Deep learning-driven wireless communication for edge-cloud computing: opportunities and challenges

Huaming Wu*, Xiangyi Li and Yingjun Deng

Abstract

Future wireless communications are becoming increasingly complex with different radio access technologies, transmission backhauls, and network slices, and they play an important role in the emerging edge computing paradigm, which aims to reduce the wireless transmission latency between end-users and edge clouds. Deep learning techniques, which have already demonstrated overwhelming advantages in a wide range of internet of things (IoT) applications, show significant promise for solving such complicated real-world scenarios. Although the convergence of radio access networks and deep learning is still in the preliminary exploration stage, it has already attracted tremendous concern from both academia and industry. To address emerging theoretical and practical issues, ranging from basic concepts to research directions in future wireless networking applications and architectures, this paper mainly reviews the latest research progress and major technological deployment of deep learning in the development of wireless communications. We highlight the intuitions and key technologies of deep learning-driven wireless communication from the aspects of end-to-end communication, signal detection, channel estimation and compression sensing, encoding and decoding, and security and privacy. Main challenges, potential opportunities and future trends in incorporating deep learning schemes in wireless communications environments are further illustrated.

Keywords: Wireless communication, Future network, Security, Edge-cloud computing, Internet of things

Introduction

Along with the incredible growth of mobile data generated in internet of things (IoT) and the explosion of complicated wireless applications, e.g., virtual reality (VR) and augmented reality (AR), the fifth-generation (5G) technology demonstrates high-dimensional, high-capacity and high-density characteristics [1, 2]. Moreover, future wireless communication systems will become ever-more demanding for edge-cloud computing since the edge servers are in proximity of the IoT devices and communicate with them via different wireless communication technologies [3, 4]. The requirements of high bandwidth and low latency for wireless communications have posed enormous challenges to the design, configuration, and optimization of next-generation networks (NGN) [5, 6]. In the meantime, massive multiple-input multiple-output (MIMO) is widely regarded as a major technology for future wireless communication systems. In order to improve the quality of wireless signal transmission, the system uses multiple antennas as multiple transmitters at the base station (BS) and receivers at a user equipment (UE) to realize the multipath transmitting, which can double the channel capacity without increasing spectrum resources or antenna transmit power. However, conventional communication systems and theories exhibit inherent limitations in the utilization of system structure information and the processing of big data. Therefore, it is urgent to establish new communication models, develop more effective solutions to address such complicated scenarios and

* Correspondence: whming@tju.edu.cn
Center of Applied Mathematics, Tianjin University, Weijin Road, Tianjin, China

further fulfill the requirements of future wireless communication systems, e.g., beyond the fifth-generation (B5G) networks.

Along with the fast convergence of communication and computing in popular paradigms of edge computing and cloud computing [7, 8], intelligent communication is considered to be one of the mainstream directions for the extensive development of future 5G and beyond wireless networks, since it can optimize wireless communication systems performance. In addition, with tremendous progress in artificial intelligence (AI) technology, it offers alternative options for addressing these challenges and replacing the design concepts of conventional wireless communications. Deep learning (DL) is playing an increasingly crucial role in the field of wireless communications due to its high efficiency in dealing with tremendous complex calculations, and is regarded as one of the effective tools for dealing with communication issues. Although deep learning has performed well in some IoT applications, "no free lunch" theorem [9] shows that a model cannot solve all problems once and for all, and we cannot learn a general model for a wide range of communication scenarios. This means that for any particular mobile and wireless network issue, we still need to adopt different deep learning architectures such as convolutional neural networks (CNN), deep neural networks (DNN) and recurrent neural networks (RNN), in order to achieve better performance of the communication systems.

As a classic model of deep learning, autoencoder is widely used in the design paradigms of communication system models. Autoencoder-based wireless communication models are drawing more and more attention [10–12]. Generative adversarial network (GAN) [13] is a promising technique, which has attracted great attention in the field of mobile and wireless networking. The architecture of GAN is composed of two networks, i.e., a discriminative model and a generative model, in which a discriminator D is trained to distinguish the real and fake samples, while the generator G is trained to fool the discriminator D with generated samples. The feature of GAN is very appropriate for training. GAN-driven models and algorithms can facilitate the development of next-generation wireless networks, especially coping with the growth in volumes of communication and computation for emerging IoT applications. However, the incorporation of AI technology in the field of wireless communications is still in its early stages, and learning-driven algorithms in mobile wireless systems are immature and inefficient. More endeavors are required to bridge the gap between deep learning and wireless communication research, e.g., customize GAN techniques for network analytics and diagnosis and wireless resource management in heterogeneous mobile environments [14].

This survey explores the crossovers and the integration of wireless communication and AI technology, aims at solving specific issues in the mobile networking domain, and greatly improve the performance of wireless communication systems. We gather, investigate and analyze latest research works in emerging deep learning methods for processing and transferring data in the field of wireless communications or related scenarios, including strengths and weaknesses. The main focus is on how to customize deep learning for mobile network applications from three perspectives: mobile data generation, end-to-end wireless communications and network traffic control that adapts to dynamic mobile network environments. Several potential deep learning-driven underlying communication technologies are described, which will promote the further development of future wireless communications.

The rest of this paper is organized as follows: we first draw an overall picture of the latest literature on deep learning technologies in the field of wireless communications. Then, we present important open issues and main challenges faced by researchers for intelligent communications. After that, several potential techniques and research topics in deep learning-driven wireless communications are pointed out. Finally, the paper is concluded.

Emerging deep learning Technologies in Wireless Communications

A list of emerging technology initiatives of incorporating AI schemes for communication research is provided by IEEE Communications Society.[1] This section selects and introduces the latest research progress of deep learning-driven wireless communication from the aspects of end-to-end communication, signal detection, channel estimation, channel estimation and compression sensing, encoding and decoding, and security and privacy.

End-to-end communications

The guiding principle in communication system design is to decompose signal processing into chains with multiple independent blocks. Each independent block performs a well-defined and isolated function, such as source coding/decoding, channel coding/decoding, modulation, channel estimation and equalization [15]. This kind of approach yields today's efficient, versatile, and controllable wireless communication systems. However, it is unclear whether the optimization of individual processing blocks can achieve optimal end-to-end performance, while deep learning can realize theoretically

[1]Machine Learning For Communications Emerging Technologies Initiative https://mlc.committees.comsoc. org/research-library.

global optimal performance. Thus, deep learning has produced far-reaching significance for wireless communication systems and has shown promising performance improvements.

As shown in Fig. 1, an autoencoder consists of an encoder and a decoder, where the input data is first processed by the encoder at the transmitter, and then it is decoded at the receiver in order to get the output. The transmitter encodes the input s as a one-hot vector, and a conditional probability density function $p(y|x)$ is applied to indicate the wireless channel. After receiving the message, the receiver selects the one with the maximum probability over all possible messages as the output $\hat{s}$ [10]. Autoencoder is mainly constructed by neural networks, i.e., an encoding network and a decoding network, the wireless communication system is divided into multiple physical layers to facilitate the propagation of information via the neural network thereon.

In addition, the idea of end-to-end learning in communication systems has also attracted widespread attention in the wireless communications community [16]. Several emerging trends for deep learning in communication physical layer were elaborated in [10]. By treating the wireless communication system as an autoencoder, redefining it as the transmitter and receiver, a local optimum of the end-to-end refactoring process can be achieved. Moreover, different conditions were set in the physical layer to simulate different transmission environments in reality.

The design paradigms of conventional wireless communication systems have to consider the influence of various uncertain factors in hardware implementation, and compensate for delay and phase, which is not efficient and scalable. In contrast, model-free training of end-to-end communication systems based on autoencoder was built by hardware implementations on software-defined radios (SDRs) [17, 18], which was simpler, faster, and more efficient. Furthermore, the first entire neural network-based communication system using SDRs was implemented in [19], where an entire communication system was solely composed of neural networks for training and running. Since such a system fully considered time-varying in the actual channel, its performance was comparable to that of existing wireless communication systems.

