IEEE Open Journal of the
Computer Society

Received 9 April 2020; accepted 30 April 2020. Date of publication 5 August 2020;
date of current version 4 June 2020. The review of this paper was arranged by Guest Editors Peng Li, Amiya Nayak, and Milos Stojmenovic.

Digital Object Identifier 10.1109/OJCS.2020.2993259

Personalized Federated Learning for Intelligent IoT Applications: A Cloud-Edge Based Framework

QIONG WU, KAIWEN HE, AND XU CHEN (Member, IEEE)

School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

CORRESPONDING AUTHOR: XU CHEN (e-mail: chenxu35@mail.sysu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1001703, in part by the National Science Foundation of China under Grant 61972432, in part by the Program for Guangdong Introducing Innovative and Enterpreneurial Teams under Grant 2017ZT07X355, and in part by the Pearl River Talent Recruitment Program under Grant 2017GC010465.

ABSTRACT Internet of Things (IoT) have widely penetrated in different aspects of modern life and many intelligent IoT services and applications are emerging. Recently, federated learning is proposed to train a globally shared model by exploiting a massive amount of user-generated data samples on IoT devices while preventing data leakage. However, the device, statistical and model heterogeneities inherent in the complex IoT environments pose great challenges to traditional federated learning, making it unsuitable to be directly deployed. In this paper, we advocate a personalized federated learning framework in a cloud-edge architecture for intelligent IoT applications. To cope with the heterogeneity issues in IoT environments, we investigate emerging personalized federated learning methods which are able to mitigate the negative effects caused by heterogeneities in different aspects. With the power of edge computing, the requirements for fast-processing capacity and low latency in intelligent IoT applications can also be achieved. We finally provide a case study of IoT based human activity recognition to demonstrate the effectiveness of personalized federated learning for intelligent IoT applications.

INDEX TERMS Edge computing, federated learning, internet of things, personalization.

I. INTRODUCTION

The proliferation of smart devices, mobile networks and computing technology have sparked a new era of Internet of Things (IoT), which is poised to make substantial advances in all aspects of our modern life, including smart healthcare system, intelligent transportation infrastructure, etc [1]. With huge amounts of smart devices connected together in IoT, we are able to get access to massive user data to yield insights, train task-specified machine learning models and utimately provide high-quality smart services and products. To reap the benefits of IoT data, the predominant approach is to collect scattered user data to a central cloud for modeling and then transfer the trained model to user devices for task inferences. This kind of approach can be ineffective as data transmission and model transfer will result in high communication cost and latency [2]. Moreover, as the user-sensitive data are required to upload to the remote cloud, it may impose great privacy leakage risk. Under the increasing stringent data privacy protection legislation such as General Data Protection Regulation (GDPR) [3], the data movement would face unprecedented difficulties. An alternative is to train and update the model at each IoT device with its local data, in isolation from other devices. However, one key impediment of this approach lies in the high resource demand for deploying and training models on IoT devices with limited computational, energy and memory resources. Besides, insufficient data samples and local data shifts will lead to an even worse model.

A sophisticated solution to deal with distributed data training is federated learning which enables to collaboratively train a high-quality shared model by aggregating and averaging locally-computed updates uploaded by IoT devices [4]. The primary advantage of this approach is the decoupling of model training from the need for direct access to the training data, and thus federated learning is able to learn a satisfactory global model without compromising user data privacy. Nevertheless, there are three major challenges in the key aspects of federated learning process in the complex IoT environments, making it unsuitable to directly deploy federated learning in IoT applications.

These three challenges faced by federated learning can be summarized as (1) device heterogeneity, such as varying storage, computational and communication capacities; (2) statistical heterogeneity like the non-IID (a.k.a. non independent and identically distributed) nature of data generated from different devices; and (3) model heterogeneity, the situation where different devices want to customize their models adaptive to their application environments. Specifically, resource-constrained IoT devices will be only allowed to train lightweight models under certain network conditions and may further result in high communication cost, stragglers, and fault tolerance issues which can not be well handled by traditional federated learning. As federated learning focuses on achieving a high-quality global model by extracting common knowledge of all participating devices, it fails to capture the personal information for each device, resulting in a degraded performance for inference or classification. Furthermore, traditional federated learning requires all participating devices to agree on a common model for collaborative training, which is impractical in realistic complex IoT applications.

To tackle these heterogeneity challenges, one effective way is to perform personalization in device, data and model levels to mitigate heterogeneities and attain high-quality personalized model for each device. Due to its broad application scenarios (e.g., IoT based personalized smart healthcare, smart home services and applications, fine-grained location-aware recommendation services, and on-premise intelligent video analytics), personalized learning has recently attracted great attention [5], [6]. We investigate the emerging personalized federated learning approaches which can be the viable alternative to traditional federated learning and summarize them into four categories: federated transfer learning, federated meta learning, federated multi-task learning and federated distillation. These approaches are able to alleviate different kinds of heterogeneity issues in the complex IoT environments and can be promising enabling techniques for many emerging intelligent IoT applications.

In this paper, we propose a synergistic cloud-edge framework named PerFit for personalized federated learning which mitigates the device heterogeneity, statistical heterogeneity, and model heterogeneity inherent in IoT applications in a holistic manner. To tackle the high communication and computation cost issues in device heterogeneity, we resort to edge computing which brings the necessary on-demand computing power in the proximity of IoT devices [2]. Therefore, each IoT device can choose to offload its computationally-intensive learning task to the edge which fulfills the requirement for fast-processing capacity and low latency. Besides, edge computing can mitigate privacy concerns by storing the data locally in proximity (e.g., in the smart edge gateway at home for smart home applications) without uploading the data to the remote cloud [7]. Furthermore, privacy and security protection techniques such as differential privacy and homomorphic encryption can be adopted to enhance the privacy protection level. For statistical and model heterogeneities, this framework also enables that end devices and edge servers jointly train a global model under the coordination of a central cloud server in a cloud-edge paradigm. After the global model is trained by federated learning, at the device side, different kinds of personalized federated learning approaches can be then adopted to enable personalized model deployments for different devices tailored to their application demands. We further illustrate a representative case study based on a specific application scenario—IoT based activity recognition, which demonstrates the superior performance of PerFit for high accuracy and low communication overhead.

