Toward an Intelligent Edge: Wireless Communication Meets Machine Learning

Guangxu Zhu, Dongzhu Liu, Yuqing Du, Changsheng You, Jun Zhang, and Kaibin Huang

Abstract

The recent revival of AI is revolutionizing almost every branch of science and technology. Given the ubiquitous smart mobile gadgets and IoT devices, it is expected that a majority of intelligent applications will be deployed at the edge of wireless networks. This trend has generated strong interest in realizing an "intelligent edge" to support AI-enabled applications at various edge devices. Accordingly, a new research area, called edge learning, has emerged, which crosses and revolutionizes two disciplines: wireless communication and machine learning. A major theme in edge learning is to overcome the limited computing power, as well as limited data, at each edge device. This is accomplished by leveraging the mobile edge computing platform and exploiting the massive data distributed over a large number of edge devices. In such systems, learning from distributed data and communicating between the edge server and devices are two critical and coupled aspects, and their fusion poses many new research challenges. This article advocates a new set of design guidelines for wireless communication in edge learning, collectively called learning-driven communication. Illustrative examples are provided to demonstrate the effectiveness of these design guidelines. Unique research opportunities are identified.

Introduction

We are witnessing phenomenal growth in global data traffic, accelerated by the increasing popularity of edge devices. According to the International Data Corporation, there will be 80 billion devices connected to the Internet by 2025, and the global data will reach 163 zettabytes, which is 10 times the data generated in 2016 [1]. The unprecedented amount of data, together with the recent breakthroughs in artificial intelligence (AI), inspire people to envision ubiquitous computing and ambient intelligence, which will not only improve our quality of life but also provide a platform for scientific discoveries and engineering innovations. In particular, this vision is driving industry and academia to vehemently invest in technologies for creating an *intelligent (network) edge*, which supports emerging application scenarios including smart city, eHealth, eBanking, intelligent transportation, and so on. This has led to the emergence of a new research area, called *edge learning*, which refers to the deployment of machine learning algorithms (including supervised, unsupervised, and reinforcement learning) at the network edge [2, 3]. The key motivation of pushing learning toward the edge is to allow rapid access to the enormous real-time data generated by the edge devices for fast AI-model training, which in turn endows the devices with human-like intelligence to respond to real-time events.

Traditionally, training an AI model, especially a deep neural network model, is computation-intensive and thus can only be supported at powerful cloud servers. Riding the recent trend in developing the *mobile edge computing* platform, training an AI model is no longer exclusive to cloud servers but also affordable at edge servers. In particular, the network virtualization architecture recently recommended by the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) is able to support edge learning on top of edge computing [4]. Moreover, the latest mobile devices are also armed with high-performance *central processing units* or *graphics processing units* (e.g., the A11 bionic chip in iPhone X), making them capable of training some small-scale AI models. The coexistence of cloud, edge, and on-device learning paradigms has led to a layered architecture for in-network machine learning, as shown in Fig. 1. Different layers possess different data processing and storage capabilities, and cater for different types of learning applications with distinct latency and bandwidth requirements.

Compared to cloud and on-device learning, edge learning has its unique strengths. First, it has the most balanced resource support (Fig. 1), which helps achieve the best trade-off between the supported AI model complexity and the model training latency. Second, given its proximity to data sources, edge learning overcomes the drawback of cloud learning that fails to process real-time data due to excessive propagation delay caused by data uploading. Furthermore, the proximity gives an additional advantage of location and context awareness. Last, compared to on-device learning, edge learning achieves much higher learning accuracy by supporting more complex models and more importantly,

This article advocates a new set of design guidelines for wireless communication in edge learning, collectively called learning-driven communication. Illustrative examples are provided to demonstrate the effectiveness of these design guidelines. Unique research opportunities are identified.

Guangxu Zhu and Changsheng You were formerly affiliated with the University of Hong Kong; Guangxu Zhu is now with the Shenzhen Research Institute of Big Data; Changsheng You is now with the National University of Singapore; Dongzhu Liu, Yuqing Du, and Kaibin Huang (corresponding author) are with the University of Hong Kong; Jun Zhang is with Hong Kong Polytechnic University.

Digital Object Identifier: 10.1109/MCOM.001.1900103

The main design objective in edge learning is the fast intelligence acquisition from the rich but highly distributed data at subscribed edge devices. This critically depends on data processing at edge servers, as well as efficient communication between edge servers and edge devices.