A conditional generative adversarial network (CGAN) was applied in [20] to construct an end-to-end wireless communication system with unknown channel conditions. The encoded signal for transmitting was treated as condition information, and the transmitter and receiver of the wireless communication system were each replaced by a DNN. CGAN acted as a bridge between the transmitter and the receiver, allowing backpropagation to proceed smoothly, thereby jointly training and

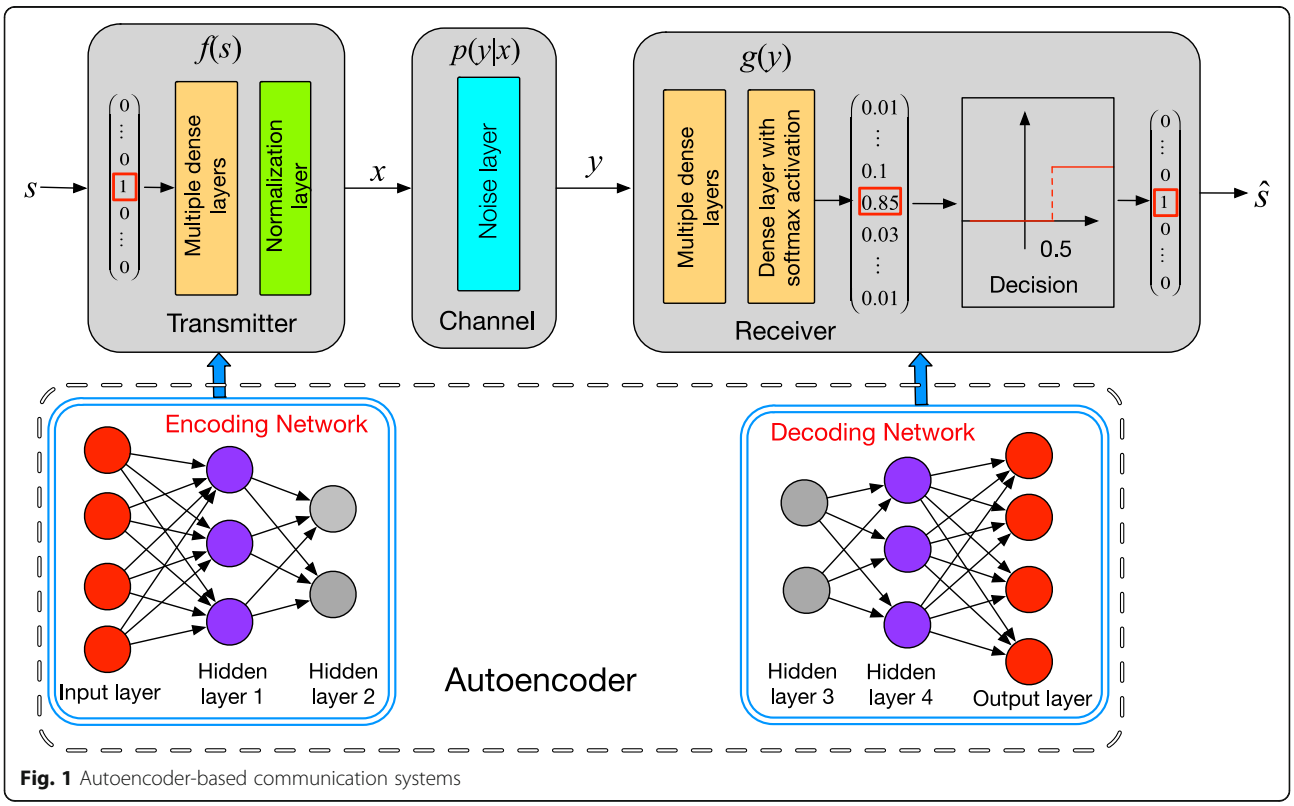

Fig. 1 Autoencoder-based communication systems

optimizing both the transmitter and receiver DNNs. This approach makes a significant breakthrough in the modeling mode of conventional wireless communications and opens up a new way for the design of future wireless communication systems.

Signal detection

Deep learning-based signal detection is getting more and more popular. Unlike the conventional model-based detection algorithms that rely on the estimation of the instantaneous channel state information (CSI) for detection, the deep learning-based detection method does not require to know the underlying channel model or the knowledge of the CSI when the channel model is known [21]. A sliding bidirectional recurrent neural network (SBRNN) was proposed in [22] for signal detection, where the trained detector was robust to changing channel conditions, eliminating the requirement for instantaneous CSI estimation.

Unlike traditional orthogonal frequency-division multiplexing (OFDM) receivers that first estimate the CSI explicitly, and then the estimated CSI is used to detect or restore the transmitted symbols, the deep learning-based method in [23] estimated the CSI implicitly and then recovered the transmitted signals directly. The estimated CSI was to solve the problem that a large amount of training data and high training cost were required due to a large increase in the number of parameters caused by DNNs.

Some recent works have suggested the use of DNNs in the context of MIMO detection and have developed model-driven deep learning networks for MIMO detection. For example, a network specifically designed for MIMO communication [24] can cope with time-varying channel in only one training phase. Instead of addressing a single fixed channel, a network obtained by unfolding the iterations of a projected gradient descent algorithm can handle multiple time-invariant and time-varying channels simultaneously in a single training phase [25]. Deep learning-based networks as demonstrated in [26] can reach near-optimal detection performance, guaranteed accuracy and robustness with low and flexible computational complexity.

Channel estimation and compression sensing

Channel estimation and compression sensing are key technologies for the real-time implementation of wireless communication systems. Channel estimation is the process of estimating the parameters of a certain channel model from the received data, while compression sensing is a technique to acquire and reconstruct sparse or compressible signals. Deep learning-based channel estimation and compression sensing methods have been suggested in several recent works [27–30].

To tackle the challenge of channel estimation when the receiver is equipped with a limited number of radio frequency (RF) chains in massive MIMO systems, a learned denoising-based approximate message passing (LDAMP) network was exploited in [27], where the channel structure can be learned and estimated from a large amount of training data. Experiment results demonstrated that the LDAMP network significantly outperforms state-of-the-art compressed sensing-based algorithms.

Motivated by the covariance matrix structure, a deep learning-based channel estimator was proposed in [28], where the estimated channel vector was a conditional Gaussian random variable, and the covariance matrix was random. Assisted by CNN and the minimum mean squared error (MMSE) estimator, the proposed channel estimator can ensure the state-of-the-art accuracy of channel estimation at a very lower computational complexity.

The basic architecture of deep learning-based CSI feedback is as shown in Fig. 2. Recently, more and more researchers have focused on the benefits of CSI feedback that the transmitter can utilize it to precode the signals before the transmission, thus we can gain the improvement of MIMO systems. The precoding technique can help to realize the high quality of restoring signals and are widely adopted in wireless communication systems. By exploiting CSI, the MIMO system can substantially reduce multi-user (MU) interference and provide a multifold increase in cell throughput. In the network of frequency division duplex (FDD) or time division duplex (TDD), the receiver UE can estimate the downlink CSI and transmit it back to the BS once they obtain it and help BS to perform precoding for the next signal. BS can also obtain the uplink CSI to help rectify the transmission at UE. The procedure of CSI feedback transmitting has drawn much attention, since high quality reconstructed CSI received by BS guarantees a good precoding, improving the stability and efficiency of the MIMO system.

Inspired by traditional compressed sensing technologies, a new CNN-based CSI sensing and recovery mechanism called CsiNet was proposed in [29], which effectively used the feedback information of training samples to sense and recover CSI, and achieved the potential benefits of a massive MIMO. The encoder of CsiNet converted the original CSI matrix into a codebook using CNN, and then the decoder restored the received codebook to the original CSI signal using the fully-connected network and refine networks.

