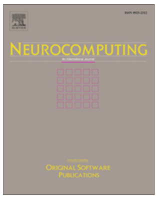
Neurocomputing 480 (2022) 110–118

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A unified framework for multi-modal federated learning ☆

Baochen Xiong [a,e], Xiaoshan Yang [b,c,e], Fan Qi [d,e], Changsheng Xu [b,c,e]

[a] *Henan Institute of Advanced Technology, Zhengzhou University, China*
[b] *NLPR, Institute of Automation, Chinese Academy of Sciences, China*
[c] *School of Artificial Intelligence, University of Chinese Academy of Sciences, China*
[d] *Hefei University of Technology, China*
[e] *Peng Cheng Laboratory, China*

ARTICLE INFO

Article history:
Received 22 March 2021
Revised 26 November 2021
Accepted 17 January 2022
Available online 22 January 2022
Communicated by Zidong Wang

Keywords:
Multi-modal
Federated learning
Co-attention

ABSTRACT

Federated Learning (FL) is a machine learning setting that separates data and protects user privacy. Clients learn global models together without data interaction. However, due to the lack of high-quality labeled data collected from the real world, most of the existing FL methods still rely on single-modal data. In this paper, we consider a new problem of multimodal federated learning. Although multimodal data always benefits from the complementarity of different modalities, it is difficult to solve the multimodal FL problem with traditional FL methods due to the modality discrepancy. Therefore, we propose a unified framework to solve it. In our framework, we use the co-attention mechanism to fuse the complementary information of different modalities. Our enhanced FL algorithm can learn useful global features of different modalities to jointly train common models for all clients. In addition, we use a personalization method based on Model-Agnostic Meta-Learning(MAML) to adapt the final model for each client. Extensive experimental results on multimodal activity recognition tasks demonstrate the effectiveness of the proposed method.

1. Introduction

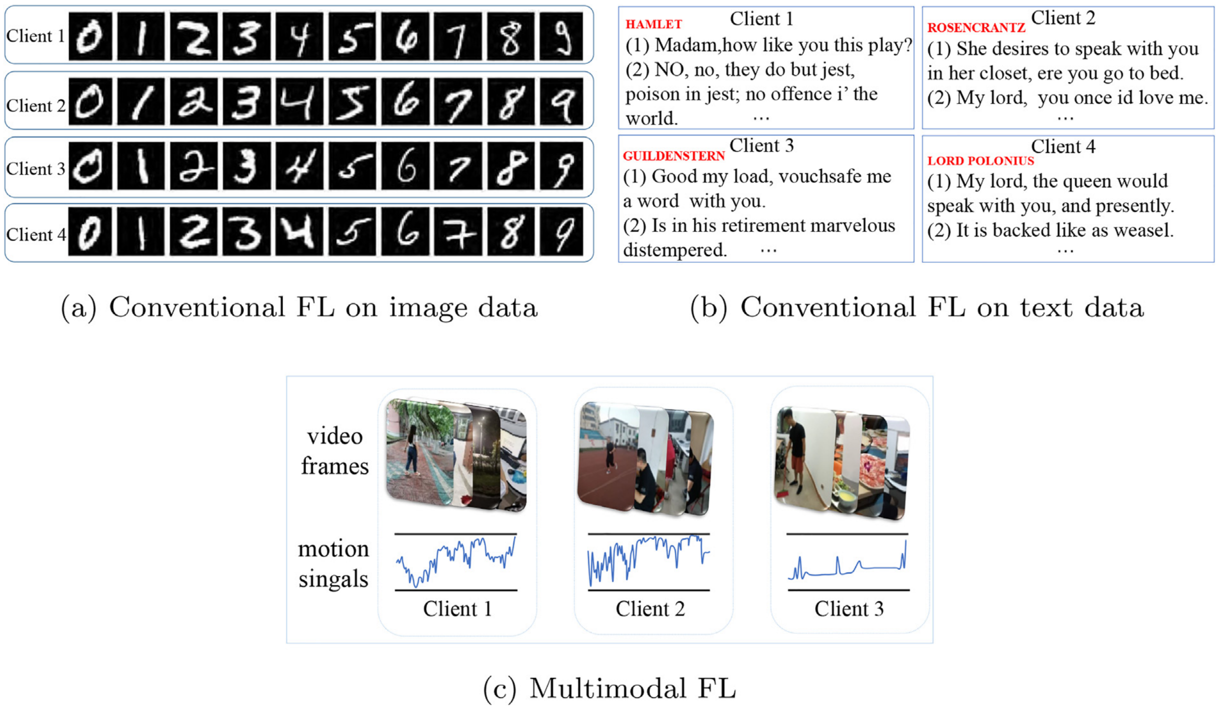
With the development of mobile Internet technology, the emergence of many electronic devices has produced more and more mobile signals. The mass production of information has made people realize the importance of protecting privacy. In the past few years, personal data protection has been continuously strengthened worldwide, including the General Data Protection Regulation (GDPR) [1] implemented by the European Union on May 25, 2018. Analyzing the data collected from mobile devices is usually based on machine learning models, such as deep neural networks. Training such models usually requires a large amount of data, which undoubtedly brings hidden dangers to private security. Due to privacy and communication bandwidth limitations, it is difficult to collect all the data in one central location. To benefit from the data collected and stored locally on client devices, federated learning (FL) is proposed as a distributed model training method that does not exchange raw data, which not only retains the privacy of the information but also saves the communication bandwidth [2–4]. This field has recently aroused great interest in research and application. For example, Google makes extensive use of FL in the Gboard mobile keyboard [5–7].

The early federated learning (FL) algorithms, e.g., FedAvg [8], focus on aggregating the gradients or parameters of local models independently learned from different clients to establish a global model. Fig. 1a shows a common practice of FL in handwritten digit recognition based on the image data of MNIST. Each client is usually assigned with a randomly selected subset of the whole data. Although the early methods are effective to solve the FL task on iid data, where the distribution between different clients is small, they cannot perform well on non-iid data. Fig. 1b shows the application of traditional FL in character prediction, where each client contains the utterances of a specific character in Shakespeare's works. In this case, the performance of conventional FL models is restricted due to the differences in word count and character's context on different clients. More recently, many improved methods [9–11] have been proposed to solve the non-iid problem. For example, PerAvg [9] allows each client to adapt to local data quickly by finding an appropriate initialization and pFedMe [11] uses Moreau envelope function to decouple personalized model optimization from global model learning.

☆ This work was supported by National Key Research and Development Program of China (No. 2018AAA0100604), National Natural Science Foundation of China (No. 61720106006, 62072455, 61721004, U1836220, U1705262, 61872424).

E-mail addresses: bcxiong@yeah.net (B. Xiong), xiaoshan.yang@nlpr.ia.ac.cn (X. Yang), cvaveryqf@gmail.com (F. Qi), csxu@nlpr.ia.ac.cn (C. Xu)

https://doi.org/10.1016/j.neucom.2022.01.063
0925-2312/© 2022 Elsevier B.V. All rights reserved.