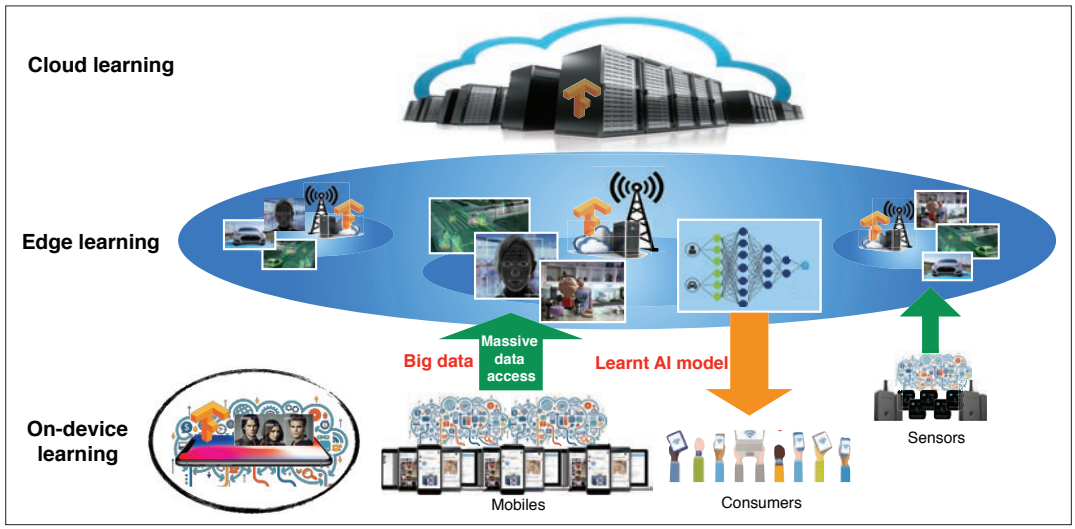

Figure 1. Layered in-network machine learning architecture.

aggregating distributed data from many devices. Due to the well-rounded capabilities, edge learning can support a wide spectrum of AI models to power a broad range of mobile applications, such as auto-driving and augmented/virtual reality (AR/VR). All these applications are real-time and thus require fast online learning. For example, in auto-driving, the AI model needs to continuously adapt to the ambient environment to make split-second decisions whether to apply the brake [5]; in AR/VR, the visuo-haptic perception cannot afford even 20 ms delay; otherwise, the user may suffer from motion sickness [6]. Nevertheless, edge learning is at its nascent stage and thus remains a largely uncharted area with many open challenges.

The main design objective in edge learning is *fast intelligence acquisition* from the rich but highly distributed data at subscribed edge devices. This critically depends on data processing at edge servers, as well as efficient communication between edge servers and edge devices. Compared to increasingly high processing speeds at edge servers, communication suffers from hostility of wireless channels (e.g., path loss, shadowing, and fading), and consequently forms the bottleneck for ultra-fast edge learning. In order to distill the shared intelligence from distributed data, excessive communication latency may arise from the need to upload to an edge server a vast amount of data generated by *millions to billions* of edge devices, as illustrated in Fig. 1. As a concrete example, Tesla's AI model for auto-driving is continuously improved using RADAR and LIDAR sensing data uploaded by millions of Tesla vehicles on the road, which can amount to about 4000 GB for one car per day. Given the enormity in data and the scarcity of radio resources, how to fully exploit the distributed data in AI model training without incurring excessive communication latency poses a grand challenge for wireless data acquisition (WDA) in edge learning.

Unfortunately, the state-of-the-art wireless technologies are incapable of tackling the challenge. The fundamental reason is that the traditional design objectives of wireless communications, namely communication reliability and data rate maximization, do not directly match that of edge learning. This means that we have to break away from the conventional philosophy in traditional wireless communication, which can be regarded as a "communication-computing separation" approach. Instead, we should exploit the coupling between communication and learning in edge learning systems. To materialize the new philosophy, we propose in this article a set of new design guidelines for wireless communication in edge learning, collectively called *learning-driven communication*. In the following sections, we shall discuss specific research directions and provide concrete examples to illustrate this paradigm shift, which cover key communication aspects including multiple access, resource allocation, and signal encoding, as summarized in Table 1. All of these new design guidelines share a common theme as highlighted below.

Guideline of learning-driven communication – fast intelligence acquisition:

Efficiently transmit learning-relevant information (data) to speed up and improve AI model training at edge servers.

Learning-Driven Multiple Access

Motivation and Principle

In edge learning, the involved training data are often privacy-sensitive and large in quantity. Thus, uploading them from devices to an edge server for centralized model training may not only raise a privacy concern but also incur prohibitive cost in communication. This motivates an innovative edge learning framework, called *federated learning*, which features *distributed learning* at edge devices and *model-update aggregation* at an edge server [7]. Federated learning can effectively address the aforementioned issues as only the locally computed model updates, instead of raw data, are uploaded to the server. A typical federated learning algorithm alternates between two phases, as shown in Fig. 2. One is to aggregate distributed model updates over a multi-access channel and apply their average to update the AI-model at the edge server. The other is to broadcast the model under training to allow edge devices to continuously refine their local models.