To further improve the correctness of CSI feedback, a real-time long short-term memory (LSTM) based CSI feedback architecture named CsiNet-LSTM was proposed in [31], where CNN and RNN are applied to extract the spatial and temporal correlation features of CSI, respectively. Using time-varying MIMO channel

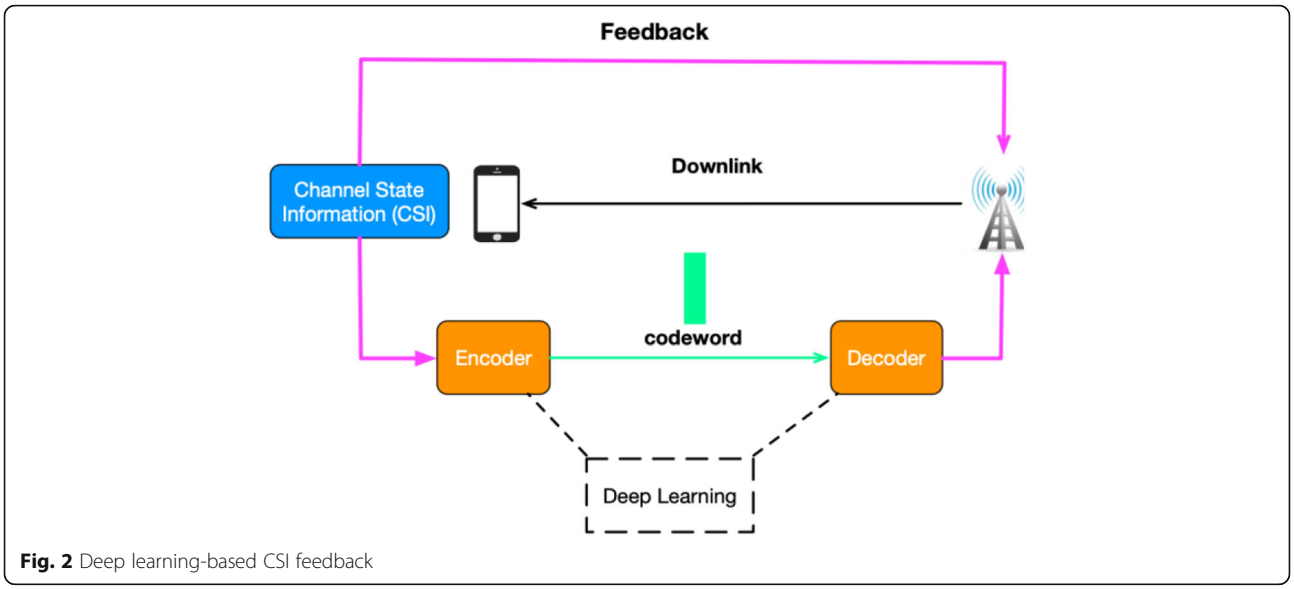

Fig. 2 Deep learning-based CSI feedback

time correlation and structural features, CsiNet-LSTM can achieve a tradeoff between compression ratio, CSI reconstruction quality, and complexity. Compared to CsiNet, the CsiNet-LSTM network can trade time efficiency for CSI reconstruction quality. Further, the deep autoencoder-based CSI feedback in the frequency division duplex (FDD) massive MIMO system was modelled in [30], which involved feedback transmission errors and delays.

As shown in Fig. 3, a novel effective CSI sensing and recovery mechanism in the FDD MIMO system was proposed in our previous work [32], referred to as ConvlstmCsiNet, which takes advantage of the memory characteristic of RNN in modules of feature extraction, compression and decompression, respectively. Moreover, we adopt depthwise separable convolutions in feature recovery to reduce the size of the model and interact information between channels. The feature extraction module is also elaborately devised by studying decoupled spatio-temporal feature representations in different structures.

Encoding and decoding

In digital communications, source coding and channel coding are typically required in data transmission. Deep learning methods have been suggested in some recent works [33–38] that can be used to improve standard source decoding and solve the problem of high computational complexity in channel decoding.

A DNN-based channel decoding method applied in [33] can directly realize the conversion from receiving codewords to information bits when considering the decoding part as a black box. Although this method shows advantages in performance improvement, learning is constrained with exponential complexity as the length of

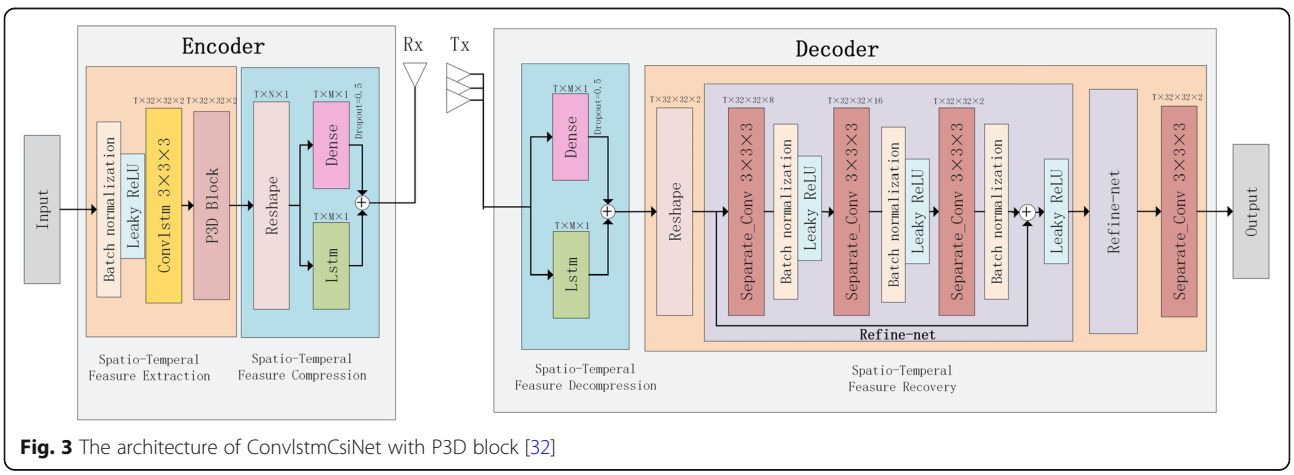

Fig. 3 The architecture of ConvlstmCsiNet with P3D block [32]

codewords increases. Therefore, it is neither fit for random codes, nor for codewords with long code lengths.

The issue of joint source encoding and channel encoding of structured data over a noisy channel was addressed in [38], a lower word error rate (WER) was achieved by developing deep learning-based encoders and decoders. This approach was optimal in minimizing end-to-end distortion where both the source and channel codes have arbitrarily large block lengths, however, it is limited in using a fixed length of information bits to encode sentences of different lengths.

Belief propagation (BP) algorithm can be combined with deep learning networks for channel decoding. Novel deep learning methods were proposed in [36, 37] to improve the performance of the BP algorithm. It demonstrated that the neural BP decoder can offer a tradeoff between error-correction performance and implementation complexity, but can only learn a single codeword instead of an exponential number of codewords. Neural network decoding was only feasible for very short block lengths, since the training complexity of deep learning-based channel decoders scaled exponentially with the number of information bits and. A deep learning polarization code decoding network with partitioned sub-blocks was proposed in [34] to improve its decoding performance for high-density parity check (HDPC) codes. By dividing the original codec into smaller sub-blocks, each of which can be independently encoded/decoded, it provided a promising solution to the dimensional problem. Furthermore, Liang et al. [35] proposed an iterative channel decoding algorithm BP-CNN, which combined CNN with a standard BP decoder to estimate information bits in a noisy environment.

Security and privacy

Due to the shared and broadcast nature of wireless medium, wireless communication systems are extremely vulnerable to attacks, counterfeiting and eavesdropping, and the security and privacy of wireless communications have received much attention [39, 40]. Moreover, wireless communication systems are becoming increasingly complex, and there is a close relationship between various modules of the system. Once a module is attacked, it will affect the operation of the entire wireless communication system.

