Multimodal federated learning: Concept, methods, applications and future directions

Wei Huang [a], Dexian Wang [b], Xiaocao Ouyang [c], Jihong Wan [d], Jia Liu [e,*], Tianrui Li [c]

[a] College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[b] School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China
[c] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[d] School of Computer Science and Technology, Guangdong University of Technology, Guangdong 510006, China
[e] School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

ARTICLE INFO

ABSTRACT

Multimodal learning mines and analyzes multimodal data in reality to better understand and appreciate the world around people. However, how to exploit this rich multimodal data without violating user privacy is a key issue. Federated learning is a privacy-conscious alternative to centralized machine learning, therefore many researchers have combined federated learning with multimodal learning to break down data barriers for the purpose of jointly leveraging multiple modal data from different clients for modeling. In order to provide a systematic summarize of multimodal federated learning, this paper describes the basic mode of multimodal federated learning, multimodal fusion based on federated learning, multimodal federated learning optimization and multimodal federated learning application, and introduces each type of multimodal federated learning methods in detail. Finally, the future research trends of multimodal federated learning are discussed and analyzed, mainly including the optimization of multimodal federated learning, privacy-preserving techniques for multimodal federated learning, multimodal federated few-shot learning & multimodal federated semi-supervised learning, and data and knowledge-driven multimodal federated learning.

1. Introduction

In recent years, with the rapid development of the Internet of Things and Internet technologies, which have accumulated huge amounts of data, artificial intelligence is also developing from unimodal intelligent information processing, such as text, speech, and vision, to multimodal general artificial intelligence [1,2]. In particular, OpenAI released the multimodal model ChatGPT-4.0 in March 2023, which made multimodal learning also become a focus in the research field [3,4]. At the same time, issues arising from the ChatGPT craze have emerged, one of the biggest concerns being the issue of privacy and security challenges of the data used for ChatGPT model training [5]. Multimodal models are trained using datasets collected from a variety of different sources, such as internet big data, social media data, user data, and so on. Most of these data are collected in ways that do not comply with existing data security protection regulations at home and abroad, and the collected data may contain sensitive information about the users, causing the risk of privacy leakage, as well as business secrets and national security issues. Therefore, in order to protect data privacy and

security, multimodal models need to take some measures before they can play a greater role in real applications.

As an emerging machine learning research direction in recent years, the concept of federated learning was first introduced by McMahan et al. in 2016 [6]. Federated learning is a machine learning setup in which multiple clients work together to train a model under the coordination of a central server in a typical federated learning setup, while keeping the training data retained in the clients without leaking it outward. Federated learning has promising applications and can be useful in areas such as recommender systems, smart cities, healthcare and finance [7–10]. Therefore, it is promising to utilize federated learning to address the privacy and security issues of multimodal model training data, and many research scholars are gradually carrying out work on combining multimodal learning with federated learning [11–15].

This paper is a survey of the research on multimodal federated learning. Section 1 introduces the relevant background of multimodal federated learning. Section 2 first presents the research progress of multimodal learning and describes the multimodal learning technology

* Corresponding author.
E-mail addresses: huangweifujian@my.swjtu.edu.cn (W. Huang), wangdexian@my.swjtu.edu.cn (D. Wang), ouyangxiaocao@my.swjtu.edu.cn (X. Ouyang), jhwan@gdut.edu.cn (J. Wan), xiaoke92@foxmail.com (J. Liu), trli@swjtu.edu.cn (T. Li).

from the aspects of multimodal representation learning, multimodal fusion learning, generic multimodal model, and multimodal application, then provides the research progress of federated learning and introduces the federated learning technology from the aspects of client-side model optimization, global model optimization, and multi-task learning. Section 3 focuses on the research progress of multimodal federated learning from the four perspectives of the basic mode of multimodal federated learning, multimodal fusion based on federated learning, multimodal federated learning optimization, and multimodal federated learning application, respectively. Section 4 looks forward to the research trend of multimodal federated learning. Section 5 is the concluding remarks.

2. Related work

2.1. Related work on multimodal learning

Multimodal data appears in various fields in real life, and multimodal data contains different data information, which has important research value. On the one hand, classical deep learning models have been maturing in unimodal applications, showing strong learning and generalization capabilities. On the other hand, there is an urgent need to study the modeling of cross-modal data as well as information fusion in multimodality in many application areas [16–20]. The research progress in this area is described below.

Multimodal representation learning. There are often certain correspondences in multimodal data, and rough correspondences in multilevel representation learning generally refer to shallow data feature correspondences [21]. For example, Wei et al. modeled relationships within and between modalities by extracting modal global features, local similarities, and proposed a cross-modal network for a graphic matching task [22]. However, deeper correspondences in multilevel representations contain correspondences of entities, concepts, semantics, etc., which require the construction of graphs to learn these correspondences [23–28]. For example, Gao et al. designed a multimodal graph neural network to describe three modalities, i.e., image, language, and number, respectively, and then perform entity matching [29]. Kim et al. utilized the symbol graph as a multimodal public semantic space to obtain the hierarchy of information of different modalities, so as to jointly represent different modalities [30]. Zeng et al. constructed the objects of video clips and query statements, respectively, with a relationship graph and then perform object matching to finalize the cross-modal retrieval task [31].

Multimodal and multiscale representations give the model the ability to extract multiscale features of different modalities [32–39]. Cheng et al. used a high-resolution feature pyramid to learn multiscale perceptual representations as a way to solve the problem of scale variation in a bottom-up multi-person pose estimation task [40]. Li et al. proposed a dynamic selection mechanism for the convolutional kernel in a convolutional neural network, where each neuron can adaptively adjust its receptive field size based on multiple scales of input information [41]. Yang et al. presented to consider the multi-scale structure itself as a feed-forward model, which is composed as a directed acyclic graph, whereby coarse and fine grained feature sharing can be easily realized [42]. Cai et al. put forward a unified multiscale deep convolutional neural network model, which designs different output layers to obtain different scales of perceptual fields, and uses feature upsampling to replace input upsampling to reduce memory and computational overhead [43].

Multimodal fusion learning. Different modal data representations and the information they contain are different, there are some information crossings and complementarities, and reasonable fusion of multimodal information can get richer information. Zheng et al. proposed a pretrained model for fusing speech and text that learns a unified multimodal representation [44]. Kumar et al. designed a modality-aware fusion model that uses multimodal contextual attention and global

information fusion modules to capture multimodal information [45]. Nagrani et al. introduced a new Transformer-based architecture that uses a "fusion bottleneck" for multilayer modal fusion, which requires the model to organize and compress the relevant information in each modality and share the necessary information [46]. Wang et al. presented a multimodal fusion method by exchanging channels between different modalities [47]. Joze et al. put forward a cross-modal fusion approach for convolutional neural network structures that fuses information from the middle layer of the network without drastically altering the underlying network structure [48]. Inspired by neuroscience ideas about multisensory integration and processing, Shanka et al. investigated the introduction of neural dependencies in the loss function for multimodal fusion and validated it on a multimodal sentiment analysis task [49]. Wang et al. performed multimodal sarcasm detection for tweets consisting of text and images and proposed a multimodal hierarchical fusion model to solve this task [50]. Liu et al. combined the characteristics of different modalities of early fusion and late fusion and offered a multi-stage bi-directional fusion scheme that can be used for a variety of architectures and tasks oriented towards fusion of images and point clouds [51].

Generalized multimodal model. Generally, multimodal learning methods design a corresponding model structure for each modality, and such targeted modality-tailored models can better extract the features of the modality, thus improving the learning performance. Another part of the work is devoted to exploring the study of generalized multimodal model, that is to develop a generalized model that learns any modality data. It can automatically handle more and larger multimodal data (dozens or even hundreds of modalities), thus eliminating the tedious work of manually specifying a model for each modality and realizing the intelligent processing of multimodal data. For example, Liu et al. designed a cross-modal collaborative representation learning framework with modality-specific and generic modality-specific feature extractors [52]. Similarly, Akbari et al. devised modality-specific and generic modality feature fusioners, which are transferable to multiple frameworks and tasks [53]. In addition, the Gato developed by Google DeepMind is a multimodal, multi-task and multi-implementation model with only one set of optimized parameters trained by a single Transformer under multiple tasks, which performs well in multimodal applications [54]. Gato is constructed in a similar way to large-scale language models such as GPT-3.