Commun. technology	Item	Conventional communication	Learning-driven communication
Multiple access	Target	Decoupling messages from users	Computing function of distributed data
	Case study	OFDMA	Model-update averaging by AirComp
Resource allocation	Target	Maximize sum-rate or reliability	Fast intelligence acquisition
	Case study	Reliability-based retransmission	Importance-aware retransmission
Signal encoding	Target	Optimal trade-offs between rate and distortion/reliability	Latency minimization while preserving the learning accuracy
	Case study	Quantization, adaptive modulation, and polar code	Grassmann analog encoding

Table 1. Conventional communication vs. learning-driven communication.

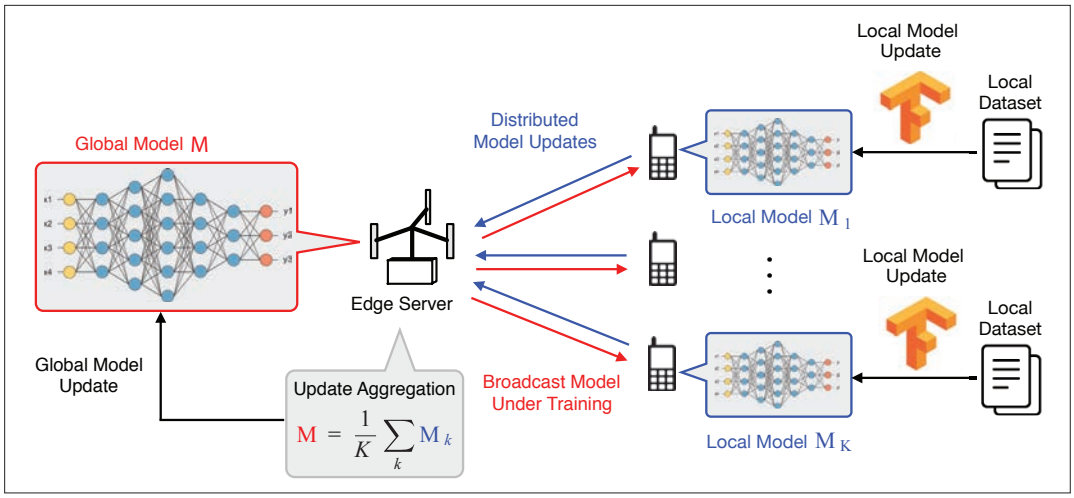

Figure 2. Federated learning using wirelessly distributed data.

Model update uploading in federated learning is bandwidth-consuming as an AI model usually comprises millions to billions of parameters. Overall, the model updates by thousands of edge devices may easily congest the air interface, making it a bottleneck for agile edge learning. The said bottleneck is arguably an artifact of the classic approach of *communication-computing separation*. Existing multiple access technologies such as orthogonal frequency-division multiple access (OFDMA) and code-division multiple access (CDMA) are purely for rate-driven communication and fail to adapt to the actual learning task. The need for enabling fast edge learning from massive distributed data calls for a new design guideline for multiple access. In this section, we present *learning-driven multiple access* as the solution, and showcase a particular technique under this new guideline.

The key innovation underpinning learning-driven multiple access is to exploit the insight that the learning task involves computing some aggregating function of multiple data samples, rather than decoding individual samples as in existing schemes. For example, in federated learning, the edge server requires the average of model updates rather than their individual values. On the other hand, the multi-access wireless channel by itself is a natural data aggregator: the simultaneously transmitted analog-waves by different devices are automatically superposed at the receiver but weighed by the channel coefficients. The above insights motivate the following design guideline that changes the traditional philosophy of "overcoming interference" to the new one of "harnessing interference."

Guideline of learning-driven multiple access:

Unique properties of wireless channels, e.g., broadcast and superposition, should be exploited for function computation over distributed data to accelerate edge learning.

Following the new guideline, the wave superposition nature of the multi-access channel suggests that by using *linear-analog modulation* and *pre-channel compensation* at the transmitter, the "interference" caused by concurrent data transmission can be exploited for fast data aggregation. This intuition has been captured by a recently proposed technique called *over-the-air computation* (AirComp) [8–10]. By allowing simultaneous transmission, AirComp can dramatically reduce the multiple access latency by a factor equal to the number of users, overcoming the communication latency bottleneck in edge learning.