The remainder of this paper is organized as follows. The following section discusses the main challenges of federated learning in IoT environments. To cope with these challenges, we advocate a personalized federated learning framework based on cloud-edge architecture and investigate some emerging solutions to personalization. Then, we evaluate the performance of personalized federated learning methods with a motivating study case of human activity recognition. Finally, we conclude the paper.

II. MAIN CHALLENGES OF FEDERATED LEARNING IN IOT ENVIRONMENTS

In this section, we first elaborate the main challenges and the potential negative effects when using traditional federated learning in IoT environments.

A. DEVICE HETEROGENEITY

There are typically a large number of IoT devices that differ in hardware (CPU, memory), network conditions (3G, 4G, WiFi) and power (battery level) in IoT applications, resulting in diverse computing, storage and communication capacities. Thus, device heterogeneity challenges arise in federated learning, such as high communication cost, stragglers, and fault tolerance [8]. In federated setting, communication costs are the principal constraints considering the fact that IoT devices are frequently offline or on slow or expensive connections [9]. In the federated learning process performing a synchronous update, the devices with limited computing capacity could become stragglers as they take much longer to report their model updates than other devices in the same round. Moreover, participating devices may drop out the learning process due to poor connectivity and energy constraints, causing a negative effect on federated learning. As the stragglers and faults issues are very prevalent due to the device heterogeneity in complex IoT environments, it is of great significance to address the practical issues of heterogeneous device communication and computation resources in federated learning setting.

B. STATISTICAL HETEROGENEITY

Consider a supervised task with features x and labels y, the local data distribution of user i can be represented as

$\mathcal{P}_i(x, y)$. Due to users' different usage environments and patterns, the personally-generated data (x, y) from different devices may naturally exhibit the kind of non-IID distributions. As $\mathcal{P}_i(x, y) = \mathcal{P}_i(y|x)\mathcal{P}_i(x) = \mathcal{P}_i(x|y)\mathcal{P}_i(y)$, user data can be non-IID in many forms, such as feature distribution skew, label distribution skew and concept shift [10]. For example, in healthcare applications, the distributions of users' activity data differ greatly according to users' diverse physical characteristics and behavioral habits (feature distribution skew). Moreover, the number of data samples across devices may vary significantly [11]. This kind of statistical heterogeneity is pervasive in complex IoT environments. To address this heterogeneity challenge, the canonical federated learning approach, FederatedAveraging (FedAvg), is demonstrated to be able to work with certain non-IID data. However, FedAvg may lead to a severely degraded performance when facing highly skewed data distributions. Specifically, on the one hand, non-IID data will result in weight divergence between federated learning process and the traditional centralized training process, which indicates that Fedvg will finally obtain a worse model than centralized methods and thus result in poor performance [12]. On the other hand, FedAvg only learns the coarse features from IoT devices, while fails in learning the fine-grained information on a particular device.

C. MODEL HETEROGENEITY

In the original federated learning framework, participating devices have to agree on a particular architecture of the training model so that the global model can be effectively obtained by aggregating the model weights gathered from local models. However, in practical IoT applications, different devices want to craft their own models adaptive to their application environments and resource constraints (i.e., computing capacity). And, they may be not willing to share the model details due to privacy concerns. As a consequence, the model architectures from different local models exhibit various shapes, making it impossible to perform naive aggregation by traditional federated learning [13]. In this case, the problem of model heterogeneity turns to become how to enable a deep network to understand the knowledge of others without sharing data or model details. Model heterogeneity inherent in IoT environments has attracted considerable research attention due to its practical significance for intelligent IoT applications.

III. CLOUD-EDGE FRAMEWORK FOR PERSONALIZED FEDERATED LEARNING

As elaborated in Section II, there exist device heterogeneity, statistical heterogeneity and model heterogeneity in IoT applications, which poses great challenges to traditional federated learning. An effective solution for addressing those heterogeneity issues can boil down to personalization. By devising and leveraging more advanced federated learning methods, we aim to enable the great flexibility such that individual devices can craft their own personalized models to meet their resource and application requirements and meanwhile enjoy the benefit from federated learning for collective knowledge sharing.

In this paper, we advocate a personalized federated learning framework for intelligent IoT applications to tackle the heterogeneity challenges in a holistic manner. As depicted in Fig. 1, our proposed PerFit framework adopts a cloud-edge architecture, which brings necessary on-demand edge computing power in the proximity of IoT devices. Therefore, each IoT device can choose to offload its intensive computing tasks to the edge (i.e., edge gateway at home, edge server at office, or 5G MEC server outdoors) via the wireless connections, thus the requirements for high processing efficiency and low latency of IoT applications can be fulfilled.

To support collaborative learning for intelligent IoT applications, federated learning (FL) is then adopted between end devices, edge servers and the remote cloud, which enables to jointly train a shared global model by aggregating locally-computed models from the IoT users at the edge while keeping all the sensitive data on device. To tackle the heterogeneity issues, we will further carry out personalization and adopt some personalized federated learning methods to fine tune the learning model for each individual device.

Specifically, the collaborative learning process in PerFit mainly consists of the following three stages as depicted in Fig. 1:

- *Offloading stage:* When the edge is trustworthy (e.g., edge gateway at home), the IoT device user can offload its whole learning model and data samples to the edge for fast computation. Otherwise, the device user will carry out model partitioning by keeping the input layers and its data samples locally on its device and offloading the remaining model layers to the edge for device-edge collaborative computing [14].
- *Learning stage:* The device and the edge collaboratively compute the local model based on personal data samples and then transmit the local model information to the cloud server. The cloud server aggregates local model information submitted by participating edges and averages them into a global model to send back to edges. Such model information exchanging process repeats until it converges after a certain number of iterations. Thus, a high-quality global model can be achieved and then transmitted to the edges for further personalization.
- *Personalization stage:* To capture the specific personal characteristics and requirements, each device will train a personalized model based on global model information and its own personal information (i.e., local data). The specific learning operations at this stage depend on the adopted personalized federated learning mechanism which will be elaborated in next section.