Running AI functions on nearby edge servers or remote cloud servers is very vulnerable to security and AI data privacy issues. Thus, offloading AI learning models and collected data to external cloud servers for training and further processing may result in data loss due to the user's reluctancy of providing sensitive data such as location information. Many research efforts have focused on bridging DL and wireless security, including adversarial DL techniques, privacy Issues of DL solutions and DL hardening solutions [41, 42], to meet critical privacy and security requirements in wireless communications.

Conventional wireless communication systems generally suffer from jamming attacks, while autoencoder-based end-to-end communication systems are extremely susceptible to physical adversarial attacks. Small disturbances can be easily designed and generated by attackers. New algorithms for making effective white-box and black-box attacks on a classifier (or transmitter) were designed in [43, 44]. They demonstrated that physical adversarial attacks were more destructive in reducing the transmitter's throughput and success ratio when compared to jamming attacks. In addition, how to keep security and enhance the robustness of intelligent communication systems is still under discussion. Defense strategies in future communication systems are still immature and inefficient. Therefore, further research on the defense mechanisms of adversarial attacks and the security and robustness of deep learning-based wireless systems is very necessary.

One possible defense mechanism is to train the autoencoder to have an antagonistic perturbation, which is a technique that enhances robustness, known as the adversarial training [45]. Adversarial deep learning is applied in [46] to launch an exploratory attack on cognitive radio transmissions. In a canonical wireless communication scenario with one transmitter, one receiver, one attacker, and some background traffic, even the transmitter's algorithm is unknown to the attacker, it can still sense a channel, detect transmission feedback, apply a deep learning algorithm to build a reliable classifier, and effectively jam such transmissions. A defense strategy against an intelligent jamming attack on wireless communications was designed in [47] to successfully fool the attacker into making wrong predicts. To avoid the inaccurate learned model due to interference of the adversary, one possible way is to use DNNs in conjunction with GANs for learning in adversarial radio frequency (RF) environments, which are capable of distinguishing between adversarial and trusted signals and sources [48].

Open challenges

This section discusses several open challenges of deep learning-driven wireless communications from the aspects of baseline and dataset, model compression and acceleration, CSI feedback and reconstruction, complex neural networks, training at different SNRs and fast learning.

Baseline and dataset

The rapid development of computer vision, speech recognition, and natural language processing have benefited

most from the existence of many well-known and effective datasets in computer science, such as ImageNet [49] and MNIST [50]. For fairness, performance comparisons between different approaches should be performed under the same experimental environment by using common datasets. In order to compare the performance of newly proposed deep learning models and algorithms, it is critical to have some well-developed algorithms serving as benchmarks. Experiment results based on these benchmarks are usually called baselines, which are very important to show the development of a research field [51]. The quality and quantity of open datasets will have a huge impact on the performance of deep learning-based communication systems.

Wireless communication systems involve inherently artificial signals that can be synthesized and generated accurately, the local bispectrum, envelope, instantaneous frequency, and symbol rate of the signal can be used as input features. Therefore, in some cases, we should pay more attention to the standardization of data generation rules rather than the data itself.

In the field of intelligent wireless communications, however, there are few existing and public datasets that can be directly applied for training. It is necessary to either create generic and reliable datasets for different communication problems or develop new simulation software to generate datasets in various communication scenarios. On the basis of such dataset or data generation software, widely used datasets similar to ImageNet and MNIST can be created. Then, we can treat them as baselines or benchmarks for further comparison and research.

Model compression and acceleration

Deep neural networks (DNN) have achieved significant success in computer vision and speech recognition, in the meanwhile, their depth and width are still boosting, which lead to a sharp increase in the computational complexity of networks. At present, the number of parameters in DNN models is very huge (parameters are generally tens of millions to hundreds of millions) and thus the amount of calculation is extremely large. Current deep learning models either rely on mobile terminals or edge-cloud server to run AI functions and are under tremendous pressure in terms of high data storage and processing demands [41]. Offloading complex compute tasks from mobile terminals to a central cloud with AI functions can alleviate the limitation of computation capacity, but also results in high latency for AI processing due to long-distance transmissions. Therefore, it is not appropriate to offload AI learning model to the central cloud server, especially for data-intensive and delay-sensitive tasks.

Some deep learning algorithms deployed on mobile terminals can only rely on cloud graphic processing units (GPUs) to accelerate computing, however, the wireless bandwidth, the communication delay, and the security of cloud computing will incur enormous obstacles. The large memory and high computational consumption required by the DNN greatly restricts the use of deep learning on mobile terminals with limited resources. Deep learning-based communication systems are also difficult to deploy on small mobile devices such as smartphones, smartwatches and tablets.

Due to the huge redundancy of the parameters in DNN models, these models can be compressed and accelerated to build a lightweight network, which is an inevitable trend in the development of related technologies in the future. Methods like low-rank factorization, parameter pruning and sharing, quantization, and knowledge distillation can be applied in DNN models. Specifically, on the one hand, it is possible to consider quantifying the parameters of DNN models to further compress the network model; on the other hand, channel pruning and structured sparse constraints can be applied to eliminate part of the redundant structure and accelerate the calculation speed [52].

Lightweight AI engines at the mobile terminals are required to perform real-time mobile computing and decision making without the reliance of edge-cloud servers, where the centralized model is stored in the cloud server while all training data is stored on the mobile terminals. In addition, learning parameter settings or updates are implemented by local mobile devices. In some cases, if the floating-point calculation or storage capacity of the network model is greatly reduced, but the performance of the existing DNNs remains essentially unchanged, such a network model can run efficiently on resource-constrained mobile devices.

CSI feedback and reconstruction

The massive multiple-input multiple-output (MIMO) system is usually operated in OFDM over a large number of subcarriers, leading to a problem of channel state information (CSI) feedback overload. Moreover, in order to substantially provide a multifold increase in cell throughput, each base station is equipped with thousands of antennas in a centralized or distributed manner [29]. Therefore, it is crucial to utilize the available CSI at the transmitter for precoding to improve the performance of FDD networks [32]. However, compressing a large amount of CSI feedback overload in massive MIMO systems is very challenging. Traditional estimation approaches like compressive sensing (CS) can only achieve poor performance on CSI compression in real MIMO system due to the harsh preconditions.

Although DL-based CSI methods outperform much than the CS ones, the price of training cost remains high, which requires large quantities of channel estimates. Once the wireless environment changes significantly, a trained model still has to be retrained [53]. In addition, a more able and efficient structure of DNN is needed. The design of CSI feedback link and precoding mode still remains an open issue that different MIMO systems should adopt their own appropriate designed CSI feedback link and precoding manner. Furthermore, DL-based CSI feedback models are still immature when adopted in real massive MIMO systems and suffer constraints of realistic factors, e.g., time-varying channel with fading, SRS measurement period, channel capacity limitation, hardware or device configuration, channel estimation and signal interference in MU systems. These challenges may hinder the general applications temporarily and will be addressed by future DL-based models with a more exquisite and advanced architecture.

Complex neural networks

Due to the widely used baseband representations in wireless communication systems, data is generally processed in complex numbers, and most of the associated signal processing algorithms rely on phase rotation, complex conjugate, absolute values, and so on [10].

Therefore, neural networks have to run on complex values rather than real numbers. However, current deep learning libraries usually do not support complex processing. While complex neural networks may be easier to train and consume less memory, they do not provide any significant advantages in terms of performance. At present, we can only think of a complex number as a real number and an imaginary number. Complex neural networks that are suitable for wireless communication models should be developed.

Training at different SNRs

Up to now, it is still not clear which signal-to-noise (SNR) ratio the deep learning model should be trained on. The ideal deep learning model should be applied to any SNR regardless of the SNR used for training or the range of SNR it is in. In fact, however, this is not the case. The results of training deep learning models under certain SNR conditions are often not suitable for other SNR ranges [10].

For example, training at lower SNRs does not reveal important structural features of wireless communication systems at higher SNRs, and similarly, training at higher SNRs can not reveal important structural features of wireless communication systems at lower SNRs. Training the deep learning model across different SNRs can also seriously affect the training time. In addition, how to construct an appropriate loss function, how to adjust parameters and data representation for wireless communication systems are still big problems that must be solved.