(1) Madam,how like you this play?
(2) NO, no, they do but jest, poison in jest; no offence i' the world. …

(1) She desires to speak with you in her closet, ere you go to bed.
(2) My lord, you once id love me. …

(1) Good my load, vouchsafe me a word with you.
(2) Is in his retirement marvelous distempered. …

(1) My lord, the queen would speak with you, and presently.
(2) It is backed like as weasel. …

(a) Conventional FL on image data

(b) Conventional FL on text data

(c) Multimodal FL

Fig. 1. Illustration of the difference between the proposed multimodal FL and the conventional FL. (a) Conventional FL on image data. (b) Conventional FL on text data. (c) Multimodal FL..

For humans, the perception of the surrounding environment always requires multiple modalities, such as visual scenes, sounds, and odors. For machine learning algorithms, many highest-performance models in computer vision and multimedia tasks, such as activity recognition [12], always involve multiple data forms, because multimodal data can provide complementary information, which will significantly improve the performance [13]. However, existing FL methods always focus on single-modal data, e.g., image or text, which may limit their generality in real application. Many devices, e.g., mobile phones, always have different types of sensors, such as cameras, motion sensors and microphones, which can receive different kinds of data (e.g., images and motion signals) for jointly learning local models. Whereas, it is still questionable whether existing FL methods can be directly applied to multimodal data. In this paper, we study the important yet under-explored task of **multi-modal federated learning**, i.e., collaboratively learning a global model on distributed local clients where each of them contains multimodal data.

In the proposed multimodal FL, local clients share the same feature space but have different samples. Therefore, it belongs to Horizontal FL [14]. The major difference from conventional Horizontal FL is that local clients in the proposed multimodal FL share a common multimodal feature space while the conventional Horizontal FL only considers single-modal feature space. Fig. 1c shows a typical example of multimodal FL for activity recognition based on multimodal data. Each client represents a specific user, where the data consists of images and motion information collected by cameras and motion sensors. The multimodal data are collected when the user is doing the predefined daily activities. We can see that the multimodal FL is much more difficult than the conventional single-modal FL. Because, in addition to resolve the distributed model learning problem in FL, the complementarity of different modalities should also be fully considered under the restriction of privacy protection. Moreover, the multimodal data on different clients may also have different distributions which brings extra challenge for distributed model learning.

To overcome the above challenges, we propose a unified multimodal FL framework, which can distributively capture the complementary information between different modalities on local clients. We propose to use co-attention mechanism in multimodal FL to simultaneously fuse the complementary information of different modalities and learn generalized features that are useful for joint training of the general model on all clients. Besides, to adapt to local data on each client under multimodal conditions, the personalized approach using model-agnostic meta-learning (MAML) is adopted to find the initial point shared by all clients, and each client can easily update the general multimodal FL model according to its own loss function. Although the initial model for each client is obtained from the server, the final model for each client is different from the other clients based on its own data. To evaluate the effectiveness of the proposed multimodal FL framework, we apply it to the multimodal activity recognition task and obtain better performances than conventional FL methods on two datasets, i.e., Multimodal Data and MMC-PCL-Activity.

The main contributions of this paper are highlighted as follows.

1. For the first time, we propose a unified framework to address a new task of multimodal federated learning, which needs to collaboratively learn a global model on distributed local clients where each of them contains multimodal data.
2. We propose a co-attention mechanism to simultaneously fuse the complementary information of different modalities and learn generalized features that are useful for joint training of the general model on all clients.
3. We validate the proposed method on two multimodal datasets in activity recognition.

The rest of this paper is organized as follows. In Section 2, we summarize the related work. The proposed multimodal FL framework is described with details in Section 3. Experimental results are reported and analyzed in Section 4. Finally, we conclude the paper with future work in Section 5.

2. Related Work

In this section, we introduce the existing work most relevant to our work, including traditional FL methods, personalized FL, and multimodal representation.

FL and Challenges: One of the first FL algorithms is FedAvg [15], which uses local SGD updates and builds a global model from a subset of clients with non-IID data. In [16,17], the authors introduce some quantitative methods to solve the communication problems of FL. It is well known that communication is one of the main bottlenecks in FL. Because data transmission with wireless client is usually very expensive and unreliable when compared with the connection between the data centres or within the data centre [18]. In [19,8,20–22], the authors mainly study the privacy protection related issues in FL. The server only accepts model parameters passed by the client. This method has significant improvements of practical privacy protection, but it does not guarantee absolute privacy security. For example, a malicious client can check all the messages (including model iterations) received from the server in the round that it participates in and tamper with the training process. In [23–25], the authors propose to perform multiple local optimizations locally and then sends the local model to the server. These methods mainly solve the problem that large-scale training of global models usually cannot be promoted well.

Recently, some works have introduced algorithms for homogeneous settings, intending to sample all client data points from the same probability distribution [26,27]. Some other works have explored the statistical heterogeneity of client data points in FL [28–33]. Different from these existing methods, we focus on the FL task of multi-modal data and give a clear definition of the new problem.

Personalized FL: A global model in FL can be viewed as a "centre point" that all clients agree to satisfy. In contrast, a personalized model is a point where clients follow different directions based on heterogeneous data distribution. In order to solve the personalization problem in FL, various methods have been proposed. [28] proposes a method of mixing global and local models, in which the L2GD algorithm combines the optimization of the local model and the global model. In [34], three personalization methods were proposed: model interpolation, data interpolation and user clustering. Due to privacy protection, user clustering and data interpolation methods require all clients' meta-features, which makes them infeasible in FL. But in [35], the method of model interpolation is used to create adaptive personalized FL algorithm. This method tries to mix the client's local model and global model. [36] proposes a personalization algorithm FedPer based on neural networks, in which the network is divided into a basic layer and a personalization layer. The basic layer is trained on the server, and both types of layers are trained on the client. [37] proposes a federated multitasking framework of MOCHA, which mainly solves system and statistical heterogeneity. For more details about FL, its challenges, and personalization approaches, we refer the readers to comprehensive surveys in [4,18].

The work more related to our research is [5], which requires different predictions between client devices to solve the next character prediction task. [38] proposes a meta-learning algorithm inspired by online convex optimization, which can improve the performance of FedAvg. [39] finds that FedAvg can be interpreted as meta-learning, and proposes a personalized FL method combining FedAvg and Reptile [40]. In particular, the method in [9] is the most relevant work to our research direction. It establishes an initialization meta-model that can be effectively updated after another gradient descent step. Different from this method, our main concern is that the client's model has better performance on local multimodal data. In addition, in our experiments, we demonstrated that [9] is not as effective as our method in multimodal cases.

