1 Introduction

This chapter introduces the background and motivations of edge learning. We also specify the main challenges faced by edge learning ranging from challenges in data and computation, to communication.

1.1 Background

Machine learning has demonstrated great promises in various fields - e.g., smart healthcare, smart surveillance, smart homes, self-driving, and smart grid - which are fundamentally altering the way individuals and organizations live, work, and interact. Big data is one of the key promotion factors that boosts machine learning development, following the significant successes and progress of machine learning models (especially deep learning models) in many domains in recent years. Big data and machine learning are enabling technologies for smart decision making, automation, and resource optimization. These technologies collectively promote intelligent services from concepts to practical applications. Traditionally, to develop these intelligent services and applications, big data should be stored and processed in the cloud data center in a centralized mode. Due to the powerful capabilities of the cloud, it has enabled great achievements in learning from big data, especially for applications that allow long response delay and all data aggregated to the cloud - e.g., e-commerce services and recommendation systems.

However, with the growing workloads related to 5G, the Internet of Things (IoT), and real-time analytics, traditional centralized learning frameworks require all training data from different sources to be uploaded to a remote data server, which incurs significant communication overhead, service latency, as well as security and privacy issues. According to a report by Cisco, nearly 847 ZB of data will be generated at the edge while the storage capability of data centers will only reach to 19.5ZB by 2021. Moreover, plenty of emerging applications require a strict guarantee of response latency, e.g., self-driving and Industry 4.0 usually require millisecond-level or even microsecond latency.

Meanwhile, traditional distributed machine learning usually ignores the privacy and security issue. In recent years, with the development and wider application of artificial intelligence (AI) technology, data privacy protection has also received more and more attention [81]. The European Union has introduced the first data privacy

protection bill - the General Data Protection Regulation (GDPR), which clarified certain provisions on data privacy protection. The "Cyber Security Law of the People's Republic of China" and "General Principles of the Civil Law of the People's Republic of China," which were implemented in 2017 also state that "network operators must not disclose, tamper with, or destroy the personally collected information, and when conducting data transactions with third parties, it is necessary to ensure that the proposed contract clearly stipulates the scope of the proposed transaction data and data protection obligations." In view of these legislations, the collection of user data must be open and transparent between enterprises and institutions. Data cannot be exchanged without user authorization [357]. The challenge that brings to traditional machine learning is: if the data cannot be communicated between institutions, a company has a limited amount of data, or a few giant companies monopolize a large amount of data. It is difficult for small companies to obtain data, which results in data islands. To tackle this problem, many researchers pay much attention to studying the design and implementation of machine learning on the edge.

Therefore, it is urgent to shift model training and inference from the cloud to the edge. In fact, the wide deployment of edge devices promotes the significant increase of computing capacity in the edge environment, far exceeding the increasing speed of network bandwidth. From this aspect, edge devices can be viewed as the extension of the cloud because of the huge computing capacity. By taking the advantages of both cloud and edge, big data analytics could be more efficient. The edge learning paradigm - i.e., distributed machine learning over edge devices - enables distributed edge nodes to cooperatively train models and conduct inferences with their locally cached data.

To explore the new characteristics and potential prospects of edge learning, we will provide a comprehensive and systematic introduction of the recent research efforts on edge learning. We hope that this book will elicit escalating attention, stimulate fruitful discussions, and inspire further research ideas in this field.

1.2 From Cloud Learning to Edge Learning

In this part, we discuss the development of edge learning.

1.2.1 From Cloud Computing to Edge Computing

In the past decade, cloud computing has become a mature computing paradigm for providing powerful computation as well as various standardized services. Traditional cloud-based methods have enabled great achievements in learning from big data, especially for those that allow long response delay and data aggregation from edge to cloud [174]. However, with the growing workloads related to 5G, autonomy, the IoT, and real-time analytics, businesses are beginning to look elsewhere for their computing needs. In addition, the cloud has unsolvable challenges in communication, computation, and storage that cannot satisfy the requirement in explosively increasing data.

These trends change the end-user experience and introduce new requirements and demands on data center and cloud-based infrastructures [81, 154, 170, 205, 209, 221, 298, 357, 386, 406].