Case Study: AirComp for Federated Learning

Experiment Settings: Consider a federated learning system with one edge server and $K = 100$ edge devices. For exposition, we consider the learning task of handwritten-digit recognition using the well-known MNIST dataset that consists of 10 categories ranging from digit "0" to "9" and a total of 60,000 labeled training data samples. To simulate the distributed mobile data, we random-

Following the new guideline, the wave superposition nature of the multi-access channel suggests that by using linear-analog modulation and pre-channel-compensation at the transmitter, the "interference" caused by concurrent data transmission can be exploited for fast data aggregation. This intuition has been captured by a recently proposed technique called over-the-air computation (AirComp).

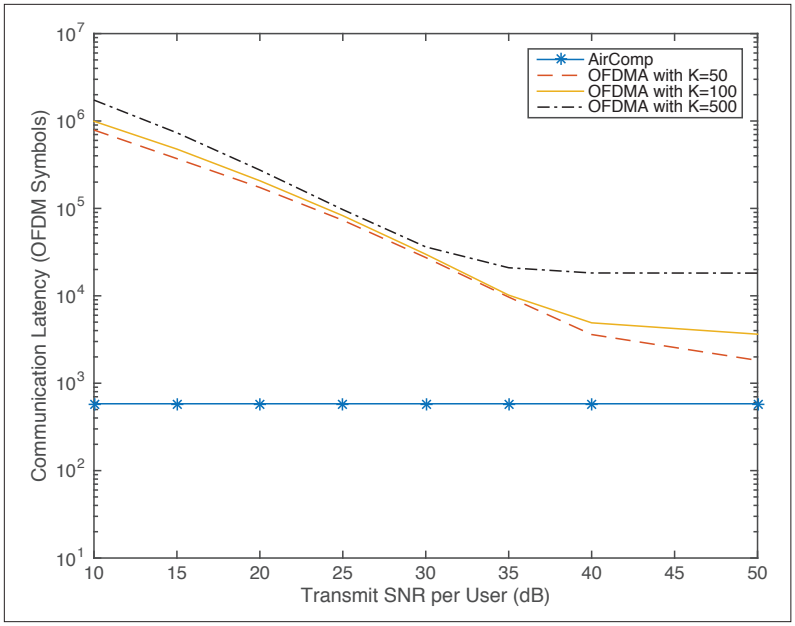

Figure 3. Performance comparison between AirComp and OFDMA in communication latency. For AirComp, model parameters are analog-modulated, and each sub-channel is dedicated for single-parameter transmission; truncated-channel inversion under the transmit power constraint is used to tackle the channel fading. For OFDMA, model parameters are first quantized into a bit sequence (16 bits per parameter). Then adaptive *M*-QAM is adopted to adapt the data rate to the channel condition such that the spectrum efficiency is maximized while the target bit error rate of 10^{-3} is maintained.

ly partition the training samples into 100 equal shares, each of which is assigned to one particular device. The classifier model is implemented using a 4-layer *convolutional neural network* with two 5×5 convolution layers, a fully connected layer with 512 units and ReLu activation, and a final softmax output layer.

AirComp vs. OFDMA: During the federated model training, in each *communication round*, local models trained at edge devices are transmitted and aggregated at the edge server over a shared broadband channel consisting of 1000 orthogonal sub-channels. We compare the proposed AirComp with the conventional OFDMA. They mainly differ in how the available sub-channels are shared. For OFDMA, the 1000 sub-channels are evenly allocated to the K edge devices, so each device uploads its local model using only fractional bandwidth that reduces as K grows. Model averaging is performed by the edge server after all local models are reliably received, and thus the communication latency is determined by the slowest device. In contrast, the AirComp scheme allows every device to use the full bandwidth so as to exploit the "interference" for direct model averaging over the air. The latency of AirComp is thus independent of the number of accessing devices.

Performance: Although AirComp is expected to be more vulnerable to channel noise, it is interesting to note that the two schemes are comparable in learning accuracy (AirComp 98.18 percent vs. OFDMA 98.45 percent). Such accurate learning of AirComp is partly due to the high expressiveness of the deep neural network, which makes the learned model robust against perturbation by channel noise. The result has a profound and refreshing implication that *reliable communication may not be the primary concern in edge learning, especially when deep neural network models are adopted*. Essentially, AirComp exploits this relaxation on communication reliability to trade for a low communication latency, as shown in Fig. 3. Without compromising the learning accuracy, AirComp achieves a significant latency reduction ranging from 10× to 1000×. The superiority in latency of AirComp over OFDMA is more pronounced in low signal-to-noise (SNR) regime and dense network scenarios.