Multimodal Representation: Sight, hearing, touch, smell, taste, etc., constitute people's perception capabilities. Losing one of the above may result in a substantial decrease in the quality of life. Based on this, the multimodal feature fusion has been widely studied. There are mainly three types of methods for multimodal feature representation in the early days. (1) Feature subspace: In [41,42], Canonical Correlation Analysis (CCA) is the most widely used method to deal with multiple modalities in the feature subspace. CCA tends to find a basis vector that can represent different modalities. The basis vector is obtained by two linear transformations, each of which corresponds to the modality of the variable. The correlation between the projection vectors is mutually maximized. (2) Semantic integration: In [43], a feature space is defined to contain semantic features, and a vector of posterior probability is used to represent each image. In [42], the semantics of each image are represented by the relevant spatial structure acquired based on CCA. (3) The kernel method: In [44], semi-supervised learning is introduced into multi-modal processing, allowing images without labels to obtain information from similar labels. The multi-kernel learning (MKL) framework assigns a kernel to the image and label, respectively.

Recently, multimodal deep learning has also received more attention [45]. A popular multimodal network architecture consists of several separate neural layers and a common hidden layer. Each modality is first fed into each layer, then projected into a joint feature space [46,47] through a common hidden layer. The joint representation will be further processed for prediction. In [48], a stacked denoising autoencoder is proposed to process multimodal data and fuse different modalities in an unsupervised form. Some activity recognition works improve recognition accuracy by combining different information from video modality and sensor signal modality. In the multimodal activity recognition study, [49] uses a multiflow extractor based on two-stage multimodal fusion technology to extract features. [50] uses several middle-level representations around the topic as necessary clues to infer activity classes. The handcrafted feature represents the content, location, and manner in which a subject interacts with a background. However, this method only fuses posterior probabilities based on various morphologies, ignoring the interaction between the extracting middle-level concepts. In [51], a multimodal data fusion framework based on LSTM is proposed. The hidden state of LSTM identifies the correlation between video features and sensor features. In [52], activity recognition is performed by connecting video and sensor signal features and using heart rate signals as self-monitoring information to enhance model performance. [53] proposes an enhanced learning framework to select video or sensor signal modalities to predict activity in different scenarios, which reduces computational effort and improves accuracy. However, the above methods only directly combine multimodal features or selectively use single-modal features in different scenarios, and they do not explore the complementarity of multimodal data under the FL framework.

3. The Proposed Method

In this section, we first give the definition of the multimodal FL problem, and then describe the proposed FL framework under multimodal conditions in detail.

3.1. Problem Definition

Suppose there is a set of multimodal data $\mathscr{D}$ (we consider two modalities in this work) distributed over n clients such that

$\mathscr{D} = \bigcup_{i=1}^{n} \mathscr{D}_i$, where $\mathscr{D}_i$ is the dataset located at client i. Each sample of client i is defined as $\left(X_i^j, Y_i^j\right) \in \mathscr{D}_i$, where X_i^j has two different modalities, Y_i^j is the corresponding sample label. The multimodal federated learning aims to learn a multimodal classification model that can correctly predict the labels of local multimodal samples. All clients need to collaborate to train the model without exchanging multimodal data.

3.2. Overview

In this section, we will overview the proposed multimodal federated learning framework (MMFed). Following the standard federated learning algorithms, such as Federated Averaging (FedAvg) [15], our framework includes operations on both the server and client: the client trains the model locally and uploads it to the server, and the server aggregates the model updates from the client through a weighted average. The proposed MMFed enables individual clients to conduct FL on multimodal samples. In our multi-modal FL framework, since each modality may contain redundant features, the most important part of the feature is dynamically selected through the attention mechanism to better capture the correlation between modalities. Therefore, different from existing FL methods that only upload the parameters of the classifer to the server, we need to upload the parameters of both the classifer and the attention module after each client updating. After the server aggregation, different clients receive the shared parameters of the attention module and the classifier. Fig. 2 depicts the overall framework of our MMFed. If we define g_i as the loss corresponding to client i, the multimodal federated learning can be formulated as solving the following optimization problem:

$$\min_{(w,v)} g(w, v) := \frac{1}{n}\sum_{i=1}^{n} g_i(w, v), \tag{1}$$

where w and v denote parameters of the attention module and the classifier, respectively. g_i represents the expected loss of the i-th client on local multimodal data,

$$g_i(w, v) := \mathbb{E}_{\left(X_i^j, Y_i^j\right) \in \mathscr{D}_i}\left[l_i\left(w, v; X_i^j, Y_i^j\right)\right], \tag{2}$$

where $l_i\left(w, v; X_i^j, Y_i^j\right)$ measures the error of the model in predicting the true label Y_i^j. As discussed in the introduction, the data for each client is multimodal.

We describe the training process as follows: (1) At the beginning of a new round of training, the server selects n clients to participate in the training, and the global model is sent to each client. (2) Each client replaces the model with a global model. (3) Clients input different modal features obtained by backbone into the attention module, and finally transfer the fused features into the classifier. (4) Clients train the model on local data using random gradient descent, and upload the trained parameters of the attention module and the classifier to the server. (5) The server aggregates parameters from different clients and updates them to get a new global model.

To explore the role of the attention module, we tried three different schemes for uploading parameters from the client to the server: sending only parameters of the classifier, sending only parameters of the attention module, and sending both of them. More details are illustrated in the experiment. In the following, Section 3.3 will illustrate the attention module for multimodal fusion. After the consideration of the heterogeneity of the data in