As a result, users and service providers are likely moving toward the edge environment; the basic concept is to reduce the amount of data sent to the cloud and to provide computation at the place nearby [42]. Recently, plenty of edge servers are deployed at the network edge. Meanwhile, mobile devices are equipped with increasingly powerful Central Processing Units (CPUs) and Graphic Processing Units (GPUs), which have capabilities to perform complex computing tasks. Thus, offloading some tasks into the network edge can mitigate the cloud burden and improve the performance of services in terms of latency and resource utilization.

1.2.2 From Distributed Machine Learning to Edge Learning

At its beginning, the machine learning model was trained in a single machine with limited hardware resources. The training algorithms of many machine learning models - including most neural networks, graphical models, etc. - can be abstracted into an iterative-convergent process, which takes a lot of computing resources to complete. Studies have shown that a CPU clocked at 2.0 GHz training a VGG-16 model on the ImageNet dataset will take several years, which is totally incompatible with the real-time nature of the model application! Due to the limited computing power of a single machine and the decentralized big training data itself, machine learning is developing in a distributed scenario, known as distributed machine learning. In distributed machine learning, multiple workers cooperate with each other with communication and train the model in parallel.

Though traditional cloud-based methods have made great achievements in learning from the big data, the cloud has limitations due to heavy resource cost, privacy issues, high latency, etc. On the other hand, the principle of edge computing naturally facilitates edge learning by leveraging resources of edge devices and mobile devices. Edge learning is a paradigm complementary to the cloud-based methods for big data analytics in the cloud-edge environment. It has been proposed and developed for moving the training and inference to the edge environment to serve delay-sensitive and privacy-sensitive applications, for which the data cannot be gathered in the cloud. In Fig. 1.1, we illustrate the cloud-edge environment. With the promotion and popularization of edge devices - e.g., smartphones, autonomous vehicles, sensors, and wearable devices - data generated at the network edge are exponentially increasing. Thus, performing the training tasks and the inference tasks of machine learning model at the network edge can mitigate the cloud burden and improve the performance of intelligent services in terms of latency and resource utilization. Distributed big data analytics in cloud-edge environments has become a new trend.

Notably, federated learning is a typical example of edge learning. Since first introduced by Google, federated learning has demonstrated to be a promising solution for future AI applications with strong privacy protection [198]. Federated learning allows users to collaboratively train a global model without sharing their own data,

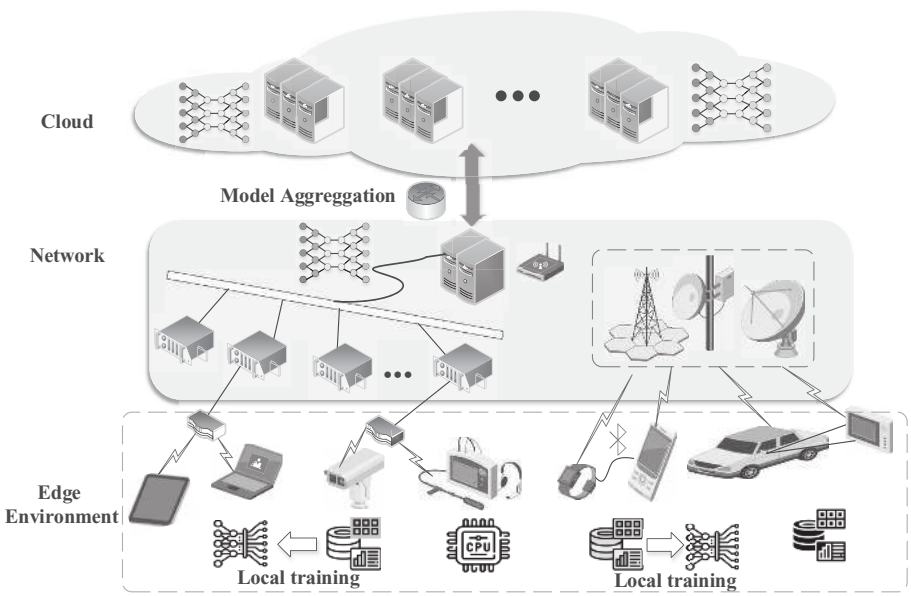


Figure 1.1 Illustration of the cloud - edge environment.

thus alleviating the risk of their privacy exposure. As such, federated learning can serve as an enabling technology for machine learning model training at mobile edge networks [170, 174, 221, 357].