Fast learning

For end-to-end training of wireless communication systems including encoders, channels, and decoders, a specific channel model is usually required. The trained model needs to be applied to its corresponding channel model, otherwise, mismatch problems will occur, which will cause severe degradation of system performance.

In real-world scenarios, however, due to many environmental factors, the channel environment often changes at any time and place, e.g., the change of the movement speed and direction of user terminals, the change of the propagation medium, the change of the refractive scattering environment. Once the channel environment changes, a large amount of training data is needed to retrain, which means that for different channel environments at each moment, such repeated training tasks need to be performed, which consumes resources and weakens the performance of the system.

Retraining is required when the system configuration changes because the system model does not have a good generalization ability. Adaptation is done on a per-task basis and is specific to the channel model [54]. Some changes in the channel environment may lead to a sharp decline in system performance. Therefore, we need to seek systems with stronger generalization ability, in order to adapt to the changing channel environment.

Potential opportunities

This section mainly describes the profound potential opportunities and the promising research directions in wireless communications assisted by the rapid development of deep learning.

Deep learning-driven CSI feedback in massive MIMO system

Recent researches indicate that applying deep learning (DL) in MIMO systems to address the nonlinear problems or challenges can indeed boost the quality of CSI feedback compression. Different from the traditional CS-based approaches, DL-based CSI methods adopt several neural network (NN) layers as an encoder replacing the CS model to compress CSI as well as a decoder to recover the original CSI, which can speed up the transmitting runtime nearly 100 times of CS ones.

The structure of autoencoder-based MIMO systems is depicted in Fig. 4, which only considers the downlink CSI feedback process, assuming that the feedback channel is perfect enough to transmit CSI with no impairments. In fact, a large part of the overload CSI serves redundant and the CSI matrix falls to sparse in the delay

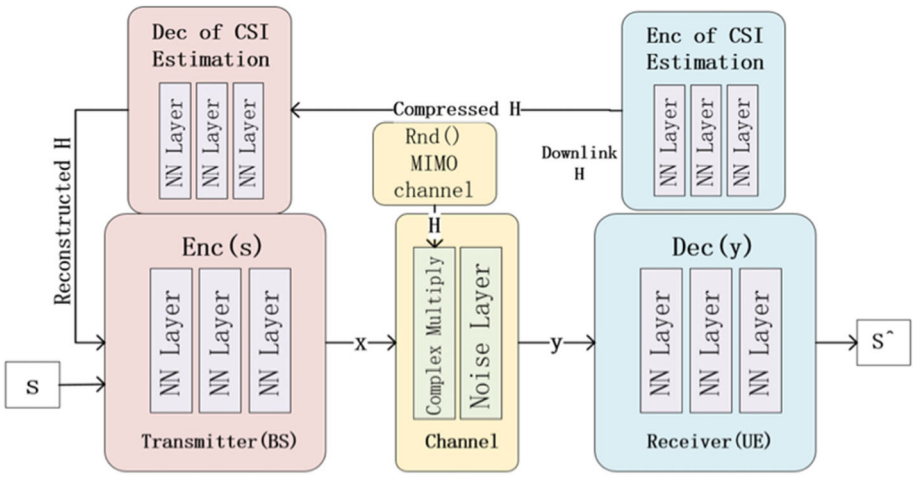

Fig. 4 The structure of autoencoder-based MIMO systems with downlink CSI feedback

domain. In order to remove the information redundancy, CNNs are applied here, which has the ability to eliminate the threshold of domain expertise since CNNs use hierarchical feature extraction, which can effectively extract information and obtain increasingly abstract correlations from the data while minimizing data preprocessing workload.

We can consider both the issues of feedback delay and feedback errors. Assume that one signal is transmitted into n time slots due to the restriction of downlink bandwidth resource, thus demanding a *n*-length time series of CSI feedback estimation within a signal transmitting period and the SRS measurement period. The time-varying channel is also under the condition of known overdue CSI or partial CSI characteristics, such as Doppler or beam-delay information. Furthermore, the feedback errors from MU interference brought by multiple UE at middle or high moving speed are also taken into account. When transmitting the compressed CSI feedback, the imperfections, e.g. additive white Gaussian noise (AWGN), in uplink CSI feedback channel would also bring feedback errors. The model is trained to minimize the feedback errors via the minimum mean square error (MMSE) detector.

The architecture of DL-based autoencoder in CSI feedback compression is also advanced via taking the advantages of RNN's memory characteristic to deal with the feature extraction in time-varying channel, which can have an active effect on time correlation exploring and better performance on CSI recovery [30]. Similarly, a DL-based autoencoder of CSI estimation method can be applied in this MIMO system, which is exposed to more practical restrictions.

In the future, we can use DL methods of CSI feedback with time-varying channel in massive MU-MIMO system to improve the compression efficiency and speed up the transmitting process, as well as develop novel theoretical contributions and practical research related to the new technologies, analysis and applications with the help of CNN and RNN.

GAN-based Mobile data augmentation

Mobile data typically comes from a variety of sources with various formats and exhibits complex correlations and heterogeneity. According to the mobile data, conventional machine learning tools require cumbersome feature engineering to make accurate inferences and decisions. Deep learning has eliminated the threshold of domain expertise because it uses hierarchical feature extraction, which can effectively extract information and obtain increasingly abstract correlations from the data while minimizing data pre-processing workload [55]. However, inefficiency in training time is an enormous challenge when applying learning algorithms in wireless systems. Traditional supervised learning methods, which learn a function that maps the input data to some desired output class label, is only effective when sufficient labeled data is available. On the contrary, generative models, e.g., GAN and variational autoencoder (VAE), can learn the joint probability of the input data and labels simultaneously via Bayes rule [56]. Therefore, GANs and VAEs are well suitable for learning in wireless environments since most current mobile systems generate unlabeled or semi-labeled data.

GANs can be used to enhance the configuration of mobile and wireless networks and help address the growth of data volumes and algorithm-driven applications to satisfy the large data needs of DL algorithms. GAN is a method that allows exploiting unlabeled data to learn useful patterns in an unsupervised manner. GANs can be further applied in B5G mobile and wireless

networks, especially in dealing with heterogeneous data generated by mobile environments.

As shown in Fig. 5, the GAN model consists of two neural networks that compete against each other. The generator network tries to generate samples that resemble the real data such that the discriminator cannot tell whether it is real or fake. After training the GAN, the output of the generator is fed to a classifier network during the inference phase. We can use GAN to generate real data according to previously collected real-world data. Furthermore, it can be used for path planning, trajectories analysis and mobility analysis.

Monitoring large-scale mobile traffic is, however, a complex and costly process that relies on dedicated probes, which have limited precision or coverage and gather tens of gigabytes of logs daily [57]. Heterogeneous network traffic control is an enormous obstacle due to the highly dynamic nature of large-scale heterogeneous networks. As for a deep learning system, it has difficulty in characterizing the appropriate input and output patterns [58].

GANs can be applied in resource management and parameter optimization to adapt to the changes in the wireless environment. To make this happen, intelligent control of network traffic can be applied to infer fine-grained mobile traffic patterns, from aggregate measurements collected by network probes. New loss functions are required to stabilize the adversarial model training process, and prevent model collapse or non-convergence problems. Further, data processing and augmentation procedure are required to handle the insufficiency of training data and prevent the neural network model from over-fitted.

Deep learning-driven end-to-end communication

The purpose of autoencoder is to make the input and the output as similar as possible, which is achieved by performing backpropagation of the error and continuing optimization after each output. Similarly, a simple wireless communication system consists of a transmitter (encoder), a receiver (decoder) through a channel, and an abundant of physical layer transmission technologies can be adopted in the wireless communication process. A communication system over an additive white gaussian noise (AWGN) or Rayleigh fading channel can be represented as a particular type of autoencoder. The purpose of wireless communication is to make the output signal and the input signal as similar as possible. However, how to adapt an end-to-end communications system trained on a statistical model to a real-world implementation remains an open question.