Research Opportunities

Robust Learning with Imperfect AirComp: Inaccurate channel estimation and non-ideal hardware at the low-cost edge devices may cause imperfect channel equalization and thus distort the aggregated data by AirComp. For practical implementation, it is important to characterize the effect of imperfect AirComp on the performance of edge learning, based on which new techniques can be designed to improve the learning robustness.

Asynchronous AirComp: Successful implementation of AirComp requires strict synchronization among all the participating edge devices. This may be hard to achieve when the devices exhibit high mobility. To enable ultra-fast data aggregation in high-mobility scenarios, new schemes operated in an asynchronous manner or with a relaxed requirement on synchronization are desirable.

Generalization to Other Edge-Learning Architectures: Apart from federated learning, it is also interesting to generalize the learning-driven multiple access to other architectures, where the edge server needs to perform more sophisticated computation over received data than simple averaging. How to exploit the superposition property of a multi-access channel to compute more complex functions is the main challenge in the generalization [9].

Learning-Driven Radio Resource Management

Motivation and Principle

With the traditional communication-computing separation approach, existing methods of radio resource management (RRM) are designed to maximize the spectral efficiency by carefully allocating the scarce radio resources such as power, frequency band, and access time. However, such an approach is no longer effective in edge learning, as it fails to exploit the subsequent learning process for further performance improvement. This motivates the following new design guideline for RRM in edge learning.

Guideline of learning-driven RRM:

Radio resources should be allocated based on the value of transmitted data for learning performance optimization.

Conventional RRM assumes that different messages have the same value for the receiver. The assumption makes sum-rate maximization a key

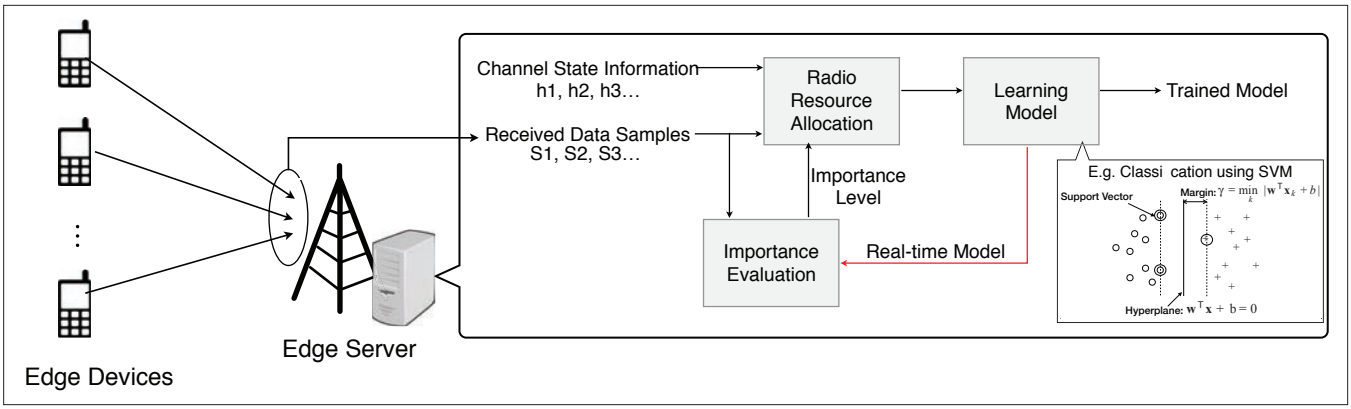

Figure 4. A communication system with learning-driven RRM.

design criterion. When it comes to edge learning, the rate-driven approach is no longer efficient as some messages tend to be more valuable than others for training an AI model.

In this part, we introduce a representative technique following the above learning-driven design guideline, called *importance-aware resource allocation*, which takes the data importance into account in resource allocation. The basic idea of this new technique shares some similarity with a key area in machine learning called *active learning*. Principally, active learning is to select important samples from a large unlabeled dataset for labeling so as to accelerate model training with a labeling budget [11]. A widely adopted measure of importance is uncertainty. Specifically, a data sample is more uncertain if it is less confidently predicted by the current model. A commonly used uncertainty measure is entropy, a notion from information theory. As its evaluation is complex, a heuristic but simple alternative is the distance of a data sample from the decision boundaries of the current model. Taking the support vector machine (SVM) as an example, a training data sample close to the decision boundary is likely to be a support vector, thereby contributing to defining the classifier. In contrast, a sample away from boundaries makes no such contribution.

Compared to active learning, learning-driven RRM has additional challenges given the volatile wireless channels. In particular, besides data importance, it needs to consider radio resource allocation to ensure a certain level of reliability in transmitting a data sample. A basic diagram of learning-driven RRM is illustrated in Fig. 4.