1.3 Edge Learning and Edge Intelligence

The integration of edge computing and machine learning results in a new interdiscipline, named edge AI or edge intelligence, which is beginning to receive a tremendous amount of interest.

Edge learning is the enabling technology to achieve edge intelligence. It is a paradigm complementary to the cloud-based methods for big data analytics in the cloud-edge environment. It is proposed and developed for moving the training and inference to the edge environment to serve delay-sensitive and privacy-sensitive applications, of which the data cannot be gathered to the cloud. This fusion of big data, edge computing and machine learning is an enabling technology for edge intelligence.

In regard to training, edge learning exploits pervasive data generated not only by user devices but also by other sensing devices and that stored in the cloud/edge servers (e.g., data from social networks). It leverages various computing entities (all the devices with computing capabilities ranging from cloud and edge servers to various edge devices) in an efficient, reliable, and robust manner. In regard to inference, trained models are properly deployed and updated in edge networks by

taking into account both resource constraints and inference latency. It should also guarantee the privacy and security of training data and machine learning models as required. Compared with the traditional cloud-centric approaches to training machine learning models, edge learning has the following advantages.

- *Enabling various dispersed computing entities in the cloud-edge environment for learning collaboratively.* Edge learning enables the use of numerous edge servers and end devices in a collaborative manner, which is expected to increase the total computing power by multiple orders of magnitude.
- *Supporting the learning of multi-source data in a resource-efficient manner.* Edge learning supports communication-efficient learning of data not only generated by mobile users and edge sensors but also data that has been collected in the cloud, which can produce much more consistent, accurate, useful information and can greatly reduce the processing delay. In addition, there is no need to upload the raw data to the edge learning server for aggregation in the locally training process; instead, just upload their model updates, which significantly improves the communication efficiency.
- *Providing privacy and security as demanded.* Edge learning strikes a quite good balance between learning accuracy and data privacy by establishing privacy policies that the full exploitation of data while guaranteeing the security and privacy of different data as required.

1.4 Challenges of Edge Learning

While edge learning has great potential for many intelligent applications - e.g., smart cities and self-driving cars - , it is quite challenging to realize it in an efficient and secure manner due to the inherent characteristics of the cloud-edge environment. We summarize the main challenges in edge learning in Fig. 1.2. First, the training efficiency is hard to achieve not only because of the distinctive methods of data storage

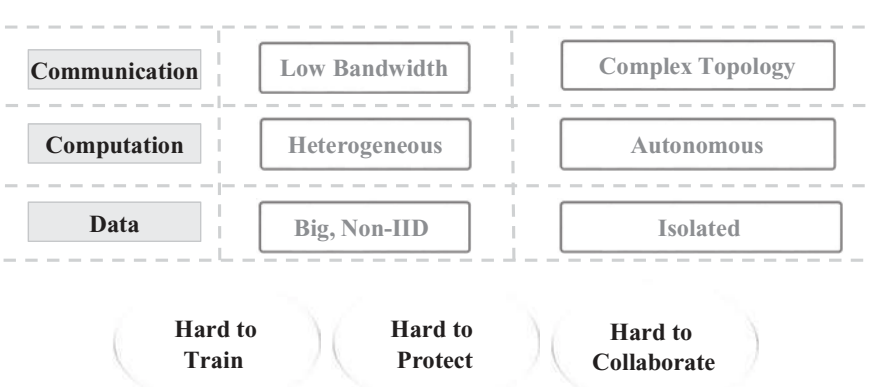


Figure 1.2 Summary of the main challenges in edge learning.