The proposed PerFit framework leverages edge computing to augment the computing capability of individual devices via computation offloading to mitigate the straggle effect. If we further conduct local model aggregation at the edge server, it also helps to reduce the communication overhead by avoiding massive devices to directly communicate with the cloud server over the expensive backbone network bandwidth [15]. Moreover, by performing personalization, we can

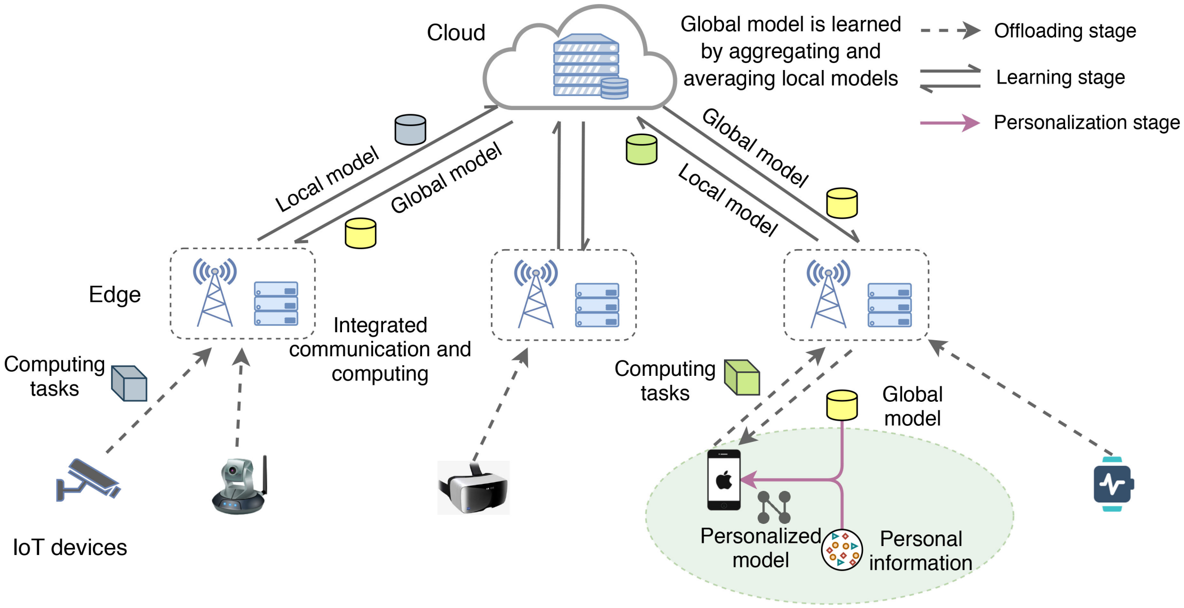


FIGURE 1. **The personalized federated learning framework for intelligent IoT applications, which supports flexible selection of personalized federated learning approaches.**

deploy lightweight personalized models at some resource-limited devices (e.g., by model pruning or transfer learning). These would help to mitigate the device heterogeneity in communication and computation resources. Also, the statistical heterogeneity and model heterogeneity can be well supported, since we can leverage personalized models and mechanisms for different individual devices tailored to their local data characteristics, application requirements and deployment environments.

Note that the adopted personalized federated learning mechanism will be the core of the collaborative learning in PerFit, which also determines the exchanging model information between the cloud server and the edges. For example, it is also allowed to transmit only part of the model parameters due to the specific setting of federated transfer learning as we will discuss in the coming section. If facing the situation where different models are trained on different IoT devices, the output class probabilities of local models can be encapsulated as its local information to send to the cloud server via federated distillation approaches. PerFit is flexible to integrate with many kinds of personalized federated methods by exchanging different kinds of model information between the edges and cloud accordingly. By addressing the heterogeneity issues inherent in the complex IoT environments and ensuring user privacy by default, PerFit can be ideal for large-scale practical deployment.

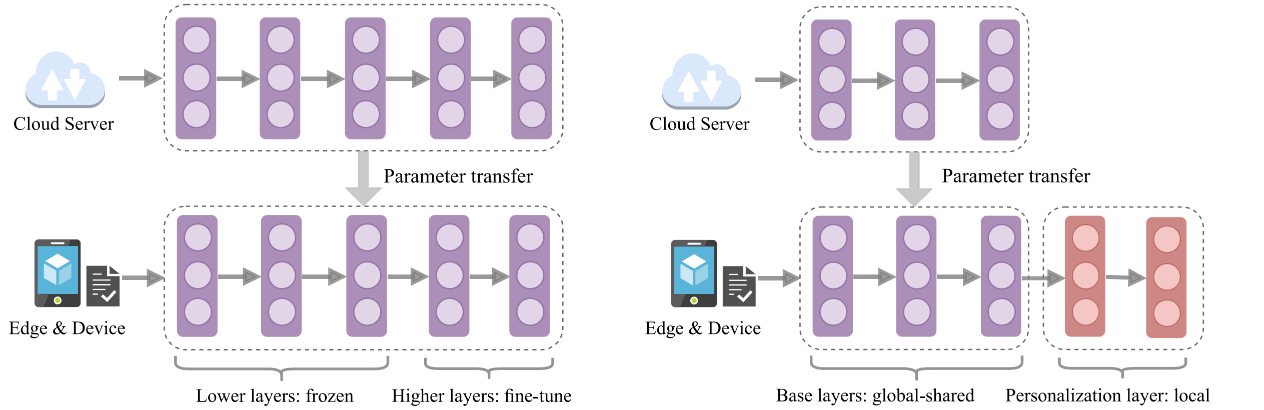
IV. PERSONALIZED FEDERATED LEARNING MECHANISMS

In this section, we review and elaborate several key personalized federated learning mechanisms that can be integrated with PerFit framework for intelligent IoT applications. These personalized federated learning schemes can be categorized by federated transfer learning, federated meta learning, federated multi-task learning, and federated distillation, which will be elaborated as follows.