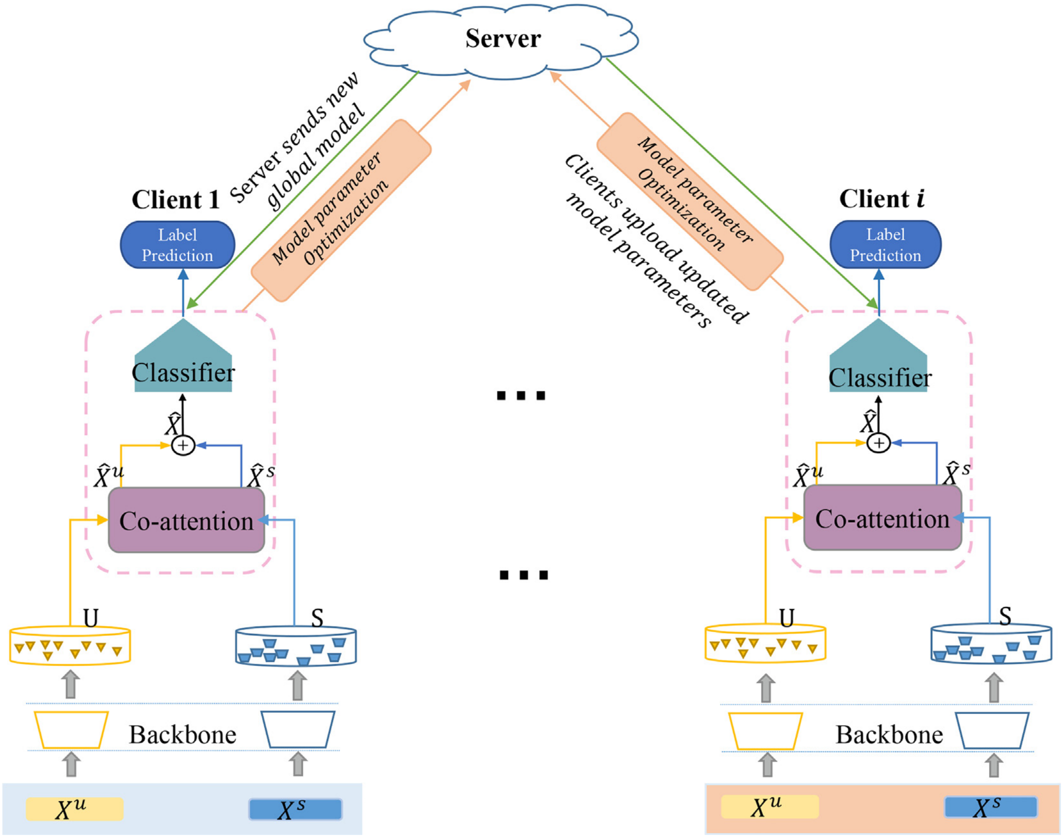

Server
Server sends new global model
Model parameter Optimization
Model parameter Optimization
Clients upload updated model parameters
Client 1
Client i
Label Prediction
Classifier
$\hat{X}$
$\hat{X}^u$
$\hat{X}^s$
Co-attention
U
S
Backbone
X^u
X^s
...

Fig. 2. Overview of the proposed multimodal federated learning framework (MMFed).

different clients, we introduce the personalization module in Section 3.4 to make each local model more suitable for the client data.

3.3. Multimodal Co-Attention

Attention module is an important part of multimodal fusion in our MMFed. It takes advantage of the internal relationship between different modalities. In this section, we will show how to explore the correlation between the features of different modalities. For two features **U** and **S** extracted from different modalities X^u and X^s of a sample X, we define $\mathbf{U} = [u_1, \ldots, u_{n_u}]^\top \in \mathbb{R}^{n_u \times d_u}$ and $\mathbf{S} = [s_1, \ldots, s_{n_s}]^\top \in \mathbb{R}^{n_s \times d_s}$. Here, n_u(or n_s) is the number of feature vectors. d_u(or d_s) denotes the size of the feature vector. Specifically, to take advantage of the complementarity and redundancy between different modalities, given the features **U** and **S**, the following operations are performed through co-attention [54]:

$$\mathbf{M} = \tanh\left(\mathbf{U}\mathbf{W}_{ca}^{b}\mathbf{S}^{\mathrm{T}}\right) \tag{3}$$

$$\mathbf{A}^u = \tanh\left(\mathbf{W}_{ca}^{u}\mathbf{U}^{\mathrm{T}} + \left(\mathbf{W}_{ca}^{s}\mathbf{S}^{\mathrm{T}}\right)\mathbf{M}\right) \tag{4}$$

$$\mathbf{A}^s = \tanh\left(\mathbf{W}_{ca}^{s}\mathbf{S}^{\mathrm{T}} + \left(\mathbf{W}_{ca}^{u}\mathbf{U}^{\mathrm{T}}\right)\mathbf{M}^{\mathrm{T}}\right) \tag{5}$$

$$\mathbf{a}^u = \mathrm{Softmax}\left(\mathbf{w}_{ca}^{u\mathrm{T}}\mathbf{A}^u\right) \tag{6}$$

$$\mathbf{a}^s = \mathrm{Softmax}\left(\mathbf{w}_{ca}^{s}{}^{\mathrm{T}}\mathbf{A}^s\right) \tag{7}$$

$\mathbf{M} \in \mathbb{R}^{n_u \times n_s}$ is the affinity matrix of **U** and **S**, which is used to convert the focus space of one modality feature into the focus space of another modality feature. $\mathrm{a}^u \in \mathbb{R}_{\mathbb{u}}^{\mathbb{n}}, \mathrm{a}^s \in \mathbb{R}_{\mathbb{s}}^{\mathbb{n}}$ are normalized attention weights of the concept features in **U** and **S**, respectively. $\mathrm{W}_{ca}^{b} \in \mathbb{R}^{d_u \times d_s}, \mathrm{W}_{ca}^{u} \in \mathbb{R}^{h_c \times d_u}, \mathrm{W}_{ca}^{s} \in \mathbb{R}^{h_c \times d_s}$, and $\mathbf{w}_{ca}^{u}, \mathbf{w}_{ca}^{s} \in \mathbb{R}_{\mathbb{c}}^{\mathbb{h}}$ are trainable parameters. In particular, h_c is the dimension of different modal attention spaces. By calculating the attention score, we can get the weighted representation $\widehat{X}^u = \mathrm{diag}(\mathbf{a}^u)\mathbf{U}$ and $\widehat{X}^s = \mathrm{diag}(\mathbf{a}^s)\mathbf{S}$ of the two features. The weighted representations of the resulting different features are fused by addition, which results in the fused feature $\widehat{X}$.

3.4. Personalized Optimization

In this section, we use the basic ideas behind the Model-Agnostic Meta-Learning (MAML) [9] to optimize parameters of both the attention module and the classifier to solve client personalization under multimodal conditions. Assuming that each client takes the initial point and updates the model parameters using a gradient descent step based on its own loss function, then Eq. (1) changes to

$$\min_{w,v} G(w, v) := \frac{1}{n}\sum_{i=1}^{n} g_i((w, v) - \alpha\nabla g_i(w, v)) \tag{8}$$

where α represents the learning rate. The above expression does not change the advantages of traditional FL. Moreover, it allows the client to update the model according to its own data to obtain a personalized model. Given the distribution of heterogeneous data on the client side, only solving Eq. (1) is not an ideal choice. Because it returns a single model, even local gradients that take several steps will not adjust rapidly to each client's local data. Instead, by resolving Eq. (8), we can get an initial model (Meta-model), which is trained to produce a model for each client that depends on local data through a one-step local gradient.

For the model parameter (w, v), our goal is to find the optimal solution of Eq. (8). Each client updates model parameters based on its own loss function. First, we rewrite the function in Eq. (8) into the average value of meta-functions $G_1, \ldots, G_n$, and the meta-function G_i for client i is defined as:

$$G_i(w, v) := g_i((w, v) - \alpha\nabla g_i(w, v)). \tag{9}$$

The gradient ∇G_i of the local function on client i is given by

$$\nabla G_i(w, v) = \left(I - \alpha\nabla^2 g_i(w, v)\right)\nabla g_i((w, v) - \alpha\nabla g_i(w, v)). \tag{10}$$

Every round of gradient calculation on all data will affect the communication efficiency. Therefore, we select a batch of data $\mathcal{B}_i$ from $\mathcal{D}_i$ and get its unbiased estimate $\tilde{\nabla} g_i(w, v, \mathcal{B}_i)$ given by

$$\tilde{\nabla} g_i(w, v, \mathcal{B}_i) := \frac{1}{|\mathcal{B}_i|}\sum_{(X_i^j, Y_i^j) \in \mathcal{B}_i} \nabla l_i\left(w, v; X_i^j, Y_i^j\right), \tag{11}$$

So we express $\nabla^2 g_i(w, v)$ in Eq. (10) with its unbiased estimate $\tilde{\nabla}^2 g_i(w, v, \mathcal{B}_i)$.