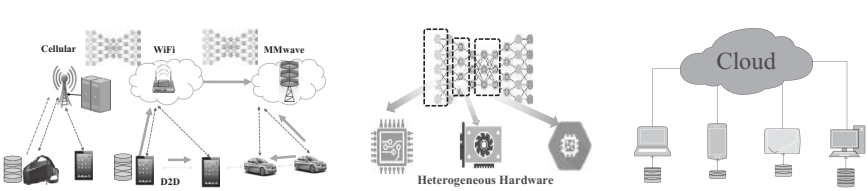
and communication of the edge facilities but also due to the greatly limited communication resource and the dynamicity of the edge environment. Second, learning on the edge needs to access data with various privacy protection requirements, posing a challenge in achieving learning accuracy while satisfying all the privacy requirements. Third, scalability is an important issue in edge learning, but it is difficult to manage a huge number of heterogeneous devices in the complex edge environment. Finally, edge devices utilize their resources and data to train a global model in edge learning. Without an appropriate incentive mechanism, edge devices may not participant in the training process.

1.4.1 Hard to Train Due to Constrained and Heterogeneous Edge Resources

The devices that participate in edge learning have heterogeneous characteristics, such as data distribution, computation abilities, and cooperation availability. There is also a much broader necessity to put forward effective methods to maximize the efficiency of the distributed learning process. In Fig. 1.3, we illustrate the challenges of edge learning from the perspective of communication, computation, and data.

First, communication is a critical challenge in edge learning. A typical learning system usually consists of a large number of mobile devices, such as mobile phones. In this scenario, the communication speed may be several times slower than local computation on most of the devices [117] [300]. In edge learning, the training process generally contains iterative transmission and aggregation of local updates - i.e., the gradient computed upon local samples -, while energy and communication resources are rather limited in edge devices. It is of great significance to achieve communication efficiency when a large number of edge devices regularly transmit local updates with millions or even billions of dimensional parameters. Thus, how to accelerate the learning process by incorporating various edge devices with resource and energy constraints is a challenging issue.

Second, edge devices are equipped with various processing units with different computing capabilities. How to make edge learning adapted to different hardware environments is challenging. Device heterogeneity significantly affects system performance due to variability in hardware (CPU, memory). For the given condition and resource constraints of each device, how to manage the computational and hardware resources to maximize the training process efficiency is crucial for us to explore.



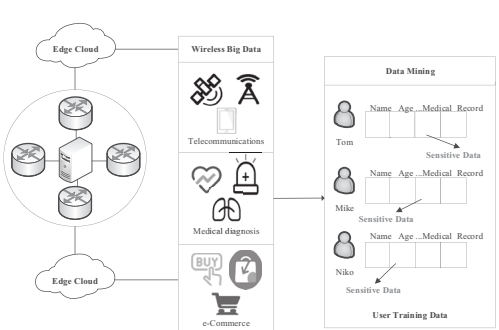
(a) Communication (b) Computation (c) Data

Figure 1.3 Challenges of edge learning from the perspective of communication, computation, and data.

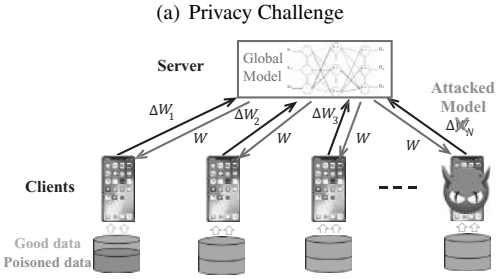
Third, traditional centralized learning frameworks require to upload all training data from different sources to a remote data server, which relies on an unrealistic assumption that their training data are independent and identically distributed (IID). This assumption is sometimes not reasonable in the real world; it is common for users' data to have a dependence relationship, so the data in each device may have different probability distributions. The data distributions generated in different devices are vary from one another, which will degrade the model accuracy under traditional distributed training algorithms. The problem of non-IID data is difficult because mobile devices usually hesitate to share their data, and it is hard to get information about data distribution.