Multimodal learning applications. Liu et al. proposed a multimodal transportation recommender system that enables multimodal representations of Point of Interest (POI), geographic distribution and weather etc [55,56]. Liu et al. further exploited the spatio-temporal correlation of transportation networks and the semantic consistency of historical routes to propose a multimodal transportation recommender system with a unified route representation [57]. Sun et al. introduced a multimodal dialog response generation task, where the model needs to generate text or images of a given contextual dialog as a response, and designed a multimodal dialog intelligences so that they can learn the dialog capabilities from a large number of text dialogs and text-image sets, respectively [58]. Zhu et al. suggested a multimodal trajectory prediction method to solve the problem of predicting the behavior of multiple intelligences in an autonomous driving task [59]. Ke et al. presented a multi-task and multi-graph learning approach for multimodal ride hailing demand and other joint spatio-temporal prediction tasks [60].

The above multimodal learning is mainly studied for centralized data. It is impossible to share and model these data directly with other organizations, considering that the data from these different sources will involve personal private information and company trade secrets, etc., so combining multimodal learning and federated learning is a research hotspot in recent years. An introduction to the state of the art in federated learning is given next.

2.2. Related work on federated learning

Federated learning is a machine learning framework that specializes in solving data silos and data privacy issues to achieve common modeling and improve the effectiveness of AI models while ensuring data privacy security and legal compliance [61]. With federated learning technology, data from all parties can be used by the central model (server), but the data itself will not be disclosed. Using the characteristics of federated learning, multi-party organizations work together to build machine learning models, thus fully protecting user privacy and data security, and achieving mutual benefits for all parties. The following describes the research progress in this field.

Optimizing client models in federated learning. Zhao et al. improved training on heterogeneous data by introducing Earth Move Distance and sharing a small portion of global data between clients [62]. Jeong et al. borrowed the idea of knowledge distillation [63] to propose the federated distillation algorithm, which incorporates the difference between local and global vectors as a regular term in local training [64]. They also proposed the FAug algorithm, which gradually transforms the original heterogeneous dataset into an IID dataset by continuously generating data through server-side training of the GAN model [64]. Yao et al. presented the FedMMD [65] and FedFusion [66] algorithms. FedMMD adds Maximum Mean Discrepancy loss to the objective function of the local model, which allows the knowledge of the global model to be better fused into the local model. FedFusion introduces a feature fusion mechanism, which effectively fuses the global features with the local features in the client's local model. Huang et al. draw on the idea of the AdaBoost algorithm [67] to dynamically adjust the number of epochs for client model training in the current iteration rounds based on the median of the loss value of each client in the previous round, to solve the problem of inconsistent distribution of data among each client [68].

Global model optimization in federated learning. More work in federated learning has focused on optimization methods and objective functions for global models, in addition to improvements on local models at the client side. Mohri et al. modified the weighted average method for global model aggregation by selecting a vector of coefficients from a subset of weight coefficients such that the loss is maximized [69]. Li et al. constructed a broader framework, Fed-Prox, based on FedAvg, which is capable of handling heterogeneous federated data while maintaining similar privacy and computational advantages [70]. The team also proposed the q-FFL method, which constructs a novel objective that can improve the fairness of accuracy distribution in federated learning [71].

Federated multi-task learning. The goal of federated multi-task learning is to provide each client with a model that best fits the local data distribution. Sattler et al. proposed the Clustered Federated Learning algorithm, which allows clients with similar data to protect each other while minimizing interference between clients with different data [72]. Shoham et al. introduced the FedCur framework, which adds a penalty term to the loss function that forces all local models to converge to a shared optimum [73]. The FedCur and FedProx algorithms use the same parameter stiffness, but FedCur incorporates the idea of multi-task learning to solve the federated learning problem of how to learn a task without interfering with the different tasks learned on the same model to improve convergence. Smith et al. presented the MOCHA framework that connects each task with different parameters and learns the independence of each device by adding a loss term, while modeling task correlations through multi-task learning using a shared representation [74].

Federated multi-task learning concerns the differences and commonalities of the training data for each client in the real world. It combines the idea of multi-task learning and considers each client as an independent task to solve the problem of non-independent and homogeneous distribution of client data, so as to train a model with personalized and adapted client problem solution for each client in the real world. The above research on federated learning mainly focuses on the exploration of the federated optimization problem. Federated learning can safely and effectively integrate multi-party data, so the combination of federated learning technology and multimodal learning technology is able to realize the real-world use of multimodal learning technology.

2.3. Related work on multimodal federated learning

Single modality data is far from enough to support various applications in today's life. With the rapid development of various sensor technologies and multimedia technologies, it is easy to obtain a variety of different modal data, such as images, audio, video, point clouds, natural language, geographic data, etc. Using multiple modal information for modeling, different modal data can provide complementary information and improve the model representation capability. Combined with federated learning, it allows multimodal learning to learn information from multiple modalities without violating data security protection regulations, making multimodal federated learning valuable for important real-world applications.

Che et al. categorized federated learning into consistent multi-modal federated learning and inconsistent multimodal federated learning based on whether the clients have the same combination of modalities, which corresponds to the basic framework of federated learning [11]. Lin et al. reviewed multimodal federated learning, which is mainly classified and discussed based on the modal distribution and modal annotations of multimodal federated learning [12]. Qi et al. proposed a multimodal data fusion method based on the framework of federated learning, which fuses time-series data from wearable sensors and visual data from cameras for solving fall detection tasks [13]. Guo et al. combined federated learning with a visual language model to take advantage of the unique benefits of multimodality in the visual language model to improve client generalization and robustness [14]. Cremonesi et al. presented the main challenges and lessons learned in designing multimodal healthcare data prediction models under a federated learning framework, which informs federated learning research efforts in healthcare [15].

The above outlines the research progress of multimodal learning and federated learning respectively, which provides a reference for the research of multimodal federated learning. Multimodal federated learning can safely and efficiently utilize information from different modal data to build a more superior model, so more and more researchers are focusing on multimodal federated learning. The content of this paper is different from the existing multimodal federated learning reviews, which classifies multimodal federated learning into three basic modes, and mainly explores and analyzes the existing multimodal federated learning methods from the research perspective of the fusion mode for modal information in the federated learning framework.

3. Multimodal federated learning

Multimodal federated learning constructs models that can process and correlate multimodal information under the premise of guaranteeing information security and data privacy. This section introduces multimodal federated learning from four perspectives: the basic mode of multimodal federated learning, multimodal fusion based on federated learning, multimodal federated learning optimization and multimodal federated learning application.

3.1. The basic mode of multimodal federated learning

Multimodal federated learning mainly focuses on learning feature representations of different modalities and uniting different clients on the server for mutual complementation and improvement of different modal features, which can be seen as a kind of server-based multimodal
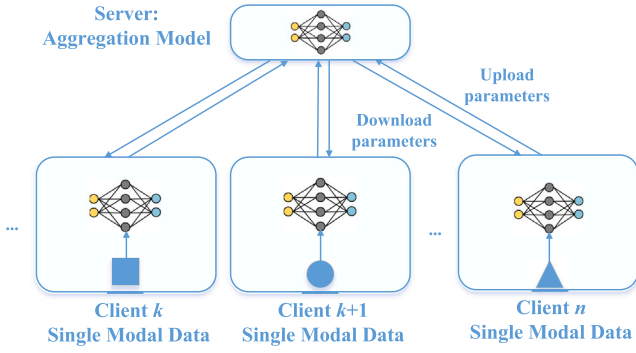
Fig. 1. Mode I-CSSM: The data on the client side is single-single modality.

Fig. 2. Mode II-CSMM: The data on the client side is single-multiple modality.

feature fusion. There are mainly three different modes in the multi-modal federated learning model, and the frameworks and related work of the three modes are specified below.