A. FEDERATED TRANSFER LEARNING

Transfer learning [16] aims at transferring knowledge (i.e., the trained model parameters) from a source domain to a target domain. In the setting of federated learning, the domains are often different but related, which makes knowledge transfer possible. The basic idea of federated transfer learning is to transfer the globally-shared model to distributed IoT devices for further personalization in order to mitigate the statistical heterogeneity (non-IID data distributions) inherent in federated learning. Considering the architecture of deep neural networks and communication overload, there are two main approaches to perform personalization via federated transfer learning.

Chen *et al.* [17] first train a global model through traditional federated learning and then transfer the global trained model back to each device. Accordingly, each device is able to build personalized model by refining the global model with its local data. To reduce the training overhead, only model parameters of specified layers will be fine-tuned instead of retraining whole model. As presented in Fig. 2(a), model parameters in lower layers of global model can be transferred and reused directly for local model as lower layers of deep networks focus on learning common and low-level features. While the model parameters in higher layers should be fine-tuned with local data as they learn more specific features tailored to current device. Besides, Feng *et al.* [18] design two personal adaptors (personal bias, personal filter) for higher layers in user's local model which can be fine-tuned with personal information.



(a) The whole trained global model in the cloud server is transferred to the device for personalization with its local data.

(b) The device model is combined with the part of model transferred from cloud server and the personalization layers owned by users locally.

FIGURE 2. **Federated transfer learning.**

Arivazhagan *et al.* [19] propose FedPer which takes a different way to perform personalization through federated transfer learning. FedPer advocates viewing deep learning models as *base* + *personalization* layers as illustrated in Fig. 2(b). Base layers act as the shared layers which are trained in a collaborative manner using the existing federated learning approach (i.e., FedAvg method). While the personalization layers are trained locally thereby enabling to capture personal information of IoT devices. In this way, after the federated training process, the globally-shared base layers can be transferred to participating IoT devices for constituting their own personalized deep learning models with their unique personalization layers. Thus, FedPer is able to capture the fine-grained information on a particular device for superior personalized inference or classification, and address the statistical heterogeneity to some extent. Besides, by uploading and aggregating only part of the models, FedPer requires less computation and communication overhead, which is essential in IoT environments.

Note that subject to the computing resource constraint of the device, model pruning and compression techniques can be further leveraged to achieve the lightweight model deployment after the personalized model is obtained.

B. FEDERATED META LEARNING

Federated learning in IoT environments generally faces statistical heterogeneity such as non-IID and unbalanced data distributions, which makes it challenging to ensure a high-quality performance for each participating IoT devices. To tackle this problem, some researchers concentrate on improving FedAvg algorithm by leveraging the personalization power of meta learning. In meta learning, the model is trained by a meta-learner which is able to learn on a large number of similar tasks and the goal of the trained model is to quickly adapt to a new similar task from a small amount of new data [20]. By regarding the similar tasks in meta learning as the personalized models for the devices, it is a natural choice to integrate federated learning with meta learning to achieve personalization through collaborative learning.

Jiang *et al.* [21] propose a novel modification of FedAvg algorithm named Personalized FedAvg by introducing a fine-tuning stage using model agnostic meta learning (MAML), a representative gradient-based meta learning algorithm. Thus, the global model trained by federated learning can be personalized to capture the fine-grained information for individual devices, which results in an enhanced performance for each IoT device. MAML is flexible to combine with any model representation that is amenable to gradient-based training. Besides, it can learn and adapt quickly from only a few data samples.

Since the federated meta learning approach often utilizes complicated training algorithms, it has higher implementation complexity than the federated transfer learning approach. Nevertheless, the learned model by federated meta learning is more robust and can be very useful for those devices with very few data samples.

C. FEDERATED MULTI-TASK LEARNING

In general, federated transfer learning and federated meta learning aim to learn a shared model of the same or similar tasks across the IoT devices with fine-tuned personalization. Along a different line, federated multi-task learning aims at learning distinct tasks for different devices simultaneously and tries to capture the model relationships amongst them without privacy risk [22]. Through model relationships, the model of each device may be able to reap other device's information. Moreover, the model learned for each device is always personalized. As shown in Fig. 3, in the training process of federated multi-task learning, the cloud server learns the model relationships amongst multiple learning tasks based on the uploaded model parameters by IoT devices. And, then each device can update its own model parameters with its local data and current model relationships. Through the alternating

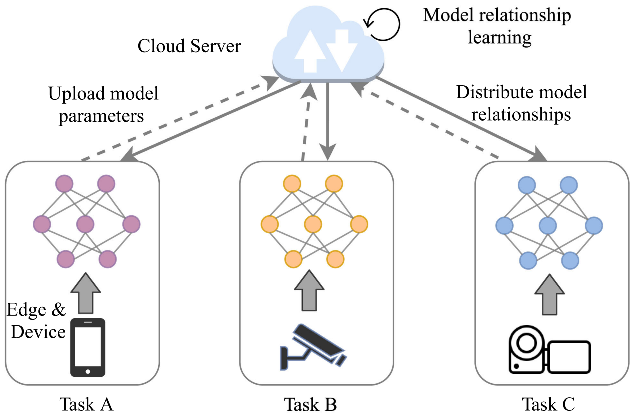


FIGURE 3. **Federated multi-task learning.**

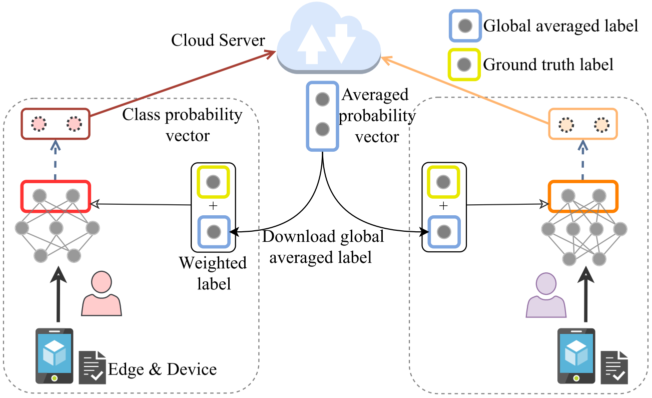


FIGURE 4. **Federated distillation.**

optimization of model relationships in the cloud server and model parameters for each task, federated multi-task learning enables participating IoT deivices to collaboratively train their local models so as to mitigate statistical heterogeneity and obtain high-quality personalized models.