Similar to the global k-th iteration of FedAvg, the server sends the k-th global model (w_k, v_k) to each client, and each client performs τ local stochastic gradient descent relative to G_i. Especially a new local sequence $\left\{(w, v)_{k+1,t}^{i}\right\}_{t=0}^{\tau}$ is generated after the client is updated, $(w, v)_{k+1,0}^{i} = (w, v)_k$ and $\tau \geqslant t \geqslant 1$,

$$(w, v)_{k+1,t}^{i} = (w, v)_{k+1,t-1}^{i} - \beta\tilde{\nabla} G_i\left((w, v)_{k+1,t-1}^{i}\right), \tag{12}$$

where β is the local learning rate (stepsize) and $\tilde{\nabla} G_i\left((w, v)_{k+1,t-1}^{i}\right)$ is an estimate of $\nabla G_i\left((w, v)_{k+1,t-1}^{i}\right)$ in Eq. (10).

After evaluating the local update $(w, v)_{k+1,\tau}^{i}$, the user sends it to the server, and then the server updates its global model by averaging the received models, i.e., $(w, v)_{k+1} = \frac{1}{n}\sum_{i \in n}(w, v)_{k+1,\tau}^{i}$.

4. Experiment

In this section, we will evaluate the proposed unified multimodal federated learning method in multimodal activity recognition. We first describe the two activity recognition datasets used in the experiment. Then we describe the details of the experiment and analyze the results.

4.1. Dataset

Multimodal Data: Multimodal Data [49] is an earlier published dataset for egocentric activity recognition. It contains 20 life-logging activities. It comprises of 200 egocentric videos augmented with sensor signals. In our experiment, we use videos with three-axis accelerations and three-axis gyroscopes as sensor signals. We split each video into multiple samples and use single activity labels. Finally we get 600 samples. To meet the requirements of FL, we suppose there are 10 clients and distribute all samples equally. Each client uses 42 samples as training and 18 samples as testing. Fig. 3a shows some sample examples in the Multimodal Data dataset.

MMC-PCL-Activity: This is a new dataset collected by ourselves. It contains sensor data and images for daily activities. The sensor data and images are collected from 14 users, and each user collects information of 15 daily activities for 31 days. In particular, sensor data includes three-axis accelerations, three-axis gyroscopes and heart rate. Sensor signals are collected through a mobile phone, which is placed in the right pocket of the user. A wrist band provides the corresponding heart rate data. In [52], heart rate is used

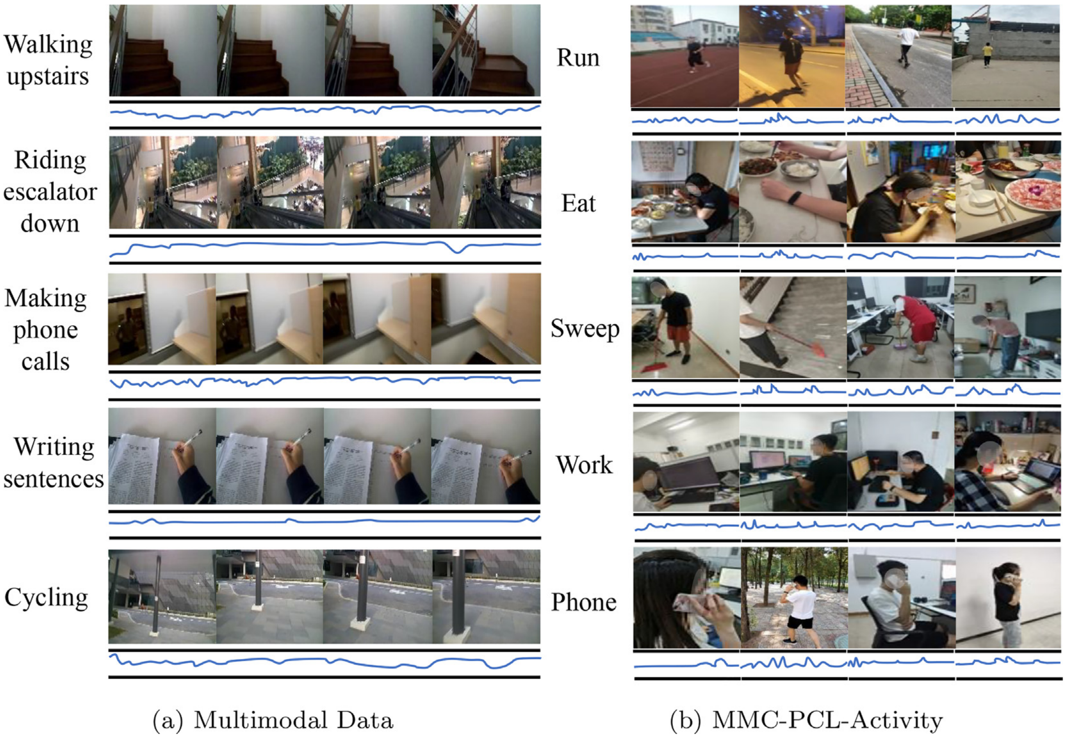


Fig. 3. Examples of the datasets (video frames and accelerometer singals). (a) Multimodal Data. (b) MMC-PCL-Activity..

as self-supervised information for activity recognition. In our experiments, we only use three-axis accelerations and three-axis gyroscopes as sensor signals. Since the original data is large-scale, we selected the first 7 days of data from the top 10 collectors. Unlike Multimodal Data, the collection method of MMC-PCL-Activity is closer to the real application of FL, and we deem each user as a client in our experiment. The collection time of sensor data varies from 30 s to 60 s, and there are individual discrepancies among different users. We matched the sensor data to the images and used the first 5 days of data from each user/client for training and the next2 days for testing. Fig. 3b shows some examples of the MMC-PCL-Activity dataset.

4.2. Implementation Details

In the experiment, the two datasets are treated differently. For the dataset of Multimodal Data with two modalities, video signal and sensor signal, we use Inception-V3 [55] to extract video frame features and extract hand-craft features from sensor signals as in [56]. For the MMC-PCL-Activity dataset with image and sensor signals, we used Resnet50 [57] and LSTM [51] to extract the features of the image and sensor, respectively. The dimension d_u(or d_s) of feature vectors is set to 2048. The dimension h_c of multimodal attention space is set to 128. We use a two-layer deep neural network (DNN) with hidden layer size of 2048. Each layer has a ReLU activation and a softmax layer is used at the end of the deep neural network.All experiments are conducted using PyTorch [58] version 1.4.0. For all baselines, we use publicly released code. For all approaches, we use the SGD optimizer with a global learning rate $\alpha = 0.01$. The SGD weight decay is set to 0.00001 and the momentum is set to 0.9. The number of local computation rounds τ is set to 50. The number of global rounds k is set to 300. For MMFed, we set the local learning rate β to 0.01.