Mode I-CSSM is a single-single modality for client data. As shown in Fig. 1, client k represents a particular client in Mode I-CSSM. There are n clients in total. The different shapes of the clients in the figure represent one modality of data. For example, the square in client k represents text data. Each client in this mode only has data of one modality, and the clients use the data of this modality to train the model and perform model aggregation through the server to realize the common information interaction between different modalities.

Yang et al. followed Mode I-CSSM to propose a novel cross-modal federated human activity recognition task, thus facilitating the large-scale application of human activity recognition models to more local devices [75]. The task considers data with only one modality per device. The task considers two challenging issues. First, how to collaboratively construct a common feature subspace for different clients with cross-modal heterogeneity. Second, all the knowledge learned from one client may not be useful for models of other clients with different modalities. Therefore, this work proposed a feature-decomposed activity recognition network that distributedly aggregates local models learned on clients with different modalities.

There is also work on proposing a multimodal federated learning model based on Mode I-CSSM combined with a vertical federated learning framework. Wei argued that feature fusion methods change the structural information of the original data while losing some feature information that may have important judgmental value, so he proposed a multimodal heterogeneous data mining model based on federated learning [76]. The model first trains the classification model of each modal data independently, then applies an aggregation algorithm to analyze the direction of their gradient descent, followed by feeding the results back to each client, and finally each client uses the gradient to update their respective models. The model not only cleverly solves the heterogeneity problem, but also mines the information between different modal data, which promotes the robustness and discriminative nature of the model.

Mode II-CSMM is single-multiple modality for client data. As shown in Fig. 2, client k represents one of the clients of Mode II-CSMM, which stores data in three modalities (using circles, triangles, and squares to represent speech, video, and text data, respectively). While client k + 1 has only speech data. In this model, some clients have data in multiple modalities, and some clients have data in only one modality. Clients learn the feature representation of a single modality or the feature representation of multiple modalities, and realize the information interaction between the same modality or different modalities under different clients through the server aggregation model.

For the second mode, Zhao et al. proposed a multimodal semi-supervised federated learning framework can handle data from different modalities and single modality [77]. The client of the framework
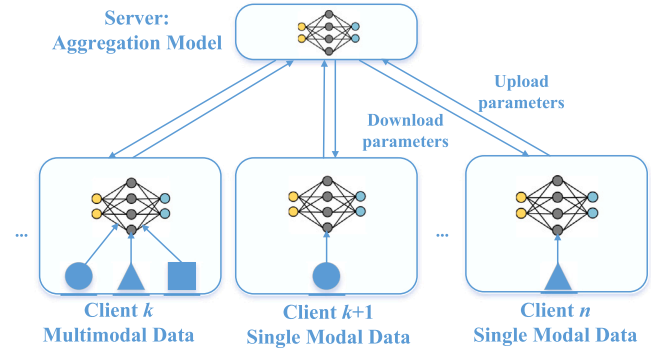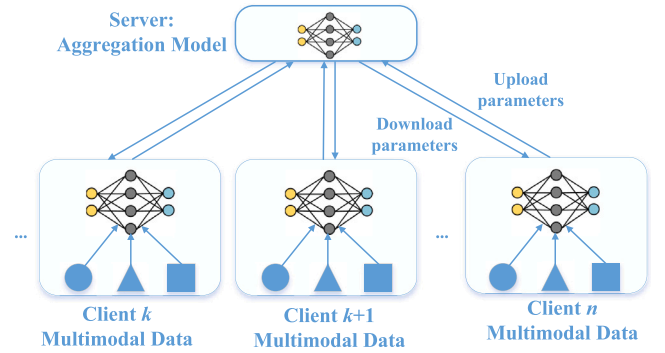

Fig. 3. Mode III-C3M: The data on the client side is multiple-multiple modality.

extracts shared or associated hidden representations from multiple modalities, and the server uses the multimodal FedAvg algorithm to aggregate local autoencoders trained in different data modalities. Experiments on a human action recognition dataset show that the framework can jointly model multiple clients with different modal data and improve model accuracy.

Mode III-C3M is multiple-multiple modality for client data. As shown in Fig. 3, client k represents a certain client in Mode III-C3M, which stores data in three modalities (represented using circles, triangles, and squares, respectively). Each client in this mode has data in multiple modalities, and the clients use these multimodal data to train models and perform model aggregation through the server to realize the information interaction of multimodal data under different clients.

Zong et al. presented a framework for federated cross-modal retrieval in a distributed data storage scenario [78]. The basic framework of the model follows the third mode, where each client has data with multiple modalities. The specific training process of the model is described as follows, first the cross-modal retrieval model is trained and its local data is used to learn the common space of multiple modalities in each client; then the central server aggregates the common subspaces of multiple clients; and finally each client updates the common subspace of the local model based on the common subspace aggregated on the server.

The above describes the three basic modes of multimodal federated learning, and gives an architectural diagram and relevant examples for each mode. Multimodal federated learning in which each client is allowed to have either single or multiple modalities of data, federated learning in this form is more in line with practical applications and promotes the use of models in many different scenarios. Multimodal fusion can integrate information from multiple modalities to obtain a consistent common model output. It is a fundamental problem in the multimodal domain, and thus is analyzed next for multimodal fusion in federated learning.
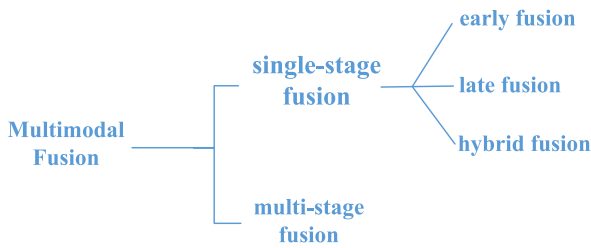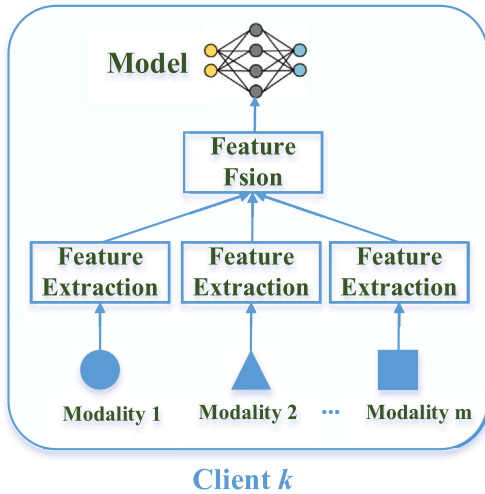
Fig. 4. The way of multimodal fusion.

Fig. 5. Early fusion model.

3.2. Multimodal fusion based on federated learning

Multimodal fusion is an essential step in the modeling process of multimodal federated learning models in order to allow the model to learn the data information of different modalities as well as to capture the hidden values between modalities. Multimodal fusion is mainly categorized into single-stage fusion and multi-stage fusion, as shown in Fig. 4. Single-stage multimodal fusion can be categorized into three forms, i.e., early fusion, late fusion, and hybrid fusion. The federated learning based multimodal fusion approach mainly focuses on model construction with multimodal fusion on the client and then further aggregated models on the server.

As shown in Fig. 5, the model of the client can be modeled by early fusion, where m represents the data of the *m*th modality. Early fusion generally directly fuses the raw data of different modalities, or extracts the modal feature representations of different input data, and then fuses the extracted features to achieve the integration of information from different modalities.

There have been some research works on applying multimodal federated learning models with early fusion approaches on the client side. Mo et al. presented a multi-source heterogeneous data fusion method based on Tucker decomposition in federated learning by constructing a higher-order tensor with heterogeneous spatial dimensionality properties to capture the high-dimensional features of heterogeneous data [79]. The method is Mode II-CSMM in the basic model of multimodal federated learning, i.e., the client data is in the form of single-multiple modality. As an example of combining the basic Mode II-CSMM of multimodal federated learning with the early fusion approach applied on the client, a framework diagram for constructing multimodal federated learning is given in Fig. 6.