As shown in Fig. 6, we can extend the above single channel model to two or more channels, where multiple transmitter and multiple receivers are competing for the channel capacity. As soon as some of the transmitters and receivers are non-cooperative, adversarial training strategies such as GANs could be adopted. We can perform joint optimization for common or individual performance metrics such as block error rate (BLER). However, how to train two mutually coupled autoencoders is still a challenge. One

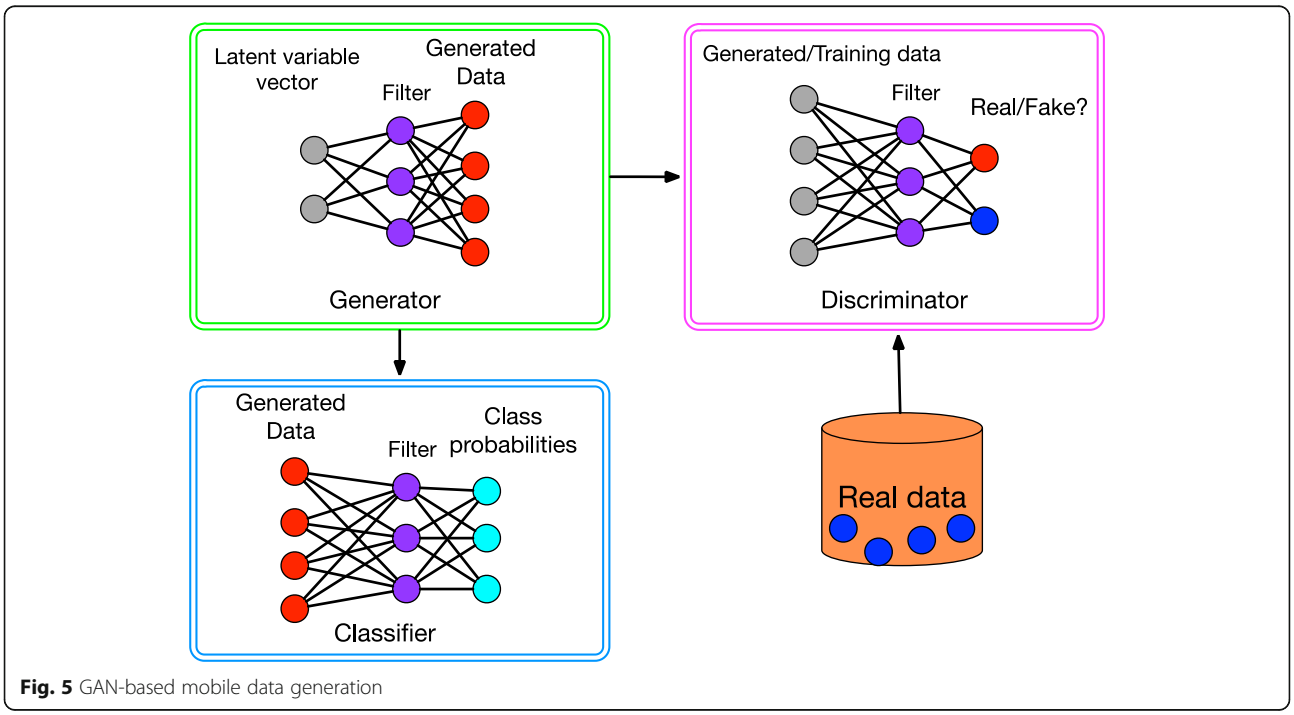

Fig. 5 GAN-based mobile data generation

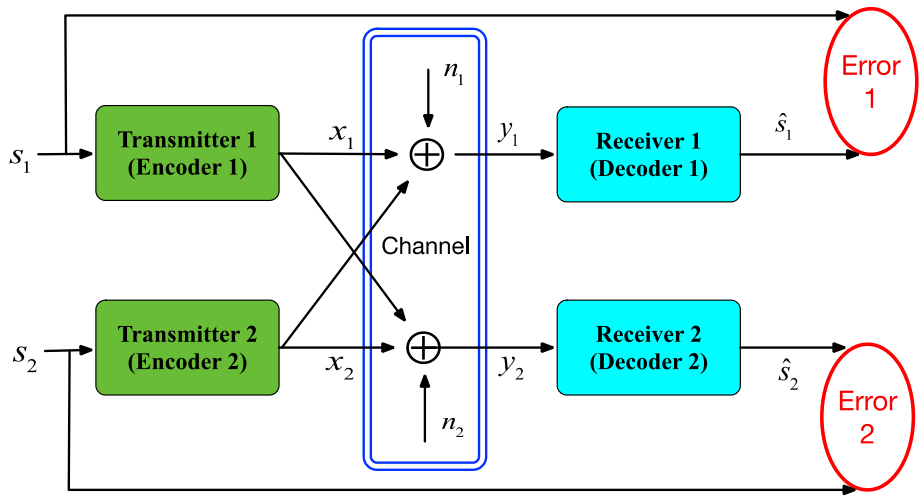

Fig. 6 Autoencoder-based MIMO System

suggestion is to assign dynamic weights to different autoencoders and minimize the weighted sum of the two loss functions.

The diagram of the energy-based generative adversarial network (EBGAN) [59] in wireless communications is depicted in Fig. 7. We use an encoder instead of a transmitter, and a decoder instead of a receiver for intelligent communications. The generative network is applied to generate the canonicalized signal, and then fed into the discriminative network for further classification. Inverse filtering can be applied to simplify the task of the learned discriminative network. Similarly, the purpose of EBGAN-based end-to-end communication is to make the output signal and the input signal as close as possible.

The discriminator D is structured as an autoencoder:

$$D(x) = \|Dec(Enc(x)) - x\| \tag{1}$$

where $Enc(\cdot)$ and $Dec(\cdot)$ denote the encoder function and decoder function, respectively.

Given a positive margin m, a data sample x and a generated sample $G(z)$, the discriminator loss LD and the generator loss LG are formally defined by:

$$\begin{aligned}\mathcal{L}_D(x,z) &= D(x) + \{m - D(G(z))\}_+ \\ &= \|Dec(Enc(x)) - x\| \\ &+ \{m - \|Dec(Enc(G(z))) - G(z)\|\}_+\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_G(z) &= D(G(z)) \\ &= \|Dec(Enc(G(z))) - G(z)\|\end{aligned} \tag{3}$$

where $\{\cdot\}_+$ is the hinge loss function, the generator is trained to produce contrastive samples with minimal energies, while the discriminator is trained to assign high energies to generated samples [59].

Most mathematical models in wireless communication systems are static, linear, and Gaussian-compliant optimization models. However, a realistic communication system has many imperfect and non-linear problems, e.g., nonlinear power amplifiers, which can only be

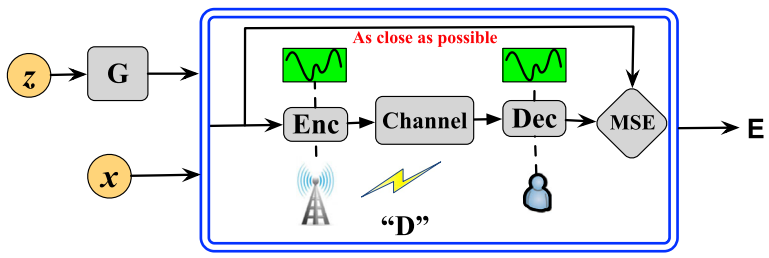

Fig. 7 EBGAN with an autoencoder discriminator in wireless communications

captured by these models. The EBGAN-based wireless communication system no longer requires a mathematical linear processing model that can be optimized for specific hardware configurations or spatially correlated channels. With the help of EBGAN, we can learn about the implementation details of the transmitter and receiver and even the information coding without any prior knowledge.