Smith *et al.* [8] develop a distributed optimization method MOCHA through a federated multi-task learning framework. For high communication cost, MOCHA allows the flexibility of computation which yields direct benefits for communication as performing additional local computation will result in fewer communication rounds in federated settings. To mitigate stragglers, the authors propose to approximately compute the local updates for devices with limited computing resources. Besides, asynchronous updating scheme is also an alternative approach for straggler avoidance. Furthermore, by allowing participating devices periodically dropping out, MOCHA is robust to fault tolerance. As device heterogeneity inherent in complex IoT environments is critical to the performance of federated learning, federated multi-task learning is of great significance for intelligent IoT applications. Nevertheless, as federated multi-task learning produces one model per task, it requires that all clients (e.g., IoT devices) participate in every iteration which is impractical in IoT applications. To tackle this issue, we believe that cluster-based federated multi-task learning is a promising direction in research.

D. FEDERATED DISTILLATION

In original federated learning framework, all clients (e.g., participating edges and devices) have to agree on a particular architecture of the model trained on both the global server and local clients. However, in some realistic business setting, like healthcare and finance, each participant would have capacity and desire to design its own unique model, and may not be willing to share the model details due to privacy and intellectual property concerns. This kind of model heterogeneity poses new challenge to traditional federated learning.

To tackle this challenge, Li *et al.* [23] propose FedMD, a new federated learning framework that enables participants to independently design their own models by leveraging the power of knowledge distillation. In FedMD, each client needs to translate its learned knowledge to a standard format which can be understood by others without sharing data and model architecture. And, then a central server collects these knowledges to compute a consensus which will be further distributed to the participating clients. The knowledge translation step can be implemented by knowledge distillation, for example, using the class probabilities produced by client model as the standard format as shown in Fig. 4. In this way, the cloud server aggregates and averages the class probabilities for each data sample and then distributes to clients to guide their updates. Jeong *et al.* [24] propose federated distillation where each client treats itself as a student and sees the mean model output of all the other clients as its teacher's output. The teacher-student output difference provides the learning direction for the student. Here, it is worthnoting that, to operate knowledge distillation in federated learning, a public dataset is required because the teacher and student outputs should be evaluated using an identical training data sample. Moreover, federated distillation can significantly reduce the communication cost as it exchanges not the model parameters but the model outputs [25].

E. DATA AUGMENTATION

As user's personally-generated data naturally exhibits the kind of highly-skewed and non-IID distribution which may greatly degrade the model performance, there are emerging works focusing on data augmentation to facilitate personalized federated learning. Zhao *et al.* [12] propose a data-sharing strategy by distributing a small amount of global data containing a uniform distribution over classes from the cloud to the edge clients. In this way, the highly-unbalanced distribution of client data can be alleviated to some extent and then the model performance of personalization can be improved. However, directly distributing the global data to edge clients will impose great privacy leakage risk, this approach is required to make a trade-off between data privacy protection and performance improvement. Moreover, the distribution difference between global shared data and user's local data can also bring performance degradation.

To rectify the unbalanced and non-IID local dataset without compromising user privacy, some over-sampling techniques and deep learning approaches with generative ability are adopted. For example, Jeong *et al.* [24] propose federated augmentation (FAug), where each client collectively trains a generative model, and thereby augments its local data towards yielding an IID dataset. Specifically, each edge client recognizes the labels being lacking in its data samples, referred to as target labels, and then uploads few seed data samples of these target labels to the server. The server oversamples the uploaded seed data samples and then trains a generative adversarial network (GAN). Finally, each device can download the trained GAN's generator to replenish its target labels until reaching a balanced dataset. With data augmentation, each client can train a more personalized and accurate model for classification or inference based on the generated balanced dataset. It is worthnoting that the server in FAug should be trustworthy so that users are willing to upload their personal data.

V. CASE STUDY

In this section, we first describe the experiment settings and then evaluate different personalized federated learning approaches with different kinds of heterogeneities in terms of accuracy and comminication size.

A. DATASET DESCRIPTION AND IMPLEMENTATION DETAILS

In the experiments, we focus on human activity recognition task based on a publicly accessible dataset called MobiAct [26]. Each volunteer participating in the generation of MobiAct dataset wears a Samsung Galaxy S3 smartphone with accelerometer and gyroscope sensors. The tri-axial linear accelerometer and angular velocity signals are recorded by embedded sensors while volunteers perform predefined activities. We use an 1-second sliding window for feature extraction since one second is enough to perform an activity. There are ten kinds of activities recorded in MobiAct, such as walking, stairs up/down, falls, jumping, jogging, step in a car, etc. To practically mimic the environment of federated learning, we randomly select 30 volunteers and regard them as different clients. For each client, we take a random number of samples for each activity and finally, each client has 480 samples for model training. In this way, the personal data of different clients may exhibit the kind of non-IID distributions (statistical heterogeneity). The test data for each client is composed of 160 samples under a balanced distribution.

In order to meet the needs of different clients for customizing their own models (model heterogeneity) in IoT applications, we design two kinds of models for training on the clients: 1) a Multi-Layer Perceptron network composed of three fully-connected layers with 400, 100 and 10 neural units (521,510 total parameters), which we refer to as the 3NN, 2) a convolutional neural network (CNN) with three 3×3 convolutional layers (the first with 32 channels, the second with 16, the last with 8, each of the first two layers followed by a 2×2 max-pooling layer), a fully-connected layer with 128 units and $ReLu$ activation, and a final $Softmax$ output layer (33,698 total parameters). Cross-entropy loss and Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 are used for the training of both 3NN and CNN.