The framework performs feature extraction (including audio feature extraction, visual feature extraction, and text feature extraction)

on local multimodal data in client model training. Different feature extraction modules correspond to different modal feature extraction sub-networks, where the audio modal information and video modal information are feature extracted using the acoustic analysis framework and the facial expression analysis framework, respectively, and the text modal information is feature extracted using the long and short-term memory network and the convolutional neural network. Then the modal features from one to three are fused using the feature fusion module, which introduces the tensor decomposition theory, constructs the higher-order memory unit of the spatial dimensional characteristics of heterogeneous data, and uses the memory unit to fuse the data information of different modalities. For the fused data, the framework uses the traditional fully connected layer to make decisions on the basis of global features. After the model training is completed, it is uploaded to the server for model aggregation, and the average aggregation algorithm is applied to the feature fusion module, the feature decision module, and the feature extraction sub-module to complete the model aggregation, finally the updated model is re-distributed to the client for the next round of training.

There are also research efforts to apply multimodal federated learning models with early fusion approaches on the client side. For example, for sensory data that exists in multiple modalities (text, video, audio, etc.), Wang et al. used a multimodal Transformer approach to merge multimodal data prior to subsequent operations, fusing multichannel information by directly attending to low-level features of other channels [80]. Xiong et al. employed a common attention mechanism in client model training to fuse complementary information from different modalities, and learn useful global features from different modalities in order to jointly train a common model for all clients and adapt each client's model using a personalized approach based on model agnostic meta-learning [81]. Psaltis et al. performed multimodal fusion at different granularity levels and proposed a federation aggregation mechanism [82]. Nandi et al. presented a federated learning based algorithm for real-time sentiment classification of multimodal data streams, where the popular feature concatenation is used for feature fusion in client-side model training, which is also an early fusion approach [83,84]. Salehi et al. put forward a multimodal federated learning framework where data from multiple modalities (e.g., LiDAR, GPS, and camera imagery) are fused using deep learning at the client side, and the server combines model weights from multiple clients and then propagates the final fused architecture back to the client side [85]. For the fusion of multimodal image representations of different visions and languages, Liu et al. introduced a federated learning framework to fuse various types of image representations in different tasks and conducted experiments under three modes of federated learning (horizontal federated learning, vertical federated learning, and federated transfer learning) to demonstrate the effectiveness of the method [86].

As shown in Fig. 7, if the model of the client takes a late fusion approach to modeling, it first trains a specific model for different modal data, then fuses the predictions of each modal model, and finally makes a prediction of the result.

As shown in Fig. 8, the client can also be modeled using hybrid fusion. The hybrid fusion approach combines early and late fusion, which has the advantages of both approaches, while also increasing the complexity of the model structure and the difficulty of training.

The above three fusion methods are single-stage methods. Fig. 9 shows a multi-stage bidirectional fusion approach, where the feature extraction method for each branch is selected to be suitable according to the modal data to be extracted in that branch. For example, a convolutional neural network is used to extract video and image features and a language model is used to extract text features. Feature fusion integrates the feature vectors of different modalities through fusion methods, such as splicing operations, weighted summation methods, and attention-based fusion methods, etc. Each branch needs to judge the correlation between modalities when performing feature fusion and select a suitable method for feature fusion in the current branch.
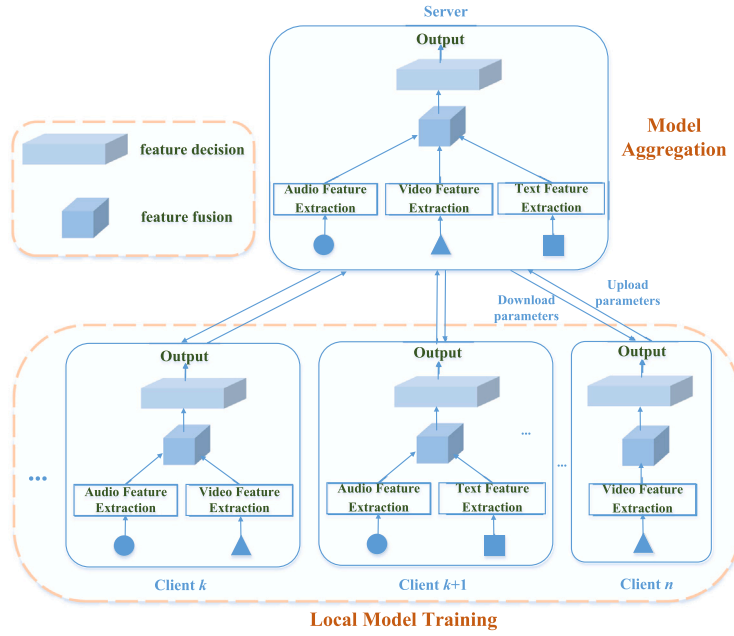
Fig. 6. An early fusion framework based on the multimodal federal learning Mode II-CSMM.
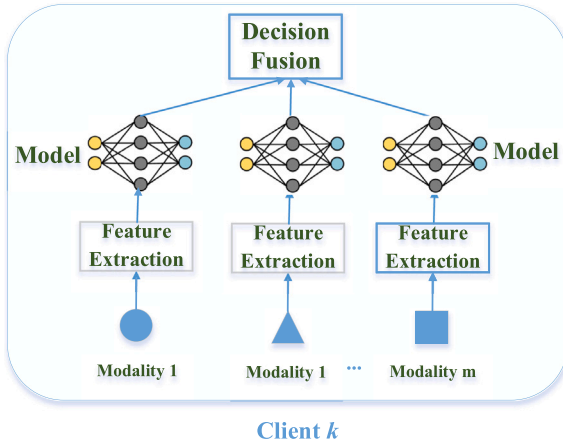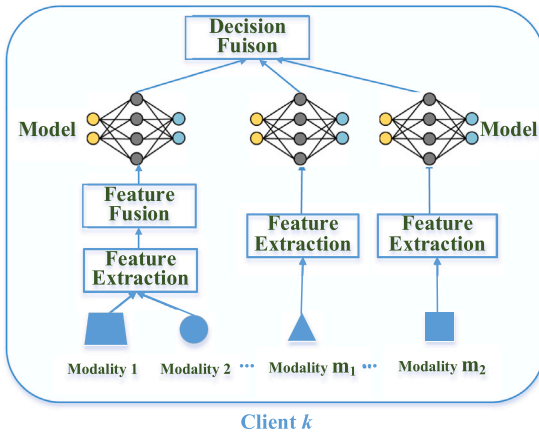

Fig. 7. Late fusion model.

Fig. 8. Hybrid fusion model.

Multi-stage bidirectional fusion methods have multiple stages of feature fusion steps, and integrate the features of multiple modalities through each stage of the feature fusion steps to realize the mutual complementation and transmission of different modal data information. Liu et al. designed this bidirectional fusion framework in order to fully utilize the characteristics of each modality as well as the intermodal complementarities, associating the two branches bidirectionally to achieve complementary information transfer [51].

There has been some research work on multimodal fusion methods based on federated learning, but it is still in its infancy. Most of the approaches are modeled using multimodal early fusion approach at the client side. Research work on using late fusion approach, hybrid fusion approach and multi-stage fusion approach at the client side is still to be investigated. The implementation of these fusion approaches is more difficult compared to the early fusion approaches, and considering the limited resources (e.g., CPU, memory, energy consumption, etc.) of the client in federated learning, the work to use these fusion approaches on the client side has not been realized yet. However, deploying federated learning frameworks in different real-world scenarios can be targeted to use appropriate fusion approaches for applications, and we believe that more work will focus on different federated learning-based multimodal fusion approaches in the future.

3.3. Multimodal federated learning optimization

Federated learning can safely and effectively unite multi-party data, but it still faces some problems (e.g., heterogeneity, communication burden, etc.), while multimodal federated learning also suffers from these problems, and federated learning in order to effectively merge multimodal information will make these problems more prominent. For example, different clients in multimodal federated learning distribute data in different modalities, which makes the statistical heterogeneity among clients more serious. In order to make the multimodal federated learning model have higher accuracy and faster convergence speed, as well as to ensure the security, robustness and efficiency of the model, it is necessary to design a reasonable optimization method, so some researchers have focused on the optimization of multimodal federated learning.