1.4.2 Hard to Protect Due to Vulnerable Edge Devices

Security and privacy issues are drawing more and more attention. We illustrate the privacy and security challenges of edge learning in Fig. 1.4. In the cloud-edge environment, security and privacy are far more difficult to guarantee than in the cloud. Since the data are generated and stored in users' own devices, privacy and security are important issues that should be considered in edge learning. The training data may contain privacy-sensitive information, such as location, health records, and manufacturing information. In traditional machine learning, directly uploading the original datasets to the centralized server or exchanging the training data among edge devices can have a high risk of privacy leakage. Although many studies have proposed sharing model updates instead of the raw data [198], edge learning still faces sensitive information leakage when communicating model updates during the training process.



(a) Privacy Challenge



(b) Security Challenge

Figure 1.4 Challenge of edge learning from the perspective of privacy and security.

For instance, a third-party or centralized server may extract memory information of worker nodes from the uploaded model weights. How to guarantee the communication security throughout the training process is a significant problem to be resolved. Furthermore, the trustworthiness of the worker nodes needs to be considered to avoid malicious attacks.

1.4.3 Hard to Manage Due to Complex Edge Environment

Collaborative learning architecture enables learning beyond the cluster environment. Edge learning enables the use of all the cloud, edge servers, end devices in a collaborative manner, which is expected to increase the total computing power by multiple orders of magnitude.

In the scenario of big data analytics, operating large-scale machine learning applications often results in distributed processing and parallel computing, and handling the collaboration between edge nodes - especially in the heterogeneous environment - has become a promising research direction for both algorithm design and system implementation. We intend to elaborate an efficient distributed platform, which is compatible with the heterogeneous environment and fully exploits the capacity of edge devices when conducting machine learning applications. To achieve this goal, the major challenge is managing a huge number of heterogeneous devices in the complex edge environment to achieve a scalable edge learning system.

1.4.4 Hard to Collaborate Due to Lack of Participant

Another main challenge in edge learning is data islands, i.e., each client maintains its local data and has no incentive for contributing data to model training if no reward is granted. Thus, we must motivate a large number of clients to participate in edge learning to break the limitations inherent in isolated data islands. The power of the existing edge-based machine learning systems relies heavily on the quality of worker nodes' local model updates. Therefore, a fair incentive mechanism needs to be developed to achieve reliable participation.

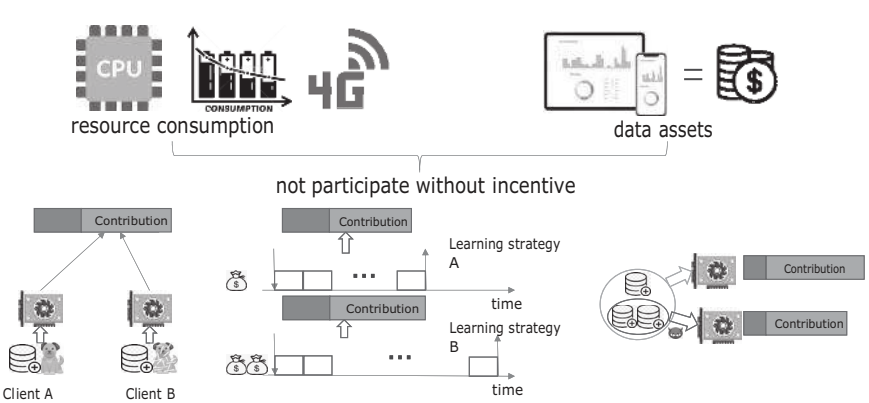


Figure 1.5 Challenges in edge device collaboration.

How to build a beneficial ecosystem for sustainable development of edge learning is a crucial issue. In Fig. 1.5, we illustrate the difficulty of collaboration due to lack of incentive. We face two main challenges: (1) from the worker nodes' perspective, how to recruit and retain more participants to improve the model efficiency, and (2) from the server-side perspective, how to evaluate each participant's contribution during the training process.

1.5 The Scope and Organization of This Book

In this section, we discuss the scope and organization of this book. An overview of organization is shown in Fig. 1.6.

