table 4: The comparison of global performance of FedMEKT under different number of EKT steps on UR Fall Detection Dataset

		# of EKT Steps R		
		1	2	3
Acce-Rgb	Acce	57.14	**69.32**	65.64
	Rgb	57.35	**60.21**	58.29
Acce-Depth	Acce	46.44	**69.25**	57.47
	Depth	60.12	**65.85**	56.90
Rgb-Depth	Rgb	80.82	73.81	**83.98**
	Depth	58.22	**70.33**	58.49

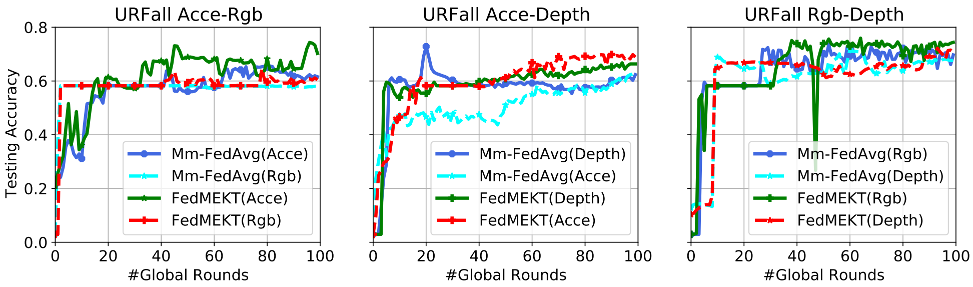

Figure 2: Global Performance of URFall dataset.

Table 5: The comparison of global performance of FedMEKT under different proxy data size on UR Fall Detection Dataset

		Size of Proxy Data D_r		
		100	500	1000
Acce-Rgb	Acce	58.25	57.51	**69.32**
	Rgb	57.34	58.22	**60.21**
Acce-Depth	Acce	58.22	58.90	**69.25**
	Depth	57.35	59.88	**65.85**
Rgb-Depth	Rgb	58.23	64.87	**73.81**
	Depth	64.17	59.59	**70.33**

tion, $R = 3$ achieves the best performance, which means that the R steps may depend on different scenarios. Hence, we may have different EKT steps R numbers on other datasets.

FedMEKT with Different Proxy Data Size Since our proposed method leverage the proxy data to exchange knowledge between server, in this experiment, we validate our proposed method on different sizes of proxy dataset D_r. Table 5 illustrates the effect of proxy data size in the FedMEKT algorithm. As the results show, in most scenarios, we can increase the performance by increasing the size of proxy data D_r. In our experiment in UR Fall Dataset, we utilize 1000 samples for the proxy dataset, which is $1/10$ of the total data, and achieve the best performance.

5 Conclusion

In this work, we proposed a novel multimodal federated learning framework under the semi-supervised setting by developing the multimodal embedding knowledge transfer scheme. Through extensive simulations, our method FedMEKT obtains a more stable and better performance in local and global linear evaluation than the MM-FedAvg algorithm without exchanging model parameters, thus could save more communication costs when the model is large (million of parameters) and guaranteeing better privacy protection. Moreover, ablation studies show the effectiveness of our proposed method. In future work, we will continue extending this research with multimodal fusion design and extend to more number of modalities which can bridge the gap of FL deployment in future personalized applications.

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*, pp. 91–98. Springer, 2014.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J Oedegaard, and Jim Tørresen. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51:1–26, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Abhishek Kumar, Trisha Mittal, and Dinesh Manocha. Mcqa: Multimodal co-attention based network for question answering. *arXiv preprint arXiv:2004.12238*, 2020.

Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine*, 117(3):489–501, 2014.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion. In *First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*, 2018.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

Giuseppe Lugano. Virtual assistants and self-driving cars. In *2017 15th International Conference on ITS Telecommunications (ITST)*, pp. 1–5. IEEE, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Urja Pawar, Donna O'Shea, Susan Rea, and Ruairi O'Reilly. Explainable ai in healthcare. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pp. 1–2. IEEE, 2020.

Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*, 2015.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–7473, 2019.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022.

Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems*, 34:10092–10104, 2021.

Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on iot data. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 43–54. IEEE, 2022.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.