The current federated learning optimization methods mainly study the optimization of local model versus global model [62,63,69,70],
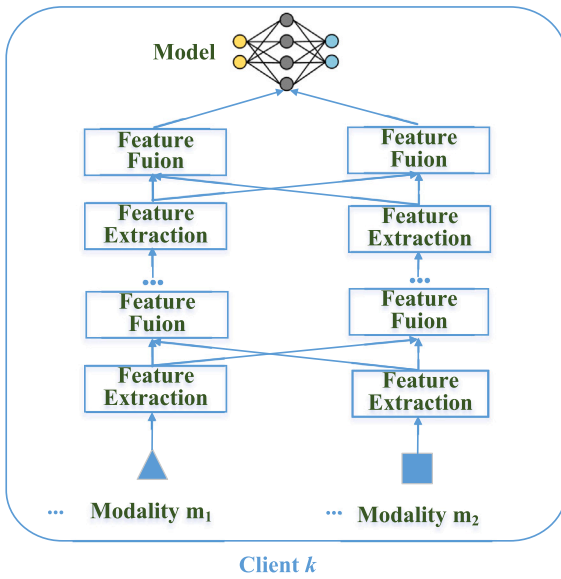
Fig. 9. Multi-stage fusion model.

the efficiency of the model [87–90] and the optimization of model communication [91–93]. Among them, the optimization of local and global model mainly considers the accuracy problem and convergence problem of the model, as well as improving the performance of the model. This can be achieved by optimizing the objective function, designing personalized models, and multi-task learning. Model efficiency aims to effectively improve the training efficiency of federated learning under the premise of ensuring that the model performance does not degrade, and to reduce the data transmission between each client or between the client and the server. This can be achieved by methods such as model compression, model distillation, and model pruning. Communication optimization of the model takes into account the communication between the server and the client, and ensures limited communication during model training by designing efficient communication protocols, asynchronous update schemes, and reducing the number of communication rounds.

For the optimization of local and global model in multimodal federated learning, Chen et al. attributed the significant performance degradation of multimodal federated learning models to the fact that both locally updated and globally aggregated model suffered from overfitting and inconsistent generalization rates [94]. In order to balance the generalization rates of local update and global aggregation, they proposed a new hierarchical gradient hybrid training algorithm, which calculates the optimal mix of modalities and the optimal weights of the local model at the same time by adapting an adaptive measure of modal overfitting and generalization phenomena, and conducted rigorous theoretical analysis and practical tests.

Optimizing for model efficiency in multimodal federated learning, Ji et al. considered different inference speeds due to different computational capabilities of training devices and modal complexity, and proposed a vertical federated learning model to detect human states using multimodal data. The model utilizes lightweight feature extraction submodels while ensuring data privacy and tolerating accuracy degradation [95]. In addition, a fast and secure module is designed that effectively reduces the amount of data during network transmission, decreasing the model's dependence on the network and the risk of data interception by third parties.

For the optimization of model communication in multimodal federated learning, Wei et al. utilized artificial intelligence techniques to address key challenges in cross-modal communication, adopting a federated learning paradigm first to address sparse data collection

and privacy protection in the description of immersive experiences for multimodal services. The reinforcement learning paradigm is then used to construct a joint optimization framework for co-transmission of audio, visual and haptic streams. Finally, a transfer learning paradigm is used to extract, migrate, and fuse features from different modalities to deal with problems such as corruption, loss, or asynchronous arrival of signals at the receiver side in cross-modal communication, thus enabling effective cross-modal communication that supports artificial intelligence [96].

There has been much work on federated learning optimization. Due to the multimodal complexity of data in multimodal federated learning, it brings more serious data heterogeneity and communication burden problems to the federation learning training process, and the direct application of federation learning optimization methods to multimodal federated learning is not necessarily effective, so the research on multimodal federation learning optimization needs further research and development. The research and development of multimodal federated learning optimization plays an important role in applying multimodal federated learning to real-life applications, which can solve the problem of data confidentiality in practical applications and effectively integrate the data of multiple participants in multiple modalities. The practical applications of multimodal federated learning are described next.

3.4. Multimodal federated learning application

The application of multimodal federated learning involves a variety of different domains, such as the medical field, automatic driving, visual question and answer (Q&A), action recognition, and emotion classification. Existing applications of multimodal federated learning are summarized in Table 1, along with the corresponding reference, basic mode, and model for each application.

Federated learning techniques have been widely used in medical big data mining and processing [105–107]. Due to the security and privacy protection requirements and ownership of medical big data, it is difficult to collaborate and integrate big data distributed in different medical institutions, and it is hard for machine learning methods with centralized data training mode to mine knowledge with clinical value, which limit the development space of smart healthcare. With the help of the federated learning infrastructure framework, it can break the data barriers between medical institutions and promote the popularization of artificial intelligence technology in the field of smart medical industry [108]. While medical big data is generally characterized by large scale, diverse types, and rich potential information, for this reason some researchers have applied multimodal federated learning to the medical field. To enable telemedicine centers lacking advanced diagnostic facilities to use multimodal data safely and efficiently, Qayyum et al. utilized a framework of clustered federated learning to automatically diagnose COVID-19, thus alleviating the pressure on the global healthcare system since the occurrence of novel coronaviruses [97]. Parekh et al. explored multimodal, multitask federated learning in medical imaging by experimentally demonstrating the feasibility of a cross-domain federated learning model as well as the possibility of applying federated learning in medical imaging [98]. In response to melanoma disease analysis, Agbley et al. fused two modalities of data, i.e., skin lesion images and their corresponding clinical data, and compared the performance of the global federated model with the results of a centralized learning scenario [99]. Wang et al. suggested a federated multimodal unsupervised brain image synthesis model, FedMed-GAN, which facilitates the development of medically generated images and ensures privacy security. It can be used for human cognitive activities and certain pathology studies [100]. To address the data heterogeneity of federated learning applied to liver image segmentation, Bernecker et al. proposed the FedNorm algorithm based on modal normalization technique and its federated learning algorithm extending FedNorm+ [101]. Specifically, FedNorm normalizes features at the

Table 1
The application of multimodal federated learning.

Application	Reference	Basic mode	Model
Medical field	[97,98]	Mode I-CSSM	Multimodal fusion based on federated learning server
	[99–101]	Mode III-C3M	Multimodal fusion based on federated learning
Automatic driving	[102]	Mode II-CSMM	Multimodal fusion based on federated learning
Visual Q&A	[86]	Mode III-C3M	Multimodal fusion based on federated learning
	[103]	Mode II-CSMM	The optimization of multimodal federated learning
Action recognition	[81]	Mode III-C3M	Multimodal fusion based on federated learning
	[94]		The optimization of multimodal federated learning
Human activity recognition	[82]	Mode III-C3M	Multimodal fusion based on federated learning
	[75]	Mode I-CSSM	Multimodal fusion based on federated learning server
Human state detection	[95]	Mode I-CSSM	The optimization of multimodal federated learning
Cross-modal retrieval	[78]	Mode III-C3M	Multimodal fusion based on federated learning server
	[104]	Mode II-CSMM	The optimization of multimodal federated learning
Emotion classification	[79]	Mode II-CSMM	Multimodal fusion based on federated learning
	[83,84]	Mode III-C3M	
Blockchain and edge computing	[80]	Mode III-C3M	Multimodal fusion based on federated learning
Vehicle sector selection	[85]	Mode III-C3M	Multimodal fusion based on federated learning

client level, while FedNorm+ uses modal information of individual slices in feature normalization.