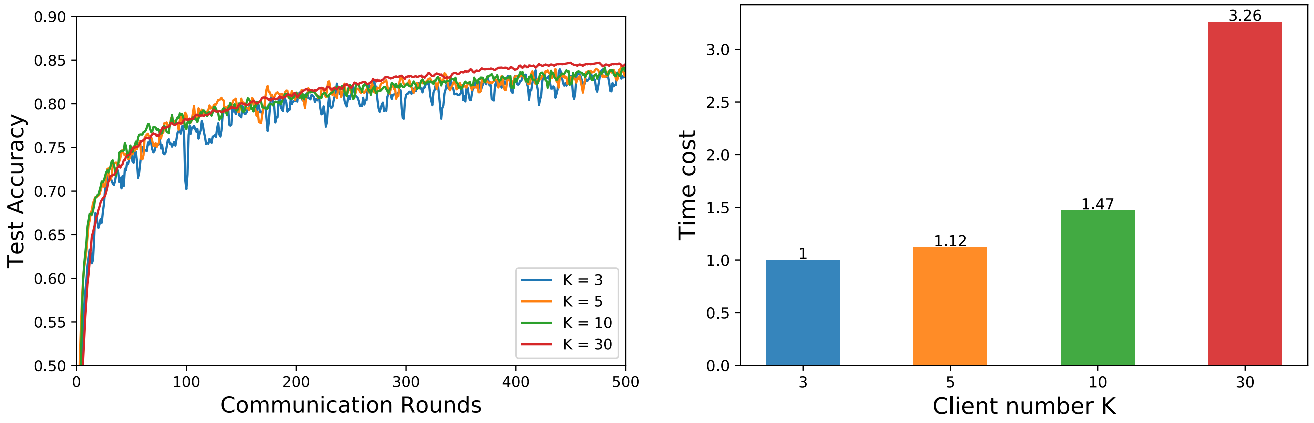
B. EXPERIMENTAL RESULTS

1) COMPARING METHODS

We compare the performance of personalized federated learning with both centralized scheme and traditional federated learning. For centralized methods, we adopt the widely-used machine learning approaches in human activity recognition task such as support vector machine (SVM) [27], k-nearest neighbor (kNN) [28], and random forest (RF) [29]. Besides, centralized 3NN (c3NN) and centralized CNN (cCNN) are also used for comparison. As centralized approaches require a large amount of data, we collect all the training data of 30 users for model learning. In traditional federated settings, each client trains a local model (e.g., 3NN or CNN in our experiment) with its personal-generated data. FedAvg method [9], which aggregates local model updates on each client and then sends them to a cloud server that performs model averaging in an iterative way, is applied to train the global model. Then, the well-trained global model in the cloud is directly distributed to clients for human activity recognition. As for personalized FL, we study the performance of the two widely-adopted approaches: federated transfer learning (FTL) and federated distillation (FD). For FTL, each client will fine tune the model downloaded from the cloud server with its personal data. While in FD, each client can customize its own model according to its own requirements. Note that each client is able to offload its learning task from its device to the edge in proximity (e.g., edge gateway at home) for fast computation in our cloud-edge paradigm.

2) PERFORMANCE EVALUATION

As elaborated in Section II-A, due to the device heterogeneity (communication and computing resources constraints of IoT devices), there are only a few clients participating in the global model learning in each communication round. Thus, we first experiment with the number of participating clients K in each round. We set K equal to 3, 5, 10, and 30, which means that $\frac{1}{10}$, $\frac{1}{6}$, $\frac{1}{3}$ and 100% of users participating in the federated learning process in each communication round. As depicted in Fig. 5(a), for all values of K, the test accuracy improves with the number of communication rounds increases and the test accuracies are similar when the training process converges. However, when K is small, the learning curve exists erratic fluctuation to some extent. As K increases, the learning curve becomes smoother and smoother. Although the test accuracies are similar, the training time for each value of K varies dramatically as demonstrated in Fig. 5(b). For example, the training time for $K = 30$ is 3.26 times longer than that in the $K = 3$ case. We make a trade-off between the stability and



(a) Test Accuracy

(b) Time Cost

FIGURE 5. The test accuracy and time cost under different number of participating clients in each communication round. We choose K = 5 by making a trade off between the stability and the efficiency of the learning algorithm.

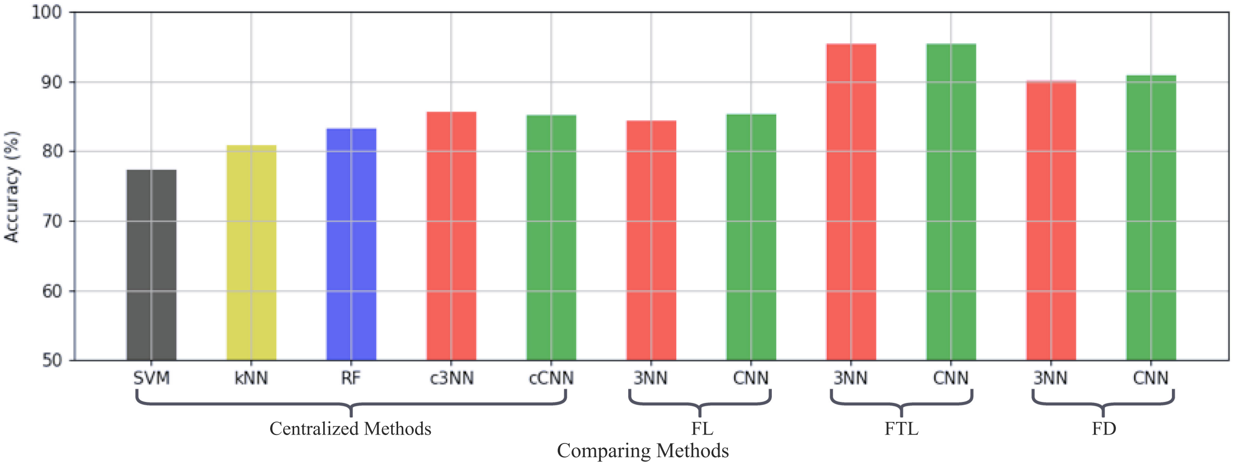


FIGURE 6. The accuracy of different learning methods in human activity recognition.

the efficiency for the training process and fix $K = 5$ for the following experiments. For each method, we compute the average of test accuracy by repeating the training and prediction processes five times.