4.3. Evaluation

We use accuracy metrics to evaluate the performance of the proposed MMFed framework on the above two datasets. We also perform several ablation studies to show the effectiveness of each component of the proposed framework. We have implemented our approach and several baseline methods based on the same FL experiment settings. FedAvg (vision) and FedAvg (sensor) are the classical FL methods [8] applied on the single-modal video/image data and sensor data, respectively. A single-modal baseline is important, because most activity recognition tasks only consider single-mode data. FedAvg(union) [8] is performed by fusing the single-modal features of video/image data and sensor data with sum-pooling. Similarly, we did the same experiment for PerAvg [9] and pFedMe [11]. PerAvg [9] is based on the MAML framework, which allows each client to adapt to local data quickly by finding an appropriate initialization model. Finally, we also compare our method with pFedMe [11] which uses the Moreau envelope function to decompose personalized model optimization from global model learning to accommodate the problem of statistical diversity.

4.3.1. Parameter Analysis

Several hyperparameters, such as τ, B, and β, play an important role in our method. To understand how they affect the performance of the proposed MMFed, we conduct various experiments on the MMC-PCL-Activity dataset.

Effects of local computation rounds τ: the larger value of τ tends to allow the client to perform more local computations. This reduces the number of communications between the server and the client. Therefore, we use multiple values of τ to observe the performance of our method, and the results are shown in Fig. 4. The results show that a larger τ is conducive to improve the performance. But computing and communication cost also need to be considered. To balance this trade-off, we fix $\tau = 50$.

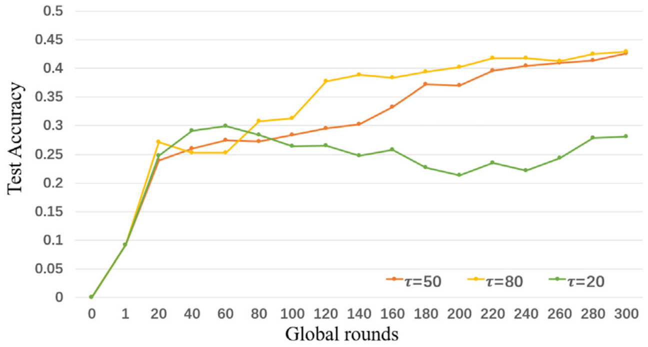


Fig. 4. Effect of τ on the performance of our method.

Table 1
Performance comparison of different methods on two datasets.

Model	MMC-PCL-Activity	Multimodal Data
	Accuracy	Accuracy
FedAvg(vision)	31.5	44.9
FedAvg(sensor)	35.0	36.1
FedAvg(union) [8]	36.2	48.4
PerAvg(vision)	32.2	45.2
PerAvg(sensor)	32.4	37.5
PerAvg(union) [9]	35.2	47.7
pFedMe(vision)	30.0	43.3
pFedMe(sensor)	32.7	40.5
pFedMe(union) [11]	35.9	46.9
MMFed	**44.1**	**53.2**

Effects of Batch Size B: In Fig. 5, as the size of the batch size increases, the accuracy of the test set increases quickly. However, a huge B will not only reduce the performance of MMFed, but also requires higher computations at the local users. So in the experiment, the value of B is configured as a constant value equal to 110.

Effects of local learning rate β: Fig. 6 illustrates how β affects the accuracy of the global model. The results show that the performance of the our model decreases when the β value is small. However, it is also important to carefully adjust β to prevent the model's gradient from diverging. We finally set the value of β to 0.01 to stabilize the training of our model.

4.3.2. Comparison to Baseline

The results on two different datasets are shown in Table 1. The results show that FedAvg, PerAvg, and pFedMe have low accuracy on single-modal data. Activity recognition using visual information alone is difficult. Because different activities may be performed in the same scene. The results also indicate that it is difficult to use sensor data alone to complete the FL task. Because some activities, such as working on a computer, reading, watching TV, are ambiguous from the perspective of motion feature. By analyzing the results, single-modal FedAvg, PerAvg, and pFedMe do not perform well in activity recognition. The performances of FedAvg (union), PerAvg (union), and pFedMe(union) methods are better than that of single-modal, which indicates that the fusion of multimodal data is important in FL tasks. It is worth noting that the Multimodal Data dataset has better performance than the MMC-PCL-Activity. Because the data distribution among different clients on the MMC-PCL-Activity dataset is extremely unbalanced. At the same time, due to the uncertain collection time, the data of each client will add additional noise to the sensor data. On the MMC-PCL-Activity dataset, Table 1 shows that the our method improves the performance by 8.9%, 8.2%, and 7.9% when compared with PerAvg, pFedMe, and FedAvg under the same experimental settings. On the Multimodal data dataset, our method improves the performance by 5.5%, 6.3%, and 4.8% when compared with PerAvg, pFedMe, and FedAvg. The results show that our method is more effective than existing methods in resolving the multimodal FL task.

In Fig. 7, we show detailed results of our method on each client. For the MMC-PCL-Activity dataset, the first client has the highest accuracy. Because the data in the first client is more evenly distributed among different classes. Similarly, the second and fourth clients have good recognition accuracy. The seventh client only has the recognition accuracy of about 39%. This is due to that the seventh client has less data. These results indicate that too little data and unevenly distributed data may lead to poor performance. For the Multimodal Data dataset, the accuracies of different clients are close to each other. This is due to that the same number of sam-

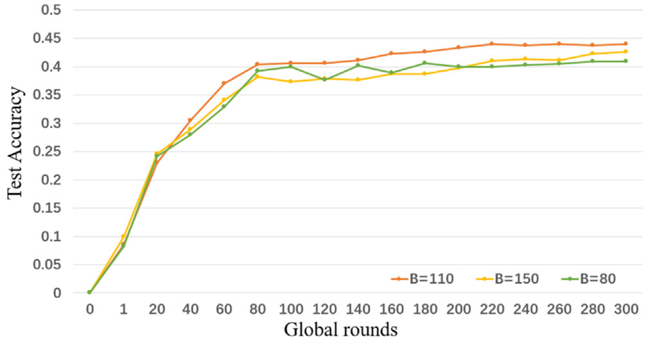


Fig. 5. Effect of B on the performance of our method.

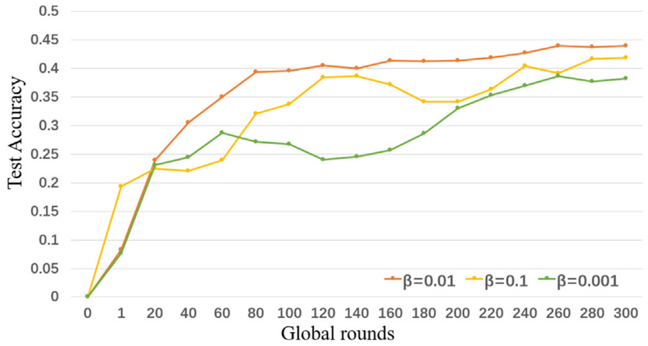


Fig. 6. Effect of β on the performance of our method.