Multimodal federated learning is also applied to several fields, such as autonomous driving, visual Q&A tasks, and action recognition. For privacy-preserving computation and collaboration in autonomous driving, Tian et al. presented a hierarchical structure of intelligent vehicle Transformers, which can effectively represent and fuse the different modal input data of intelligent vehicles and realize the secure sharing of data resources among different vehicles [102]. For retrieval and visual Q&A tasks, since existing federated learning approaches that extend to multimodal data rely on model aggregation at the single-modal level, Yu et al. introduced a multimodal federated learning framework Contractive Representation Ensemble and Aggregation for Multimodal FL (CreamFL), which is the first multimodal federated learning framework based on knowledge distillation that supports heterogeneous modalities and model architectures between servers and clients [103].

4. Discussion

From the above analysis for the basic mode of multimodal federated learning, multimodal fusion based on federated learning, the optimization of multimodal federated learning and the application of multimodal federated learning, it can be seen that the work of combining multimodal and federated learning is still in its infancy, and the federated learning can solve the shortcomings and defects of the multimodal problems to a certain degree and improve the performance of the model, so there is still a lot of room for exploration in this field. The following summarizes four directions worth studying.

The optimization of multimodal federated learning. Many researchers have devoted themselves to make the federated learning model have higher accuracy and faster convergence speed, as well as to ensure the security, robustness and efficiency of the model, and have proposed many optimization methods on federated learning [109,110]. However, in order to make the multimodal federated learning model can be really used in practical applications, it is more necessary to design optimization methods for multimodal federated learning. For example, optimization methods for studying heterogeneity and communication burden problems in multimodal federated learning are a must for the research process of multimodal federated learning. Therefore, optimization for multimodal federated learning is also a very valuable research direction.

Privacy-preserving techniques for multimodal federated learning. The current attacks in federated learning are increasingly diversified and the technology continues to evolve, while most of the existing defense strategies in federated learning are designed for specific attack methods, making most federated learning models unable to cope with the serious challenges of data security and privacy protection [111,112]. Moreover, single-modal federated learning defense strategies are not necessarily applicable to multimodal federated learning. How to extend the single-modal federated learning defense strategy to multimodal federated learning is also a worthy research question.

Multimodal federated few-shot learning & multimodal federated semi-supervised learning. Current multimodal federated learning methods generally have a strong data dependency problem, i.e., they must rely on a large amount of labeled data for deep learning training. However, it is more difficult to obtain a large amount of data and its labeling in real and complex environments, which is prone to problems such as insufficient data volume and scarce labeled data [104, 113]. It is challenging to refine the feature representation of a small number of samples, merge the information knowledge among multiple modalities, improve the effectiveness of task-specific processing in the case of insufficient samples, and extend the generalization requirements for different types of task processing under the federated learning framework [114,115]. Therefore, research on multimodal federated few-shot learning and multimodal federated semi-supervised learning, which solve the label sparsity problem and satisfy the requirements of small-sample, multimodal, high-performance, and cross-task, is also a direction worth exploring in this field.

Data and knowledge-driven multimodal federated learning. The data between different clients under the federated learning framework suffers from the problems of unbalanced distribution of multimodal data and non-uniformity of data information between modalities. In addition, the information of different modalities between clients is not necessarily useful for another client. How to unify multimodal information between clients and effectively utilize multimodal information from other clients are also key issues. Data and knowledge federated-driven approaches can integrate knowledge with data of different modalities to obtain more comprehensive, accurate and meaningful data features. Therefore, under the framework of federated

learning, the study of multimodal models and methods that effectively incorporate knowledge can realize the deep integration of data and knowledge and improve the robustness of models to data noise.

5. Conclusions

Multimodal federated learning not only promotes the development of multimodal learning in related neighborhoods, such as finance and healthcare, but also enhances the ability to mine knowledge from multimodal data and the application possibilities of multimodal models. This paper systematically reviewed the existing multimodal federated learning methods and introduced them from four main aspects, including the basic mode of multimodal federated learning, multimodal fusion based on federated learning, the optimization of multimodal federated learning and the application of multimodal federated learning. And it also gives an outlook on the future direction of multimodal federated learning, including the optimization of multimodal federated learning, privacy-preserving techniques for multimodal federated learning, multimodal federated few-shot learning and multimodal federated semi-supervised learning, and data and knowledge-driven multimodal federated learning. We hope that this paper can contribute to the continued development and progress of this research area.

CRediT authorship contribution statement

Wei Huang: Writing – review & editing. **Dexian Wang:** Resources. **Xiaocao Ouyang:** Conceptualization. **Jihong Wan:** Data curation. **Jia Liu:** Writing – review & editing. **Tianrui Li:** Supervision.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript.

Data availability

No data was used for the research described in the article.

Acknowledgments

References

[1] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, et al., Towards artificial general intelligence via a multimodal foundation model, Nature Commun. 13 (1) (2022) 3094.

[2] Y. Ling, F. Wu, S. Dong, Y. Feng, G. Karypis, C.K. Reddy, International workshop on multimodal learning-2023 theme: Multimodal learning with foundation models, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 5868–5869.

[3] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al., A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, in: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 675–718.

[4] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, Y. Tang, A brief overview of ChatGPT: The history, status quo and potential future development, IEEE/CAA J. Autom. Sin. 10 (5) (2023) 1122–1136.

[5] D. Guo, H. Chen, R. Wu, Y. Wang, AIGC challenges and opportunities related to public safety: a case study of ChatGPT, J. Saf. Sci. Resil. 4 (4) (2023) 329–339.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.

[7] O.A. Wahab, G. Rjoub, J. Bentahar, R. Cohen, Federated against the cold: A trust-based federated learning approach to counter the cold start problem in recommendation systems, Inform. Sci. 601 (2022) 189–206.

[8] S. Banabilah, M. Aloqaily, E. Alsayed, N. Malik, Y. Jararweh, Federated learning review: Fundamentals, enabling technologies, and future applications, Inf. Process. Manage. 59 (6) (2022) 103061.

[9] Z. Zheng, Y. Zhou, Y. Sun, Z. Wang, B. Liu, K. Li, Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges, Connect. Sci. 34 (1) (2022) 1–28.

[10] J. Guo, Z. Liu, S. Tian, F. Huang, J. Li, X. Li, K.K. Igorevich, J. Ma, TFL-DT: A trust evaluation scheme for federated learning in digital twin for mobile networks, IEEE J. Sel. Areas Commun. 41 (11) (2023) 3548–3560.

[11] L. Che, J. Wang, Y. Zhou, F. Ma, Multimodal federated learning: A survey, Sensors 23 (15) (2023) 6986.

[12] Y.-M. Lin, Y. Gao, M.-G. Gong, S.-J. Zhang, Y.-Q. Zhang, Z.-Y. Li, Federated learning on multimodal data: A comprehensive survey, Mach. Intell. Res. 20 (4) (2023) 539–553.

[13] P. Qi, D. Chiaro, F. Piccialli, FL-FD: Federated learning-based fall detection with multimodal data fusion, Inf. Fusion 99 (2023) 101890.

[14] T. Guo, S. Guo, J. Wang, Pfedprompt: Learning personalized prompt for vision-language models in federated learning, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1364–1374.

[15] F. Cremonesi, V. Planat, V. Kalokyri, H. Kondylakis, T. Sanavia, V.M.M. Resinas, B. Singh, S. Uribe, The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform, J. Biomed. Inform. 141 (2023) 104338.

[16] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

[17] S. Chen, P.-L. Guhur, C. Schmid, I. Laptev, History aware multimodal transformer for vision-and-language navigation, Adv. Neural Inf. Process. Syst. 34 (2021) 5834–5847.

[18] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, Y. Song, Parameter efficient multimodal transformers for video representation learning, in: Proceedings of the International Conference on Learning Representations, 2021, pp. 1–17.

[19] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, J. Berant, MultiModalQA: complex question answering over text, tables and images, in: Proceedings of the International Conference on Learning Representations, 2020, pp. 1–12.

[20] Z. Ma, J. Li, G. Li, Y. Cheng, UniTranSeR: A unified transformer semantic representation framework for multimodal task-oriented dialog system, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 103–114.

[21] Y.-W. Chen, Y.-H. Tsai, M.-H. Yang, End-to-end multi-modal video temporal grounding, Adv. Neural Inf. Process. Syst. 34 (2021) 28442–28453.