Fig. 6 illustrates the test accuracy of 30 clients under different learning approaches. For centralized methods, deep learning based methods (c3NN, cCNN) can all achieve a high accuracy than traditional machine learning based methods (SVM, kNN, and RF). Under the coordination of a central cloud server, the edge clients in traditional federated learning (FL-CNN) are able to collectively reap the benefits of each other's information without compromising data privacy and achieve a competitive average accuracy of 85.22% similar to cCNN. The slight performance degradation in FL-3NN and FL-CNN compared with the centralized fashion results from the statistical heterogeneity inherent in federated learning settings. With personalized federated learning, both FTL and FD can capture user's fine-grained personal information and obtain a personalized model for each participant, leading to a higher test accuracy. For example, FTL-3NN can reach 95.37% accuracy, which is 11.12% higher than that of FL-3NN.

Furthermore, we take a more detailed observation to evaluate the performance of personalized federated learning. As shown in Fig. 7, we adopt boxplot to graphically depict the six-number summary of the accuracies of 30 paticipating users, which consists of the smallest observation, lower quartile, median, upper quartile, largest observation, and the mean represented by green triangle. We can see that although the average performance of FL-CNN is similar with cCNN, the global model trained by FL may perform poorly on some clients. For example, the accuracy of some clients may be lower than 70% while some clients can reach a high accuracy of more than 95%. With personalization performed by each client with its own data, the accuracies of 30 clients vary in a very small scale which indicates that personalization

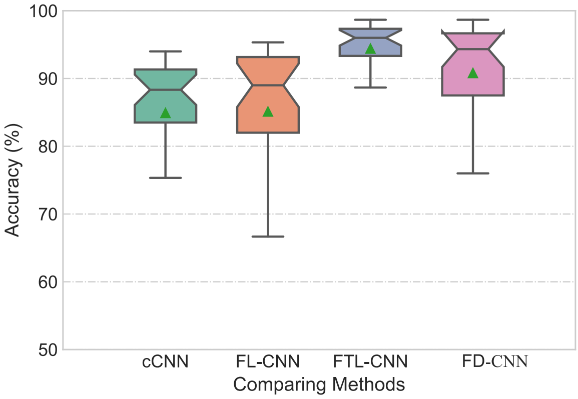


FIGURE 7. The accuracy distribution of different clients predicted by CNN under different learning schemes.

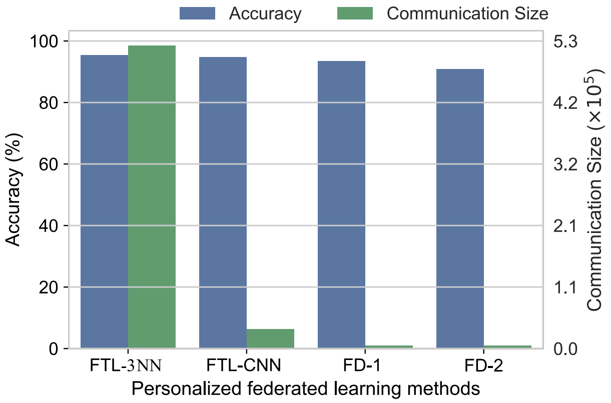


FIGURE 8. The accuracy and communication size of different implementations for federated transfer learning and federated distillation.

can significantly reduce the performance degradation caused by non-IID distribution. FD-CNN approach has an accuracy improvement of 5.69% compared with FL-CNN and the performance differences between different clients have also been narrowed. This observatiion indicates that PFL can benefit most of the participating clients and thus will encourage user engagement.

The critical nature of communication constraints in cloud-edge scenarios also needs to be considered in federated setting because of limited bandwidth, slow and expensive connections. We compare both the accuracy and communication data size of different training models for FTL and FD. In FTL-3NN and FTL-CNN, we utilize 3NN and CNN as the model trained on both the cloud and the edge clients, respectively. For federated distillation, we consider two cases of model heterogeneity: (1) FD-1: 10 clients choose 3NN as their local models while the remaining 20 clients choose CNN; (2) FD-2: the local models of 20 clients are 3NN and the models for remaining 10 clients are CNN. As depicted in Fig. 8, all the four personalized federated learning methods can achieve a high accuracy of more than 90%. However, the communication sizes vary dramatically. As all these methods can converge within hundreds of communication rounds, we only compare the communication size in each communication round. The commnication payload size for FTL depends on the model parameter number which are 521,510 and 33,698 for FTL-3NN and FTL-CNN, respectively. While the communication size for FD is proportional to the output dimension which is 10 in our human activity recognition task. In each communication round, we randomly select 500 samples from the globally-shared data and transmit the outputted class scores predicted by each participating device to the cloud server, thus the communication size for both FD-1 and FD-2 is 5000. Fig. 8 states that we are able to achieve superior prediction performance with lightweight models and small communication overhead, which is of great significance for supporting large-scale intelligent IoT applications.

VI. CONCLUSION

In this paper, we propose PerFit, a personalized federated learning framework in a cloud-edge architecture for intelligent IoT applications with data privacy protection. PerFit enables to learn a globally-shared model by aggregating local updates from distributed IoT devices and leveraging the merits of edge computing. To tackle the device, statistical, and model heterogeneities in IoT environments, PerFit can naturally integrate a variety of personalized federated learning methods and thus achieve personalization and enhanced performance for devices in IoT applications. We demonstrate the effectiveness of PerFit through a case study of human activity recognition task, which corroborates that PerFit can be a promising approach for enabling many intelligent IoT applications.

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] Z. Zhou, Xu Chen, En Li, L. Zeng, Ke Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[3] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide*, 1st Ed., Berlin, Germany: Springer International Publishing, 2017.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[5] P. Vanhaesebrouck, A. Bellet, and M. Tommasi, "Decentralized collaborative learning of personalized models over networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2017, pp. 509–517.