Case Study: Importance-Aware Retransmission for WDA

Experiment Settings: Consider a centralized edge learning system where a classifier is trained at the edge server based on SVM, with data collected from distributed edge devices. The acquisition of high-dimensional training-data samples is bandwidth consuming and relies on a noisy data channel. On the other hand, a low-rate reliable channel is allocated for accurately transmitting small-size labels. The mismatch between the labels and noisy data samples at the edge server may lead to an incorrectly learned model. To tackle the issue, importance-aware retransmission with coherent combining is proposed in [12] to enhance the data quality. The radio resource is specified by a transmission budget of $N = 4000$ samples (new/retransmitted), which constrains the communication latency. T o train the classifier, we again use the MNIST dataset and choose the relatively less differentiable class pair of "3" and "5" to focus on a binary classification case.

Importance-Aware Retransmission: Under a transmission budget constraint, the RRM problem can be specified as: How many retransmission instances should be allocated for a given data sample? Concretely, in each communication round, the edge server should make a binary decision on either selecting a device for acquiring a new sample or requesting the previously scheduled device for retransmission to improve sample quality. Given a finite transmission budget, the decision making needs to address the *trade-off between the quality and quantity* of received samples. In SVM, data located close to the decision boundary are more critical to the model training but also easier to commit the data-label mismatch issue. Therefore, they require a higher retransmission budget to ensure a pre-specified alignment probability (defined as the possibility that the transmitted and received data lie at the same side of the decision boundary). This motivates the importance-aware retransmission scheme where the retransmission decision making is controlled by applying an adaptive SNR threshold. The adaptation is realized by weighting the threshold with a coefficient that is equal to the distance between the sample and the decision boundary. This enables intelligent allocation of the transmission budget according to data importance such that the optimal quality-quantity trade-off can be achieved.

Performance: Figure 5 presents the learning performance of the importance-aware retransmission along with two benchmark schemes, namely, conventional channel-aware retransmission with a fixed SNR threshold and the scheme without retransmission. It is observed that if there is no retransmission, the learning performance dramatically degrades after acquiring a sufficiently large number of noisy samples. This is because the strong noise effect accumulates to cause the divergence of the model, which justifies the need for retransmission. Next, one can observe that importance-aware retransmission outperforms conventional channel-aware retransmission throughout the entire training duration. The effect

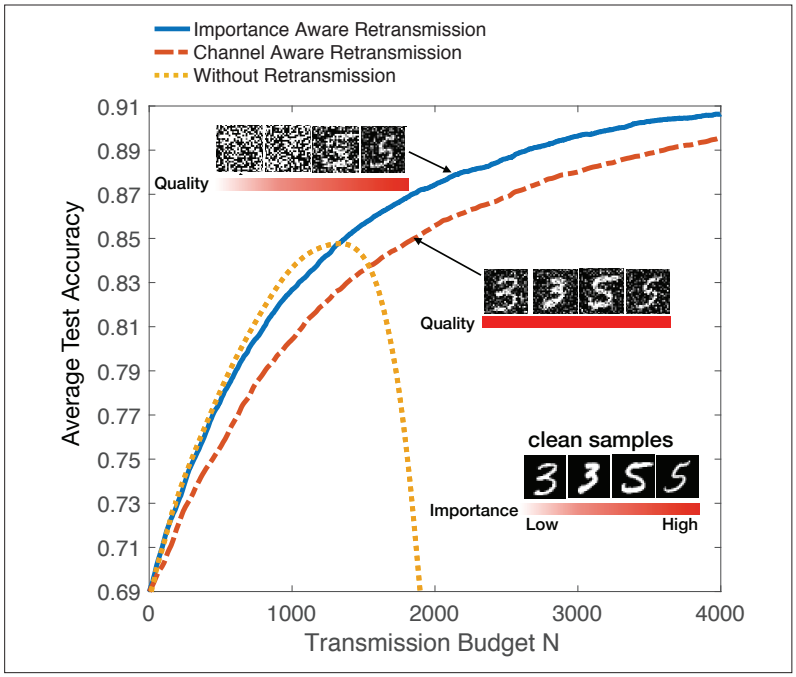

Figure 5. Classification performance for importance-aware retransmission and two baselines. The MRC combining technique is applied to coherently combine all the retransmission observations for maximizing the receive SNR. The retransmission stops when the receive SNR meets a predefined threshold. The average receive SNR is 10 dB.

of importance-aware resource allocation can be further visualized by the selected four training samples as shown in the figure: the quality varies with data importance in the proposed scheme, while the conventional channel-aware scheme strives to keep high quality for each data sample.