[22] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10941–10950.

[23] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, Multi-modal sarcasm detection via cross-modal graph convolutional network, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, Association for Computational Linguistics, 2022, pp. 1767–1777.

[24] Y. Liang, G. Huang, Z. Zhao, Joint demand prediction for multimodal systems: A multi-task multi-relational spatiotemporal graph neural network approach, Transp. Res. C 140 (2022) 103731.

[25] Y. Liang, G. Huang, Z. Zhao, Bike sharing demand prediction based on knowledge sharing across modes: A graph-based deep learning approach, in: Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems, IEEE, 2022, pp. 857–862.

[26] L. Zhang, X. Geng, Z. Qin, H. Wang, X. Wang, Y. Zhang, J. Liang, G. Wu, X. Song, Y. Wang, Multi-modal graph interaction for multi-graph convolution network in urban spatiotemporal forecasting, Sustainability 14 (19) (2022) 12397.

[27] R. Saqur, K. Narasimhan, Multimodal graph networks for compositional generalization in visual question answering, Adv. Neural Inf. Process. Syst. 33 (2020) 3070–3081.

[28] W. Zhao, X. Wu, J. Luo, Multi-modal dependency tree for video captioning, Adv. Neural Inf. Process. Syst. 34 (2021) 6634–6645.

[29] D. Gao, K. Li, R. Wang, S. Shan, X. Chen, Multi-modal graph neural network for joint reasoning on vision and scene text, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12746–12756.

[30] E.-S. Kim, W.Y. Kang, K.-W. On, Y.-J. Heo, B.-T. Zhang, Hypergraph attention networks for multimodal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14581–14590.

[31] Y. Zeng, D. Cao, X. Wei, M. Liu, Z. Zhao, Z. Qin, Multi-modal relational graph for cross-modal video moment retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2215–2224.

[32] C.-F. Chen, Q. Fan, N. Mallinar, T. Sercu, R. Feris, Big-little net: An efficient multi-scale feature representation for visual and speech recognition, in: Proceedings of the International Conference on Learning Representations, 2021, pp. 1–20.

[33] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Proceedings of the 14th Computer Vision–ECCV, Springer, 2016, pp. 483–499.

[34] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, J. Feng, Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3435–3444.

[35] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.

[36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[37] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.

[38] S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3883–3891.

[39] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[40] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5386–5395.

[41] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.

[42] S. Yang, D. Ramanan, Multi-scale recognition with DAG-CNNs, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1215–1223.

[43] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: Proceedings of the 14th Computer Vision–ECCV, Springer, 2016, pp. 354–370.

[44] R. Zheng, J. Chen, M. Ma, L. Huang, Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation, in: Proceedings of the International Conference on Machine Learning, PMLR, 2021, pp. 12736–12746.

[45] S. Kumar, A. Kulkarni, M.S. Akhtar, T. Chakraborty, When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5956–5968.

[46] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun, Attention bottlenecks for multimodal fusion, Adv. Neural Inf. Process. Syst. 34 (2021) 14200–14213.

[47] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, J. Huang, Deep multimodal fusion by channel exchanging, Adv. Neural Inf. Process. Syst. 33 (2020) 4835–4845.

[48] H.R.V. Joze, A. Shaban, M.L. Iuzzolino, K. Koishida, MMTM: Multimodal transfer module for CNN fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13289–13299.

[49] S. Shankar, Multimodal fusion via cortical network inspired losses, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 1167–1178.

[50] J. Wang, L. Sun, Y. Liu, M. Shao, Z. Zheng, Multimodal sarcasm target identification in tweets, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8164–8175.

[51] H. Liu, T. Lu, Y. Xu, J. Liu, W. Li, L. Chen, Camliflow: bidirectional camera-lidar fusion for joint optical flow and scene flow estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5791–5801.

[52] L. Liu, J. Chen, H. Wu, G. Li, C. Li, L. Lin, Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4823–4833.

[53] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, B. Gong, Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, Adv. Neural Inf. Process. Syst. 34 (2021) 24206–24221.

[54] S. Reed, K. Zolna, E. Parisotto, S.G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J.T. Springenberg, et al., A generalist agent, 2022, pp. 1–42, 11.

[55] H. Liu, Y. Tong, P. Zhang, X. Lu, J. Duan, H. Xiong, Hydra: A personalized and context-aware multi-modal transportation recommendation system, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2314–2324.

[56] H. Liu, Y. Tong, J. Han, P. Zhang, X. Lu, H. Xiong, Incorporating multi-source urban data for personalized and context-aware multi-modal transportation recommendation, IEEE Trans. Knowl. Data Eng. 34 (2) (2020) 723–735.

[57] H. Liu, J. Han, Y. Fu, J. Zhou, X. Lu, H. Xiong, Multi-modal transportation recommendation with unified route representation learning, Proc. VLDB Endow. 14 (3) (2020) 342–350.

[58] Q. Sun, Y. Wang, C. Xu, K. Zheng, Y. Yang, H. Hu, F. Xu, J. Zhang, X. Geng, D. Jiang, Multimodal dialoguer response generation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 2854–2866.

[59] D. Zhu, M. Zahran, L.E. Li, M. Elhoseiny, Halentnet: Multimodal trajectory forecasting with hallucinative intents, in: Proceedings of the International Conference on Learning Representations, 2020.

[60] J. Ke, S. Feng, Z. Zhu, H. Yang, J. Ye, Joint predictions of multi-modal ride-hailing demands: A deep multi-task multi-graph learning-based approach, Transp. Res. C 127 (2021) 103063.

[61] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210.

[62] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, V. Chandra, Federated learning with non-iid data, 2018, arXiv preprint arXiv:1806.00582.

[63] P. Chen, S. Liu, H. Zhao, J. Jia, Distilling knowledge via knowledge review, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5008–5017.

[64] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, S.-L. Kim, Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data, 2018, arXiv preprint arXiv:1811.11479.

[65] X. Yao, C. Huang, L. Sun, Two-stream federated learning: Reduce the communication costs, in: Proceedings of the 2018 IEEE Visual Communications and Image Processing, VCIP, IEEE, 2018, pp. 1–4.

[66] X. Yao, T. Huang, C. Wu, R. Zhang, L. Sun, Towards faster and better federated learning: A feature fusion approach, in: Proceedings of the 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 175–179.

[67] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: Proceedings of the European Conference on Computational Learning Theory, Springer, 1995, pp. 23–37.

[68] L. Huang, Y. Yin, Z. Fu, S. Zhang, H. Deng, D. Liu, LoAdaBoost: Loss-based AdaBoost federated machine learning with reduced computational complexity on IID and non-IID intensive care data, Plos One 15 (4) (2020) e0230706.

[69] M. Mohri, G. Sivek, A.T. Suresh, Agnostic federated learning, in: Proceedings of the International Conference on Machine Learning, PMLR, 2019, pp. 4615–4625.

[70] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.

[71] T. Li, M. Sanjabi, A. Beirami, V. Smith, Fair resource allocation in federated learning, 2019, arXiv preprint arXiv:1905.10497.

[72] F. Sattler, K.-R. Müller, W. Samek, Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints, IEEE Trans. Neural Netw. Learn. Syst. 32 (8) (2020) 3710–3722.

[73] N. Shoham, T. Avidor, A. Keren, N. Israel, D. Benditkis, L. Mor-Yosef, I. Zeitak, Overcoming forgetting in federated learning on non-iid data, 2019, arXiv preprint arXiv:1910.07796.

[74] V. Smith, C.-K. Chiang, M. Sanjabi, A.S. Talwalkar, Federated multi-task learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4427–4437.

[75] X. Yang, B. Xiong, Y. Huang, C. Xu, Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 3063–3071.

[76] X. Wei, A multi-modal heterogeneous data mining algorithm using federated learning, J. Eng. 2021 (8) (2021) 458–466.

[77] Y. Zhao, P. Barnaghi, H. Haddadi, Multimodal federated learning on iot data, in: 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation, IoTDI, IEEE, 2022, pp. 43–54.