[6] A. Bellet, R. Guerraoui, M. Taziki, and M. Tommasi, "Personalized and private peer-to-peer machine learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 473–481.

[7] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[8] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Advances Neural Inform. Process. Syst.*, 2017, pp. 4424–4434.

[9] H B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[10] P. Kairouz *et al.*, "Advances and open problems in federated learning," *CoRR*, vol., abs/1912.04977, 2019. [Online]. Available: http://arxiv.org/abs/1912.04977

[11] J. Xu and F. Wang, “Federated learning for healthcare informatics,” *CoRR*, vol. abs/1911.06270, 2019. [Online]. Available: http://arxiv.org/abs/1911.06270.

[12] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, “Federated learning with non-iid data,” *CoRR*, vol. abs/1806.00582, 2018. [Online]. Available: http://arxiv.org/abs/1806.00582

[13] T. Li, Anit K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[14] En Li, L. Zeng, Z. Zhou, and X. Chen, “Edge AI: On-demand accelerating deep neural network inference via edge computing,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.

[15] S. Luo, Xu Chen, Q. Wu, Z. Zhou, and S. Yu, “Hfel: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning,” *CoRR*, vol. abs/2002.11343, 2020. [Online]. Available: https://arxiv.org/abs/2002.11343

[16] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[17] Y. Chen, J. Wang, C. Yu, W. Gao, and X. Qin, “Fedhealth: A federated transfer learning framework for wearable healthcare,” *IEEE Intell. Syst.*, 2020.

[18] J. Feng, C. Rong, F. Sun, D. Guo, and Y. Li, “PMF: A privacy-preserving human mobility prediction framework via federated learning,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–21, 2020.

[19] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, “Federated learning with personalization layers,” *CoRR*, vol. abs/1912.00818, 2019. [Online]. Available: http://arxiv.org/abs/1912.00818

[20] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 1126–1135.

[21] Y. Jiang, J. Konečnỳ, K. Rush, and S. Kannan, “Improving federated learning personalization via model agnostic meta learning,” *CoRR*, vol. abs/1909.12488, 2019. [Online]. Available: http://arxiv.org/abs/1909.12488

[22] L. Corinzia and J. M. Buhmann, “Variational federated multi-task learning,” *CoRR*, vol. abs/1906.06268, 2019. [Online]. Available: http://arxiv.org/abs/1906.06268

[23] D. Li and J. Wang, “Fedmd: Heterogenous federated learning via model distillation,” *CoRR*, vol. abs/1910.03581, 2019. [Online]. Available: http://arxiv.org/abs/1910.03581

[24] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, “Communication-efficient on-device machine learning:Federated distillation and augmentation under non-iid private data,” *CoRR*, vol. abs/1811.11479, 2018. [Online]. Available: http://arxiv.org/abs/1811.11479.

[25] J.-H. Ahn, O. Simeone, and J. Kang, “Wireless federated distillation for distributed edge learning with heterogeneous data,” in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, 2019, pp. 1–6.

[26] G. Vavoulas, C. Chatzaki, T. Malliotakis, M. Pediaditis, and M. Tsiknakis, “The MobiAct dataset: Recognition of activities of daily living using smartphones,” in *Proc. ICT4AgeingWell*, 2016, pp. 143–151.

[27] J. A Ward, G. Pirkl, P. Hevesi, and P. Lukowicz, “Towards recognising collaborative activities using multiple on-body sensors,” in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.: Adjunct*, 2016, pp. 221–224.

[28] J. He, C. Hu, and X. Wang, “A smart device enabled system for autonomous fall detection and alert,” *Int. J. Distrib. Sensor Netw.*, vol. 12, no. 2, 2016, Art. no. 2308183.

[29] J. Yuan, Kok K. Tan, Tong H. Lee, and G. C. H. Koh, “Power-efficient interrupt-driven algorithms for fall detection and classification of activities of daily living,” *IEEE Sensors J.*, vol. 15, no. 3, pp. 1377–1387, Mar. 2014.

QIONG WU received the B.S. and M.E. degrees from the School of Data and Computer Science, Sun Yat-sen University (SYSU), Guangzhou, China in 2017 and 2019, respectively. She is currently pursuing the Ph.D. degree in the School of Data and Computer Science, SYSU. Her primary research interests include social data analysis, mobile edge computing, and federated learning.

KAIWEN HE received the B.S. degree from the School of Data and Computer Science, Sun Yat-sen University (SYSU), Guangzhou, China in 2018. She is currently pursuing the M.E. degree in the School of Data and Computer Science, SYSU. Her primary research interests include mobile data analysis, data mining, and deep learning.

XU CHEN (Member, IEEE) is a Full Professor with Sun Yat-sen University, Guangzhou, China, and the Vice Director of National and Local Joint Engineering Laboratory of Digital Home Interactive Applications. He received the Ph.D. degree in information engineering from the Chinese University of Hong Kong in 2012, and worked as a Postdoctoral Research Associate at Arizona State University, Tempe, USA, from 2012 to 2014, and a Humboldt Scholar Fellow at the Institute of Computer Science of the University of Goettingen, Germany from 2014 to 2016. He received the prestigious Humboldt research fellowship awarded by the Alexander von Humboldt Foundation of Germany, 2014 Hong Kong Young Scientist Runner-up Award, 2016 Thousand Talents Plan Award for Young Professionals of China, 2017 IEEE Communication Society Asia-Pacific Outstanding Young Researcher Award, 2017 IEEE ComSoc Young Professional Best Paper Award, Honorable Mention Award of 2010 IEEE international conference on Intelligence and Security Informatics (ISI), Best Paper Runner-up Award of 2014 IEEE International Conference on Computer Communications (INFOCOM), and Best Paper Award of 2017 IEEE Intranational Conference on Communications (ICC). He is currently an Area Editor of the IEEE OPEN JOURNAL OF THE Communications Society, an Associate Editor of the IEEE TRANSACTIONS WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) Series on Network Softwarization and Enablers.