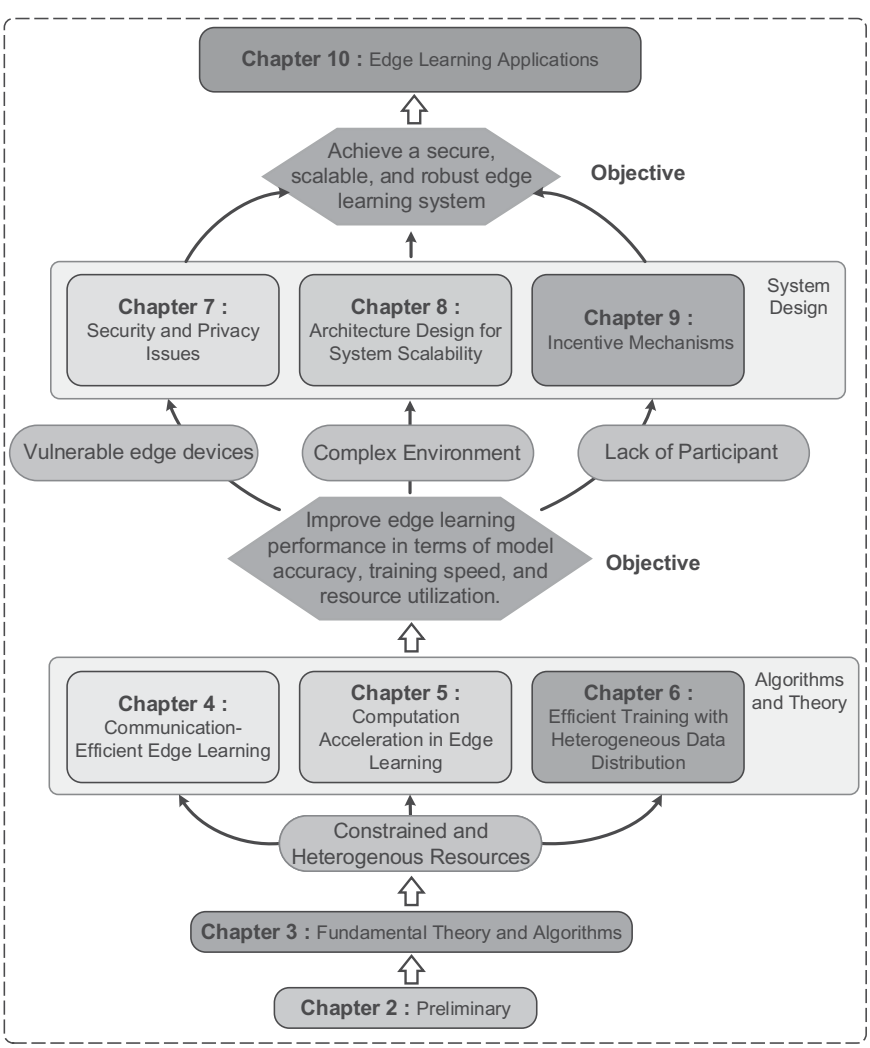


Figure 1.6 The scope and organization of this book.

Specifically, we first introduce the preliminary knowledge about edge learning, including the deep learning models, basic optimization algorithms, architectures, and synchronization modes (in Chapter 2). Then, to solve the basic optimization problem in edge learning, some fundamental theory and advanced training algorithms are explored in Chapter 3.

To deal with the constrained and heterogeneous resources in the edge environment, we introduce the edge learning technologies from the following aspects: communication-efficient technologies (Chapter 4), computation acceleration (Chapter 5), and heterogeneous data distribution (Chapter 6). Mainstream approaches are proposed to improve edge learning performance in terms of model accuracy, training speed, and resource utilization.

Apart from the investigation of algorithms and theory, we also present the details about the system design. Security guarantees and privacy protection mechanisms dealing with vulnerable edge devices are summarized in Chapter 7. In order to adapt to a complex environment, data parallelism, model parallelism, and hierarchical architecture are used in training procedure (in Chapter 8). Moreover, we explore the incentive mechanisms for edge learning to motivate the edge nodes to contribute to model training (in Chapter 9). After introducing how to design a secure, scalable, and robust edge learning system, we discuss the popular programming frameworks for edge learning and present the inspiration of how to implement learning into realistic scenarios (Chapter 10).