[78] L. Zong, Q. Xie, J. Zhou, P. Wu, X. Zhang, B. Xu, FedCMR: Federated cross-modal retrieval, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1672–1676.

[79] H. Mo, H. Zheng, M. Gao, X. Feng, Multi-source heterogeneous data fusion based on federated learning, J. Comput. Res. Dev. 59 (2) (2022) 10.

[80] W. Wang, Y. Wang, Y. Huang, C. Mu, Z. Sun, X. Tong, Z. Cai, Privacy protection federated learning system based on blockchain and edge computing in mobile crowdsourcing, Comput. Netw. 215 (2022) 109206.

[81] B. Xiong, X. Yang, F. Qi, C. Xu, A unified framework for multi-modal federated learning, Neurocomputing 480 (2022) 110–118.

[82] A. Psaltis, C.Z. Patrikakis, P. Daras, Deep multi-modal representation schemes for federated 3d human action recognition, in: Proceedings of the European Conference on Computer Vision, Springer, 2022, pp. 334–352.

[83] A. Nandi, F. Xhafa, A federated learning method for real-time emotion state classification from multi-modal streaming, Methods 204 (2022) 340–347.

[84] A. Nandi, F. Xhafa, L. Subirats, S. Fort, Federated learning with exponentially weighted moving average for real-time emotion classification, in: Proceedings of the International Symposium on Ambient Intelligence, Springer, 2022, pp. 123–133.

[85] B. Salehi, J. Gu, D. Roy, K. Chowdhury, Flash: Federated learning for automated selection of high-band mmwave sectors, in: Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications, IEEE, 2022, pp. 1719–1728.

[86] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Federated learning for vision-and-language grounding problems, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11572–11579.

[87] H. Tang, C. Yu, X. Lian, T. Zhang, J. Liu, Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression, in: Proceedings of the International Conference on Machine Learning, PMLR, 2019, pp. 6155–6165.

[88] J. Xu, W. Du, Y. Jin, W. He, R. Cheng, Ternary compression for communication-efficient federated learning, IEEE Trans. Neural Netw. Learn. Syst. 33 (3) (2020) 1162–1176.

[89] A. Huang, Y. Chen, Y. Liu, T. Chen, Q. Yang, RPN: A residual pooling network for efficient federated learning, in: Proceedings of the European Conference on Artificial Intelligence, 2020, pp. 1223–1229.

[90] M. Asad, A. Moustafa, T. Ito, Fedopt: Towards communication efficiency and privacy preservation in federated learning, Appl. Sci. 10 (8) (2020) 2864.

[91] Y. Chen, Y. Ning, M. Slawski, H. Rangwala, Asynchronous online federated learning for edge devices with non-iid data, in: Proceedings of the 2020 IEEE International Conference on Big Data, IEEE, 2020, pp. 15–24.

[92] C. Xu, Y. Qu, Y. Xiang, L. Gao, Asynchronous federated learning on heterogeneous devices: A survey, Comp. Sci. Rev. 50 (2023) 100595.

[93] W. Wu, L. He, W. Lin, R. Mao, C. Maple, S. Jarvis, SAFA: A semi-asynchronous protocol for fast federated learning with low overhead, IEEE Trans. Comput. 70 (5) (2020) 655–668.

[94] S. Chen, B. Li, Towards optimal multi-modal federated learning on non-IID data with hierarchical gradient blending, in: Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications, IEEE, 2022, pp. 1469–1478.

[95] J. Ji, D. Yan, Z. Mu, Personnel status detection model suitable for vertical federated learning structure, in: Proceedings of the 2022 6th International Conference on Machine Learning and Soft Computing, 2022, pp. 98–104.

[96] X. Wei, L. Zhou, AI-enabled cross-modal communications, IEEE Wirel. Commun. 28 (4) (2021) 182–189.

[97] A. Qayyum, K. Ahmad, M.A. Ahsan, A. Al-Fuqaha, J. Qadir, Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge, IEEE Open J. Comput. Soc. 3 (2022) 172–184.

[98] V.S. Parekh, S. Lai, V. Braverman, J. Leal, S. Rowe, J.J. Pillai, M.A. Jacobs, Cross-domain federated learning in medical imaging, 2021, arXiv preprint arXiv:2112.10001.

[99] B.L.Y. Agbley, J. Li, A.U. Haq, E.K. Bankas, S. Ahmad, I.O. Agyemang, D. Kulevome, W.D. Ndiaye, B. Cobbinah, S. Latipova, Multimodal melanoma detection with federated learning, in: Proceedings of the 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP, IEEE, 2021, pp. 238–244.

[100] J. Wang, G. Xie, Y. Huang, J. Lyu, F. Zheng, Y. Zheng, Y. Jin, FedMed-GAN: Federated domain translation on unsupervised cross-modality brain image synthesis, Neurocomputing 546 (2023) 126282.

[101] T. Bernecker, A. Peters, C.L. Schlett, F. Bamberg, F. Theis, D. Rueckert, J. Weiß, S. Albarqouni, Fednorm: Modality-based normalization in federated learning for multi-modal liver segmentation, 2022, arXiv preprint arXiv:2205.11096.

[102] Y. Tian, J. Wang, Y. Wang, C. Zhao, F. Yao, X. Wang, Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving, IEEE Trans. Intell. Veh. 7 (3) (2022) 456–465.

[103] Q. Yu, Y. Liu, Y. Wang, K. Xu, J. Liu, Multimodal federated learning via contrastive representation ensemble, 2023, arXiv preprint arXiv:2302.08888.

[104] J. Chu, J. Liu, H. Wang, H. Meng, Z. Gong, T. Li, Micro-supervised disturbance learning: A perspective of representation probability distribution, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2022) 7542–7558.

[105] I. Dayan, H.R. Roth, A. Zhong, A. Harouni, A. Gentili, A.Z. Abidin, A. Liu, A.B. Costa, B.J. Wood, C.-S. Tsai, et al., Federated learning for predicting clinical outcomes in patients with COVID-19, Nature Med. 27 (10) (2021) 1735–1743.

[106] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R.R. Colen, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (1) (2020) 12598.

[107] M. Adnan, S. Kalra, J.C. Cresswell, G.W. Taylor, H.R. Tizhoosh, Federated learning and differential privacy for medical image analysis, Sci. Rep. 12 (1) (2022) 1953.

[108] N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, S. Bakas, M.N. Galtier, B.A. Landman, K. Maier-Hein, et al., The future of digital health with federated learning, NPJ Digit. Med. 3 (1) (2020) 1–7.

[109] J. Mills, J. Hu, G. Min, Client-side optimization strategies for communication-efficient federated learning, IEEE Commun. Mag. 60 (7) (2022) 60–66.

[110] D. Qiao, G. Liu, S. Guo, J. He, Adaptive federated learning for non-convex optimization problems in edge computing environment, IEEE Trans. Netw. Sci. Eng. 9 (5) (2022) 3478–3491.

[111] M. Hao, H. Li, G. Xu, S. Liu, H. Yang, Towards efficient and privacy-preserving federated deep learning, in: Proceedings of the 2019 IEEE Eeternational Conference on Communications, IEEE, 2019, pp. 1–6.

[112] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, Y. Zhou, A hybrid approach to privacy-preserving federated learning, in: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 2019, pp. 1–11.

[113] J. Chu, H. Wang, H. Meng, P. Jin, T. Li, Restricted boltzmann machines with gaussian visible units guided by pairwise constraints, IEEE Trans. Cybern. 49 (12) (2018) 4321–4334.

[114] D. Wang, T. Li, P. Deng, F. Zhang, W. Huang, P. Zhang, J. Liu, A generalized deep learning clustering algorithm based on non-negative matrix factorization, ACM Trans. Knowl. Discov. Data 17 (7) (2023) 1–20.

[115] O. Aouedi, K. Piamrat, G. Muller, K. Singh, FLUIDS: Federated Learning with semi-supervised approach for Intrusion Detection System, in: Proceedings of the 2022 IEEE 19th Annual Consumer Communications & Networking Conference, CCNC, IEEE, 2022, pp. 523–524.