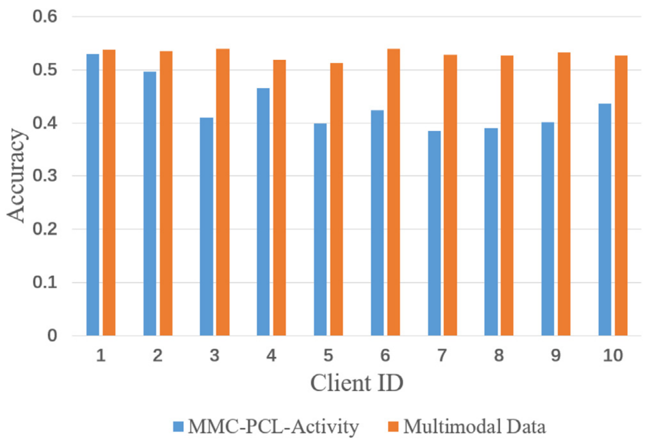


Fig. 7. Accuracy results of our method in each client.

Table 2
Ablation studies of the proposed method on two datasets.

	MMC-PCL-Activity	Multimodal Data
	Accuracy	Accuracy
MMFed A^-	35.7	46.2
MMFed O^-	39.3	50.3
MMFed P^-	40.2	49.5
MMFed C^-	43.0	50.8
MMFed	**44.1**	**53.2**

Table 3
The average training time per global round.

Methods	MMC-PCL-Activity	Multimodal Data
FedAvg	1.2 min	19s
pFedMe	1.0 min	17s
PerAvg	1.6 min	25s
MMFed	1.7 min	22s

ples are allocated in each client. In conclusion, we identify two key factors that affect activity recognition performance under FL: the number of samples in each client and the data balance.

4.3.3. Ablation Study

In this section, we conduct ablation experiments on two datasets to evaluate the effectiveness of each module in our framework. MMFed A^- is a variant of the proposed method without multimodal co-attention. MMFed O^- is a variant of the proposed method without personalized optimization module. MMFed P^- is a variant of the proposed method in which the client sends only classifier parameters to the server. Similarly, MMFed C^- is a variant of the proposed method in which the client sends only attention module parameters to the server. MMFed is the whole framework of the proposed unified multimodal federated learning. For the MMC-PCL-Activity dataset, Table 2 shows that our method improves the performance by 8.4%, 4.8%, 3.9%, and 1.1% when compared with MMFed A^-, MMFed O^-, MMFed P^-, and MMFed C^-, respectively. On the Multimodal Data dataset, the performance is improved by 7.0%, 2.9%, 3.7%, and 2.4%, respectively. This means that all three modules are indispensable in the proposed multimodal FL framework.

4.3.4. Computation Cost

We compare the training time of our model with existing FL methods. All experiments are conducted on a linux server with NVIDIA TITAN RTX and Intel(R) Xeon(R) CPU E5-2620. The average training time per global round is shown in Table 3. The training of MMFed is slower than FedAvg. Whereas, considering that the MMFed communicates the parameters of two modules to the server: the attention module and the classifier, the computation overhead of the MMFed is acceptable, especially on multimodal data.

5. Conclusion

Federated learning has emerged as a promising approach for solving data silos in many areas such as healthcare, person re-identification, and landmark classification. In this paper, we first present the definition of a new task of multimodal federation learning. To solve the key challenge of multimodal FL, i.e., distributively capturing information complementarity among different modalities, we propose a simple and effective multimodal federated learning framework (MMFed). Co-attention is used to capture complementarity information between different modalities on each local client. We also use the personalized optimization to adapt the co-attention module and the classifier module on each client. Extensive experiments show that the MMFed obtains state-of-the-art performance on two multimodal activity recognition datasets. In future work, we will consider more practical application scenarios, e.g., the training data of each local client is partially labeled.

CRediT authorship contribution statement

Baochen Xiong: Conceptualization, Methodology. **Xiaoshan Yang:** Writing - original draft. **Fan Qi:** Writing - review & editing. **Changsheng Xu:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Alekh, EU general data protection regulation: A gentle introduction, CoRR abs/1806.03253..
[2] J. Park, S. Samarakoon, M. Bennis, M. Debbah, Wireless network intelligence at the edge, Proc. IEEE 107 (11) (2019) 2204–2239.
[3] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (2019) 12:1–12:19. .
[4] T. Li, A.K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, CoRR abs/1908.07873..
[5] A. Hard, K. Rao, R. Mathews, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, D. Ramage, Federated learning for mobile keyboard prediction, CoRR abs/1811.03604..
[6] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, F. Beaufays, Applied federated learning: Improving google keyboard query suggestions, CoRR abs/1812.02903..
[7] S. Ramaswamy, R. Mathews, K. Rao, F. Beaufays, Federated learning for emoji prediction in a mobile keyboard, CoRR abs/1906.04329..
[8] H.B. McMahan, D. Ramage, K. Talwar, L. Zhang, Learning differentially private recurrent language models, in: ICLR (Poster), OpenReview.net, 2018..
[9] A. Fallah, A. Mokhtari, A.E. Ozdaglar, Personalized federated learning: A meta-learning approach, CoRR abs/2002.07948..
[10] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, in: MLSys, mlsys.org, 2020. .
[11] C.T. Dinh, N.H. Tran, T.D. Nguyen, Personalized federated learning with moreau envelopes, in: NeurIPS, 2020. .
[12] V. Mezaris, A. Scherp, R.C. Jain, M.S. Kankanhalli, Real-life events in multimedia: detection, representation, retrieval, and applications, Multim. Tools Appl. 70 (1) (2014) 1–6.
[13] T. Baltrusaitis, C. Ahuja, L. Morency, Multimodal machine learning: A survey and taxonomy, CoRR abs/1705.09406. .
[14] H. Zhu, J. Xu, S. Liu, Y. Jin, Federated learning on non-iid data: A survey, Neurocomputing 465 (2021) 371–390.
[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: AISTATS, Vol. 54 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 1273–1282..
[16] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, R. Pedarsani, Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization, in: AISTATS, Vol. 108 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 2021–2031. .
[17] X. Dai, X. Yan, K. Zhou, H. Yang, K.K.W. Ng, J. Cheng, Y. Fan, Hyper-sphere quantization: Communication-efficient SGD for federated learning, CoRR abs/1911.04655..
[18] P. Kairouz, H.B. McMahan, B. Avent, Advances and open problems in federated learning, CoRR abs/1912.04977..
[19] J.C. Duchi, M.I. Jordan, M.J. Wainwright, Privacy aware learning, J. ACM 61 (6) (2014) 38:1–38:57. .
[20] W. Zhu, P. Kairouz, B. McMahan, H. Sun, W. Li, Federated heavy hitters discovery with differential privacy, in: AISTATS, Vol. 108 of Proceedings of Machine Learning Research PMLR, 2020, pp. 3837–3847.
[21] N. Agarwal, A.T. Suresh, F.X. Yu, S. Kumar, B. McMahan, cpsgd: Communication-efficient and differentially-private distributed SGD, in: NeurIPS, 2018, pp. 7575–7586. .
[22] Z. Li, V. Sharma, S.P. Mohanty, Preserving data privacy via federated learning: Challenges and solutions, IEEE Consumer Electron. Mag. 9 (3) (2020) 8–16.
[23] J. Wang, G. Joshi, Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms, CoRR abs/1808.07576..
[24] T. Lin, S.U. Stich, K.K. Patel, M. Jaggi, Don't use large mini-batches, use local SGD, in: ICLR, OpenReview.net, 2020. .