Meta-learning to wireless communication

In real-world scenarios, it is not worthwhile to perform multi-tasks training from scratch just because of different channel models, because these tasks are closely related, they share the same encoder and decoder network structure, and their parameter changes are only affected by the channel model. Training from scratch is under the assumption that such tasks are completely independent and cannot make full use of the connections, resulting in many repetitive and redundant training steps, however, it is not true.

Meta-learning, or learning to learn [60], that is, to make the model a learner. It learns a priori knowledge in multi-tasking and then quickly applies it to the learning of new tasks, so that fast learning and few-shot learning can be realized. Meta-learning provides a way to perform multi-task learning and optimizes the system parameters toward a common gradient descent direction during training, thereby achieving the optimal generalization ability and reduced training data and/or time complexity. In the meantime, when a new task arrives, the system can train on a few rounds of iterative (or even one round of iterative) with very little training data, so that the parameters can be dynamically fine-tuned on the basis of the original learning model to adapt to the new channel model, where the dynamic parameter tuning is possible. Thus, meta-learning can be implemented for end-to-end learning of encoder and decoder with unknown or changing wireless channels, and it outperforms conventional training and joint training in wireless communication systems [54].

A specific example of meta-training methods known as model agnostic meta-learning (MAML) [61]. Its core idea is to find a common initialization point that allows for a quick adaptation towards the optimal performance on the new task. MAML updates parameters through one or more stochastic gradient descent (SGD) steps, which are calculated using only a small amount of data from the new task. Therefore, instead of training a common system model for all channel models, we can apply MAML to find a common initialization vector so that it supports fast training on any channel [54].

Conclusion

Several recent efforts have focused on intelligent communications to harvest remarkable potential benefits. We have mainly discussed the potential applicability of deep learning in the field of wireless communications for edge-cloud computing, such as model-free training method for end-to-end wireless communications, and further demonstrated their superior performance over conventional wireless communications. Implementation of many emerging deep learning technologies are still in the preliminary stage, and profound potential solutions to solving wireless communication problems have to be further studied. This survey attempts to summarize the research progress in deep learning-driven wireless communications and point out existing bottlenecks, future opportunities and trends.

In the research of B5G wireless networks and communication systems, the low efficiency of training time is a bottleneck when applying learning algorithms in wireless systems. Although deep learning is not mature in wireless communications, it is regarded as a powerful tool and hot research topic in many potential application areas, e.g., channel estimation, wireless data analysis, mobility analysis, complicated decision-making, network management, and resource optimization. It is worthwhile to investigate the use of deep learning techniques in wireless communication systems to speed up the training process and develop novel theoretical contributions and practical research related to the new technologies, analysis and applications for edge-cloud computing.

Abbreviations

AI: Artificial intelligence; AR: Augmented reality; AWGN: Additive white gaussian noise; BLER: Block error rate; BP: Belief propagation; BS: Base station; B5G: Beyond the fifth-generation; CGAN: Conditional generative adversarial network; CNN: Convolutional neural network; CSI: Channel state information; DNN: Deep neural network; EBGAN: Energy-based generative adversarial network; FDD: Frequency division duplex; GAN: Generative adversarial network; GPU: Graphics processing unit; HDPC: High-density parity check; LDAMP: Learned denoising-based approximate message passing; LSTM: Long short-term memory; MMSE: ModelAgnostic meta-learning; MAML: Minimum mean squared error; MIMO: Multiple-input multiple-output; MU: Multi-user; NGN: Next-generation network; OFDM: Orthogonal frequency-division multiplexing; RF: Radio frequency; RNN: Recurrent neural network; SDR: Software-defined radio; SBRNN: Sliding bidirectional recurrent neural network; SGD: Stochastic gradient descent; TDD: Time division duplex; IoT: Internet of things; UE: User Equipment; VR: Virtual reality; WER: Word error rate; 5G: Fifth-Generation

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Authors' informations

Huaming Wu received the B.E. and M.S. degrees from Harbin Institute of Technology, China in 2009 and 2011, respectively, both in electrical engineering. He received the Ph.D. degree with the highest honor in computer science at Free University of Berlin, Germany in 2015. He is currently an associate professor in the Center for Applied Mathematics, Tianjin University. His research interests include mobile cloud computing, edge computing, fog computing, internet of things (IoTs), and deep learning.

Xiangyi Li received the B.S. in Applied Mathematics from Tianjin University, China. She is currently a M.S. student majoring in applied mathematics at Tianjin University, China. Her research interests include deep learning, wireless communications and generative models.

Yingjun Deng received the B.S. in Applied Mathematics (2009) and M.S. in Computational Mathematics (2011) from Harbin Institute of Technology, China. He got his Ph.D. in Systems Optimization and Dependability from Troyes University of Technology, France in 2015. He worked as a postdoctoral fellow, respectively at University of Waterloo in Canada (2015–2016), and Eindhoven University of Technology in Netherlands (2018–2019). He became a lecturer since 2016 in the Center for Applied Mathematics, Tianjin University, China. His research interests include applied statistics, deep learning, prognostic and health management, and predictive maintenance.

Authors' contributions

HW designed the survey and led the write up of the manuscript. XL contributed part of the writing of the manuscript. YD took part in the discussion of the work described in this paper. All authors have read and approved the manuscript.

Funding

This work is partially supported by the National Key R & D Program of China (2018YFC0809800), the National Natural Science Foundation of China (61801325), the Huawei Innovation Research Program (HO2018085138), the Natural Science Foundation of Tianjin City (18JCQNJC00600), and the Major Science and Technology Project of Tianjin (18ZXRHSY00160).

Availability of data and materials

No

Competing interests

The authors declare that they have no competing interests.

Received: 11 February 2020 Accepted: 30 March 2020
Published online: 10 April 2020