Research Opportunities

Cache-Assisted Importance-Aware RRM: With sufficient storage space, edge devices may pre-select important data from the locally cached data before uploading, which can result in faster convergence of the AI-model training. However, the conventional importance evaluation based on data uncertainty may lead to undesired selection of outliers. How to incorporate data representativeness into the data importance evaluation by exploiting the local data distribution is the key issue to be addressed.

Multi-User RRM for Faster Intelligence Acquisition: Multiple access technologies allow simultaneous data uploading from multiple users. The resultant batch data acquisition can enhance overall efficiency, as it reduces the frequency of updating an AI model under training. However, due to the correlation of data across devices, accelerating model training may come at a cost of unnecessarily processing redundant information. Therefore, how to efficiently exploit data diversity in the presence of inter-user correlation is an interesting topic to study.

Learning-Driven RRM in Diversified Scenarios: In the case study presented above, importance-aware RRM assumes the need of uploading raw data. However, in a more general edge learning system, what is uploaded from edge devices to the edge server is not necessarily the raw data but can be other learning-related contents (e.g., model updates in federated learning). This makes the presented data-importance-aware RRM design not directly applicable and calls for new design.

Learning-Driven Signal Encoding

Motivation and Principle

In machine learning, feature extraction techniques are widely applied in pre-processing raw data so as to reduce its dimensions as well as improve the learning performance. There are numerous feature extraction techniques. For example, *principal component analysis* is a popular technique for identifying a latent feature space and using it to reduce data samples to their low-dimensional features; *linear discriminant analysis* finds the most discriminant feature space to facilitate data classification. A common theme shared by feature extraction techniques is to reduce a training dataset into low-dimensional features that simplify learning and improve its performance. Too aggressive or too conservative dimensionality reduction may degrade the learning performance. Furthermore, the choice of a feature space directly affects the performance of a targeted learning task. These make designing feature extraction techniques a challenging but important topic in machine learning.

In wireless communication, techniques of source-and-channel encoding are developed to "preprocess" transmitted data, but for a different purpose, namely efficient and reliable delivery. Source coding samples, quantizes, and compresses the source signal such that it can be represented by a minimum number of bits under a constraint on signal distortion. This gives rise to a *rate-distortion trade-off*. On the other hand, for reliable transmission, channel coding introduces redundancy into a transmitted signal for protecting it against noise and hostility of wireless channels. This results in the *rate-reliability trade-off*. Designing joint source-and-channel coding essentially involves the joint optimization of the two mentioned trade-offs.

Since both are data preprocessing operations, it is natural to integrate feature extraction and source-and-channel encoding so that the inherent data geometric structure can be exploited for communication-efficient coding design in edge-learning systems. This gives rise to a new topic of *learning-driven signal encoding* with the following design guideline.

Guideline of learning-driven signal encoding:

Signal encoding at an edge device should be designed by jointly optimizing feature extraction, source coding, and channel encoding so as to accelerate edge learning.

Case Study: Grassmann Analog Encoding

In this subsection, an example technique following the above guideline, called Grassmann analog encoding (GAE), is introduced. GAE represents a raw data sample in the Euclidean space by a subspace, which can be interpreted as a feature, via projecting the sample onto a Grassmann manifold. The operation reduces the data dimensionality but also distorts the data sample by causing *degree-of-freedom* loss. In return, the direct transmission of GAE

encoded data samples using linear-analog modulation not only supports blind multiple-input multiple-output (MIMO) transmission without *channel state information* but also provides robustness against fast fading. The feasibility of blind transmission is attributed to the same principle as the classic non-coherent MIMO transmission [13]. On the other hand, the GAE encoded dataset retains its original cluster structure and thus its usefulness for training a classifier at the edge server.

The effectiveness of GAE has been demonstrated via a case study in prior work [13], where the learning performance of a centralized edge learning system using GAE for fast analog (data) transmission has been benchmarked against that using two high-rate coherent schemes: digital and analog MIMO transmission. Significant performance gain has been observed in high-mobility scenarios.

Research Opportunities

Gradient-Data Encoding: In federated learning, the computed gradients may have a high dimensionality, so its transmission could be communication-inefficient. Fortunately, it is found that by exploiting the inherent sparsity structure, the gradient update can be truncated appropriately without significantly degrading the training performance [14]. This inspires the design of gradient compression techniques to reduce communication latency.

Channel-Aware Feature Extraction: Traditional channel-aware signal processing can also be jointly designed with the feature extraction in edge learning systems. Particularly, a recent study [15] has shown the inherent analogy between the feature extraction process for classification and non-coherent communication. This suggests the possibility of exploiting the channel characteristics for efficient feature extraction, giving rise to a new research area called channel-aware feature extraction.