[25] S.U. Stich, Local SGD converges fast and communicates little, in: ICLR (Poster), OpenReview.net, 2019. .
[26] F. Zhou, G. Cong, On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization, in: IJCAI, ijcai.org, 2018, pp. 3219–3227. .
[27] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, V. Chandra, Federated learning with non-iid data, CoRR abs/1806.00582..
[28] F. Hanzely, P. Richtárik, Federated learning of a mixture of global and local models, CoRR abs/2002.05516..
[29] A.K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, V. Smith, On the convergence of federated optimization in heterogeneous networks, CoRR abs/1812.06127..
[30] S.P. Karimireddy, S. Kale, M. Mohri, S.J. Reddi, S.U. Stich, A.T. Suresh, SCAFFOLD: stochastic controlled averaging for on-device federated learning, CoRR abs/1910.06378..
[31] F. Haddadpour, M. Mahdavi, On the convergence of local descent methods in federated learning, CoRR abs/1910.14425..
[32] A. Khaled, K. Mishchenko, P. Richtárik, Tighter theory for local SGD on identical and heterogeneous data, in: AISTATS, Vol. 108 of Proceedings of Machine Learning Research PMLR, 2020, pp. 4519–4529.
[33] X. Li, K. Huang, W. Yang, S. Wang, Z. Zhang, On the convergence of fedavg on non-iid data, in: ICLR, OpenReview.net, 2020. .
[34] Y. Mansour, M. Mohri, J. Ro, A.T. Suresh, Three approaches for personalization with applications to federated learning, CoRR abs/2002.10619..
[35] Y. Deng, M.M. Kamani, M. Mahdavi, Adaptive personalized federated learning, CoRR abs/2003.13461..
[36] M.G. Arivazhagan, V. Aggarwal, A.K. Singh, S. Choudhary, Federated learning with personalization layers, CoRR abs/1912.00818..
[37] V. Smith, C. Chiang, M. Sanjabi, A.S. Talwalkar, Federated multi-task learning, in: NIPS, 2017, pp. 4424–4434. .
[38] M. Khodak, M. Balcan, A.S. Talwalkar, Adaptive gradient-based meta-learning methods, in: NeurIPS, 2019, pp. 5915–5926. .
[39] Y. Jiang, J. Konecný, K. Rush, S. Kannan, Improving federated learning personalization via model agnostic meta learning, CoRR abs/1909.12488..
[40] A. Nichol, J. Achiam, J. Schulman, On first-order meta-learning algorithms, CoRR abs/1803.02999..
[41] D.R. Hardoon, S. Szedmák, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[42] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM Multimedia, ACM, 2010, pp. 251–260.
[43] N. Rasiwasia, P.J. Moreno, N. Vasconcelos, Bridging the gap: Query by semantic example, IEEE Trans. Multim. 9 (5) (2007) 923–938.
[44] M. Guillaumin, J.J. Verbeek, C. Schmid, Multimodal semi-supervised learning for image classification, in: CVPR, IEEE Computer Society, 2010, pp. 902–909. .
[45] T. Baltrusaitis, C. Ahuja, L. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.
[46] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: visual question answering, in: ICCV, IEEE Computer Society, 2015, pp. 2425–2433..
[47] Y. Mroueh, E. Marcheret, V. Goel, Deep multimodal learning for audio-visual speech recognition, in: ICASSP, IEEE, 2015, pp. 2130–2134.
[48] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: ICML, Omnipress, 2011, pp. 689–696..
[49] S. Song, V. Chandrasekhar, B. Mandal, L. Li, J. Lim, G.S. Babu, P.P. San, N. Cheung, Multimodal multi-stream deep learning for egocentric activity recognition, in: CVPR Workshops, IEEE Computer Society, 2016, pp. 378–385. .
[50] P. Hsieh, Y. Lin, Y. Chen, W.H. Hsu, Egocentric activity recognition by leveraging multiple mid-level representations, in: ICME, IEEE Computer Society, 2016, pp. 1–6. .
[51] E.A. Bernal, X. Yang, Q. Li, J. Kumar, S. Madhvanath, P. Ramesh, R. Bala, Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors, IEEE Trans. Multim. 20 (1) (2018) 107–118.
[52] K. Nakamura, S. Yeung, A. Alahi, L. Fei-Fei, Jointly learning energy expenditures and activities using egocentric multimodal signals, in: CVPR, IEEE Computer Society, 2017, pp. 6817–6826. .
[53] R. Possas, S.M. Pinto-Caceres, F. Ramos, Egocentric activity recognition on a budget, in: CVPR, IEEE Computer Society, 2018, pp. 5967–5976. .
[54] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: NIPS, 2016, pp. 289–297. .
[55] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, IEEE Computer Society, 2016, pp. 2818–2826. .
[56] M. Ma, H. Fan, K.M. Kitani, Going deeper into first-person activity recognition, in: CVPR, IEEE Computer Society, 2016, pp. 1894–1903. .
[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, IEEE Computer Society, 2016, pp. 770–778. .
[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: NeurIPS, 2019, pp. 8024–8035.

Baochen Xiong received the bachelor’s degree in Software engineering from Zhengzhou University. He is currently pursuing the master’s degree at Zhengzhou University.

Xiaoshan Yang received the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include multimedia analysis and computer vision.

Fan Qi received the Ph.D. degree in Computer Science and Technology from Hefei University of Technology, in 2021. She is currently an Associate Professor at Tianjin University of Technology. Her research interests include multimedia sentiment analysis and computer vision.

Changsheng Xu (M’97-SM’99-F’14) is a Professor in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. He has hold 50 granted/pending patents and published over 400 refereed research papers in these areas. Dr. Xu has served as associate editor, guest editor, general chair, program chair, area/track chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops, including IEEE Trans. on Multimedia, ACM Trans. on Multimedia Computing, Communications and Applications and ACM Multimedia conference. He is IEEE Fellow, IAPR Fellow and ACM Distinguished Scientist.