References

1. Ma Z, Xiao M, Xiao Y, Pang Z, Poor HV, Vucetic B (2019) High-reliability and low-latency wireless communication for internet of things: challenges, fundamentals, and enabling technologies. IEEE Internet Things J 6(5):7946–7970
2. Liu G, Wang Z, Hu J, Ding Z, Fan P (2019) Cooperative NOMA broadcasting/multicasting for low-latency and high-reliability 5g cellular v2x communications. IEEE Internet Things J 6(5):7828–7838
3. Xu X, Liu X, Xu Z, Dai F, Zhang X, Qi L (2019) Trust-oriented IoT service placement for smart cities in edge computing. IEEE Internet Things J
4. Lai P, He Q, Cui G, Xia X, Abdelrazek M, Chen F, Hosking J, Grundy J, Yang Y (2019) Edge user allocation with dynamic quality of service. In: International Conference on Service-Oriented Computing. Springer, Cham, pp 86–101
5. Qi L, Chen Y, Yuan Y, Fu S, Zhang X, Xu X (2020) A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems. World Wide Web 23, pp 1275-1297
6. Xu X, Chen Y, Zhang X, Liu Q, Liu X, Qi L (2019) A blockchain-based computation offloading method for edge computing in 5G networks. Software: Practice and Experience. Wiley, pp 1–18
7. Xu X, Zhang X, Gao H, Xue Y, Qi L, Dou W (2020) Become: Blockchain-enabled computation offloading for IoT in mobile edge computing. IEEE Trans Ind Inform 16(6):4187-4195
8. Wu H, Wolter K (2018) Stochastic analysis of delayed mobile offloading in heterogeneous networks. IEEE Trans Mob Comput 17(2):461–474
9. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82
10. O'Shea T, Hoydis J (2017) An introduction to deep learning for the physical layer. IEEE Trans Cogn Commun Netw 3(4):563–575
11. Felix A, Cammerer S, Dörner S, Hoydis J, Ten Brink S (2018) OFDM-autoencoder for end-to-end learning of communications systems. In: 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, pp 1–5
12. Jang Y, Kong G, Jung M, Choi S, Kim I-M (2019) Deep autoencoder based CSI feedback with feedback errors and feedback delay in FDD massive MIMO systems. IEEE Wireless Commun Lett 8(3):833–836
13. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems-volume 2. MIT Press, pp 2672–2680
14. Wu H, Han Z, Wolter K, Zhao Y, Ko H (2019) Deep learning driven wireless communications and mobile computing. Wirel Commun Mob Comput 2019:1–2
15. O'Shea TJ, Erpek T, Clancy TC (2017) Deep learning based MIMO communications. arXiv preprint arXiv 1707:07980
16. Qin Z, Ye H, Li GY, Juang B-HF (2019) Deep learning in physical layer communications. *IEEE Wirel Commun* 26(2):93–99
17. Aoudia FA, Hoydis J (2018) End-to-end learning of communications systems without a channel model. In: 2018 52nd Asilomar Conference on Signals, Systems, and Computers. IEEE, pp 298–303
18. Aoudia FA, Hoydis J (2019) Model-free training of end-to-end communication systems. IEEE J Selected Areas Commun 37(11):2503–2516
19. Dörner S, Cammerer S, Hoydis J, ten Brink S (2018) Deep learning based communication over the air. IEEE J Selected Top Signal Process 12(1):132–143
20. Ye H, Li GY, Juang B-HF, Sivanesan K (2018) Channel agnostic end-to-end learning based communication systems with conditional GAN. In: 2018 IEEE Globecom Workshops (GC Wkshps). IEEE, pp 1–5
21. Wang T, Wen C-K, Wang H, Gao F, Jiang T, Jin S (2017) Deep learning for wireless physical layer: opportunities and challenges. China Communications 14(11):92–111
22. Farsad N, Goldsmith A (2018) Neural network detection of data sequences in communication systems. IEEE Trans Signal Process 66(21):5663–5678
23. Ye H, Li GY, Juang B-H (2018) Power of deep learning for channel estimation and signal detection in OFDM systems. IEEE Wireless Commun Lett 7(1):114–117
24. He H, Wen C-K, Jin S, Li GY (2018) A model-driven deep learning network for MIMO detection. In: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, pp 584–588
25. Samuel N, Diskin T, Wiesel A (2017) Deep MIMO detection. In: 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), IEEE, pp 1–5
26. Samuel N, Diskin T, Wiesel A (2019) Learning to detect. IEEE Trans Signal Process 67(10):2554–2564
27. He, H., Jin, S., Wen, C., Gao, F., Ye Li, G., Xu, Z.: Model-driven deep learning for physical layer communications. IEEE Wireless Commun, 1–7 (2019)
28. Neumann D, Wiese T, Utschick W (2018) Learning the MMSE channel estimator. IEEE Trans Signal Process 66(11):2905–2917
29. Wen C-K, Shih W-T, Jin S (2018) Deep learning for massive MIMO CSI feedback. IEEE Wireless Commun Lett 7(5):748–751
30. Lu C, Xu W, Shen H, Zhu J, Wang K (2019) MIMO channel information feedback using deep recurrent network. IEEE Commun Lett 23(1):188–191
31. Wang T, Wen C-K, Jin S, Li GY (2019) Deep learning-based CSI feedback approach for time-varying massive MIMO channels. IEEE Wireless Commun Lett 8(2):416–419
32. Li X, Wu H (2020) Spatio-temporal representation with deep neural recurrent network in MIMO CSI feedback. IEEE Wireless Communications Letters
33. Gruber, T., Cammerer, S., Hoydis, J., Ten Brink, S.: On deep learning-based channel decoding. In: 2017 51st Annual Conference on Information Sciences and Systems (CISS), pp. 1–6 (2017). IEEE
34. Cammerer S, Gruber T, Hoydis J, ten Brink S (2017) Scaling deep learning-based decoding of polar codes via partitioning. In: GLOBECOM 2017-2017 IEEE Global Communications Conference. IEEE, pp 1–6
35. Liang F, Shen C, Wu F (2018) An iterative BP-CNN architecture for channel decoding. IEEE J Selected Top Signal Process 12(1):144–159
36. Nachmani E, Be'ery Y, Burshtein D (2016) Learning to decode linear codes using deep learning. In: 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, pp 341–346
37. Nachmani E, Marciano E, Lugosch L, Gross WJ, Burshtein D, Be'ery Y (2018) Deep learning methods for improved decoding of linear codes. IEEE J Selected Top Signal Process 12(1):119–131
38. Farsad N, Rao M, Goldsmith A (2018) Deep learning for joint source-channel coding of text. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 2326–2330
39. Meng T, Wolter K, Wu H, Wang Q (2018) A secure and cost-efficient offloading policy for mobile cloud computing against timing attacks. Pervasive Mobile Comput 45:4–18
40. Tian Q, Lin Y, Guo X, Wen J, Fang Y, Rodriguez J, Mumtaz S (2019) New security mechanisms of high-reliability IoT communication based on radio frequency fingerprint. IEEE Internet Things J 6(5):7980–7987

41. Nguyen, D.C., Cheng, P., Ding, M., Lopez-Perez, D., Pathirana, P.N., Li, J., Seneviratne, A.: Wireless AI: enabling an AI-governed data life cycle (2020)
42. Sagduyu YE, Shi Y, Erpek T, Headley W, Flowers B, Stantchev G, Lu Z (2020) When wireless security meets machine learning: Motivation, challenges, and research directions. arXiv preprint arXiv 2001:08883
43. Sadeghi M, Larsson EG (2019) Physical adversarial attacks against end-to-end autoencoder communication systems. IEEE Commun Lett 23(5):847–850
44. Sadeghi M, Larsson EG (2019) Adversarial attacks on deep-learning based radio signal classification. IEEE Wireless Commun Lett 8(1):213–216
45. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. Stat 1050:20
46. Shi Y, Sagduyu YE, Erpek T, Davaslioglu K, Lu Z, Li JH (2018) Adversarial deep learning for cognitive radio security: jamming attack and defense strategies. In: 2018 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, pp 1–6
47. Erpek T, Sagduyu YE, Shi Y (2019) Deep learning for launching and mitigating wireless jamming attacks. IEEE Trans Cognitive Commun Netw 5(1):2–14
48. Roy D, Mukherjee T, Chatterjee M (2019) Machine learning in adversarial RF environments. IEEE Commun Mag 57(5):82–87
49. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 248–255
50. Deng L (2012) The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag 29(6): 141–142
51. Zhang MM, Shang K, Wu H (2019) Learning deep discriminative face features by customized weighted constraint. Neurocomputing 332:71–79
52. Liu C, Wu H (2019) Channel pruning based on mean gradient for accelerating convolutional neural networks. Signal Process 156:84–91
53. Qing C, Cai B, Yang Q, Wang J, Huang C (2019) Deep learning for CSI feedback based on superimposed coding. IEEE Access 7:93723–93733
54. Simeone O, Park S, Kang J (2020) From learning to meta-learning: Reduced training overhead and complexity for communication systems. arXiv preprint arXiv:2001–01227
55. Zhang C, Patras P, Haddadi H (2019) Deep learning in mobile and wireless networking: a survey. IEEE Commun Surv Tutorials 21(3):2224–2287
56. Jagannath J, Polosky N, Jagannath A, Restuccia F, Melodia T (2019) Machine learning for wireless communications in the internet of things: a comprehensive survey. Ad Hoc Netw 93:101913
57. Mohammadi, M., Al-Fuqaha, A., Oh, J.-S.: Path planning in support of smart mobility applications using generative adversarial networks. In: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), pp. 878–885 (2018). IEEE
58. Wang M, Cui Y, Wang X, Xiao S, Jiang J (2018) Machine learning for networking: workflow, advances and opportunities. IEEE Netw 32(2):92–99
59. Zhao J, Mathieu M, LeCun Y (2017) Energy-based generative adversarial network. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon
60. Andrychowicz M, Denil M, Gomez S, Hoffman MW, Pfau D, Schaul T, Shillingford B, De Freitas N (2016) Learning to learn by gradient descent by gradient descent. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 3981–3989
61. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. Sydney, JMLR.org, pp 1126–1135