Concluding Remarks

Edge learning, sitting at the intersection of wireless communication and machine leaning, enables promising AI-powered applications and brings new research opportunities. The main aim of this article is to introduce a set of new design guidelines to the wireless communication community for the upcoming era of edge intelligence. The introduced learning-driven communication techniques, including multiple access, resource allocation, and signal encoding, can break the communication latency bottleneck and lead to fast edge learning.

Acknowledgment

The work was supported in part by Hong Kong Research Grants Council under the Grants 17208319, 17209917 and 17259416, and Shenzhen Peacock Plan under Grant KQTD2015033114415450. Dr. J. Zhang was supported by a start-up fund of Hong Kong Polytechnic University (Project ID P0013883)

References

[1] N. Poggi, "3 Key Internet of Things Trends to Keep Your Eye On in 2017"; https://preyproject.com/blog/en/3-keyinternet-of-things-trends-to-keep-your-eye-on-in-2017/, 2017.
[2] H. Li, K. Ota, and M. Dong, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," *IEEE Network*, vol. 32, no. 1, Jan. 2018, pp. 96–101.
[3] S. Wang *et al.*, "When Edge Meets Learning: Adaptive Control for Resource Constrained Distributed Machine Learning," *Proc. INFOCOM*, Honolulu, HI, 2018.
[4] ITU-T, "Unified Architecture for Machine Learning in 5G and Future Networks"; https://www.itu.int/en/ITUT/ focusgroups/ml5g/Documents/ML5G-delievrables.pdf, Feb. 2019.
[5] S.-C. Lin *et al.*, "The Architectural Implications of Autonomous Driving: Constraints and Acceleration," *Proc. ASPLOS*, Williamsburg, VA, 2018.
[6] M. S. Elbamby *et al.*, "Toward Low-Latency and Ultra-Reliable Virtual Reality," *IEEE Network*, vol. 32, no. 2, Apr. 2018, pp. 78–84.
[7] H. B. McMahan *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proc. AISTATS,* Fort Lauderdale, FL, Apr. 2017.
[8] G. Zhu and K. Huang, "MIMO Over-the-Air Computation for High-Mobility Multi-Modal Sensing," *IEEE IoT J.*, vol. 6, no. 4, Aug. 2018, pp. 6089–6103.
[9] G. Zhu, Y. Wang, and K. Huang, "Broadband Analog Aggregation For Low-Latency Federated Edge Learning," *IEEE Trans. Wireless Commun.* (Early Access), Oct. 2019.
[10] M. M. Amiri and D. Gunduz, "Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air"; https://arxiv.org/abs/1901.00844, accessed Jan. 2019.
[11] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, 2012, pp. 1–114.
[12] D. Liu *et al.*, "Wireless Data Acquisition for Edge Learning: Importance Aware Retransmission"; http://arxiv.org/abs/1812.02030, accessed Mar. 2019.
[13] Y. Du and K. Huang, "Fast Analog Transmission for High-Mobility Wireless Data Acquisition in Edge Learning," *IEEE Wireless Commun. Letters*, vol. 8, no. 2, Apr. 2018.
[14] Y. Lin *et al.*, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," *Proc. ICLR*, Vancouver, Canada, May 2018.
[15] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the Grassmann Manifold: Fundamental Limits of Subspace Classifiers," *IEEE Trans. Info. Theory*, vol. 61, no. 4, Apr. 2015, pp. 2133–47.

Biographies

Guangxu Zhu [S'14] received his Ph.D. degree from the University of Hong Kong. His research interests include edge intelligence, distributed learning, and 5G systems.

Dongzhu Liu [S'17] received her PhD degree from the University of Hong Kong. Her research interests include edge learning and content delivery networks.

Yuqing Du [S'18] is pursuing a Ph.D. degree at the University of Hong Kong. His research interests include edge intelligence and distributed learning.

Changsheng You [S'15] received his Ph.D. degree from the University of Hong Kong. His research interests include edge computing/learning, and UAV communications.

Jun Zhang [M'10, SM'15] is an assistant professor in the Department of EIE at Hong Kong Polytechnic University. His research interests include 5G systems, edge computing/learning, and big data analytics.

Kaibin Huang [M'08, SM'13] is an associate professor with the Department of EEE, University of Hong Kong. His research interests include mobile edge computing, distributed learning, and 5G systems.

A recent study has shown the inherent analogy between the feature extraction process for classification and the non-coherent communication. This suggests the possibility to exploit the channel characteristics for efficient feature-extraction, giving rise to a new research area called channel-aware feature-extraction.