SCIENCE CHINA
Information Sciences

• REVIEW •

February 2019, Vol. 62 021301:1–021301:13
https://doi.org/10.1007/s11432-018-9596-5

AI for 5G: research directions and paradigms

Xiaohu YOU[1], Chuan ZHANG[1*], Xiaosi TAN[1], Shi JIN[1] & Hequan WU[2]

[1]*National Mobile Communications Research Laboratory, Southeast University, Nanjing* 210096*, China;*
[2]*Chinese Academy of Engineering, Beijing* 100088*, China*

Received 23 July 2018/Revised 1 August 2018/Accepted 12 September 2018/Published online 26 October 2018

Abstract Wireless communication technologies such as fifth generation mobile networks (5G) will not only provide an increase of 1000 times in Internet traffic in the next decade but will also offer the underlying technologies to entire industries to support Internet of things (IOT) technologies. Compared to existing mobile communication techniques, 5G has more varied applications and its corresponding system design is more complicated. The resurgence of artificial intelligence (AI) techniques offers an alternative option that is possibly superior to traditional ideas and performance. Typical and potential research directions related to the promising contributions that can be achieved through AI must be identified, evaluated, and investigated. To this end, this study provides an overview that first combs through several promising research directions in AI for 5G technologies based on an understanding of the key technologies in 5G. In addition, the study focuses on providing design paradigms including 5G network optimization, optimal resource allocation, 5G physical layer unified acceleration, end-to-end physical layer joint optimization, and so on.

Keywords 5G mobile communication, AI techniques, network optimization, resource allocation, unified acceleration, end-to-end joint optimization

Citation You X H, Zhang C, Tan X S, et al. AI for 5G: research directions and paradigms. Sci China Inf Sci, 2019, 62(2): 021301, https://doi.org/10.1007/s11432-018-9596-5

1 Introduction

Fifth generation mobile networks (5G) implement the next generation of mobile telecommunication standards that aim to meet the demands of mobile communication in 2020 and beyond. 5G aims to provide a complete wireless communication system with diverse applications. Specifically, 5G is responsible for supporting three generic services that are classified as enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) (also referred to as mission-critical communications). These applications suggest new performance criteria for latency, reliability, connection and capacity density, system spectral efficiency, energy efficiency, and peak throughput that must be addressed using the 5G technology. To meet these criteria, ongoing studies are being conducted in many areas primarily focusing on key technologies including massive multiple-input multiple-output (MIMO), new radio access technology (RAT), heterogeneous ultra-densification networks (UDN), channel coding and decoding (e.g., polar codes), and mmWave access [1–4]. In addition, 5G networks will inevitably be heterogeneous with multiple modes and devices implemented through one unified air interface tailored for specific applications. Therefore, architectures such as the dense HetNet are involved and 5G systems are going to be virtualized and implemented over cloud data centers.

* Corresponding author (email: chzhang@seu.edu.cn)

Network slicing will be a major feature of a 5G network, including the use of a new air interface designed to dynamically optimize the allocation of network resources and utilize the spectrum efficiently [5].

The 5G technology standards are in development and are in progress to becoming complete and mature [6, 7]. In December 2017, the 3G Partnership Project (3GPP) officially announced new standards for 5G new radio (NR), which include standards for 5G non-standalone architecture (NSA) and eMBB [8]. On June 14, 2018, 3GPP formally completed the standalone (SA) version of the 5G NR standard, marking a long-awaited target date for 5G standardization [9]. These announced standards effectively set the stage for launching full-scale and cost-effective developments in 5G networks. Compared to current 4G networks, 5G NR: (1) enhances the MIMO systems using massive MIMO technology; (2) makes complete time slot structures and resource block (RB) allocation of the orthogonal frequency-division multiplexing (OFDM), proposing a more flexible air interface; (3) will introduce the non-orthogonal multiple access (NOMA) to support the Internet of things (IoT) in the near future; (4) follows previous distributed antenna systems [10], splits the wireless functions into distributed units (DU) and central units (CU), and applies network virtualization and network slicing techniques based on cloud computing.

Overall, 5G networks will tailor the provisioning mechanisms for more applications and services, which makes it more challenging in terms of complicated configuration issues and evolving service requirements. Before 5G, studies on communication systems mainly aim to achieve satisfactory data transmission rates and supportive mobility management. In the 5G era, communication systems will gain the abilities to interact with the environment, and the targets are expanded to joint optimizations of ever-increasing numbers of key performance indicators (KPIs), including latency, reliability, connection density, and user experience [11]. Meanwhile, new features such as the dynamic air interface, virtualized network, and network slicing introduce complicated system designs and optimization requirements to address the challenges related to network operation and maintenance. Fortunately, such problems can be considered in the field of artificial intelligence (AI), which provides brand new concepts and possibilities beyond traditional methods. Therefore, AI has recently regained attention in the field of communications in both academia and industry. 3GPP and ITU have both proposed research projects on 5G with AI techniques involved [12, 13].

AI is dedicated to allowing machines and systems to function with intelligence levels similar to that of humans. The field of AI research was born in the 1950s; it experienced advancements and challenges, and has been revived in recent years because of rapid developments in modern computing and data storage technologies. The general problem of simulating intelligence involves sub-problems like reasoning, inference, data fitting, clustering, and optimization, and these sub-problems make use of approaches such as genetic algorithms [14] and artificial neural networks (ANN) [15, 16]. Specifically, AI learning techniques have constructed a universal framework for various problems and have made tremendous progress, resulting in state-of-the-art techniques across various fields.

AI learning tasks are typically classified into two broad categories, supervised and unsupervised learning, depending on the availability of labels of training data for the learning system. Another learning approach, reinforcement learning, is not exactly a supervised learning approach nor an unsupervised learning approach, and so it can be listed as a new category.

- **Supervised learning.** Sample data pairs of inputs and desired outputs are fed into the computer, and the goal is to learn a general function that relates the inputs to the outputs and further detects the unknown outputs of the future inputs. One typical example of supervised learning is illustrated in Figure 1, in which labeled data pairs are fed into a multi-layer deep neural network (DNN) to train the weights between the nodes in the DNN. The training is performed offline, and after convergence, the trained DNN will be ready for recognition and inference of new inputs.
- **Unsupervised learning.** In unsupervised learning, no labels are provided to the learning algorithm and the structure in its input must be determined on its own. Self-organizing map (SOM) is an example of training using unsupervised learning. In SOM, unlabeled data are fed into a neural network to produce a low-dimensional (usually two-dimensional), discretized representation of the input space of the training samples called a map (as illustrated in Figure 2). This method is used for dimensionality reduction.
- **Reinforcement learning.** This technique is based on an alternative interaction between "Agent"

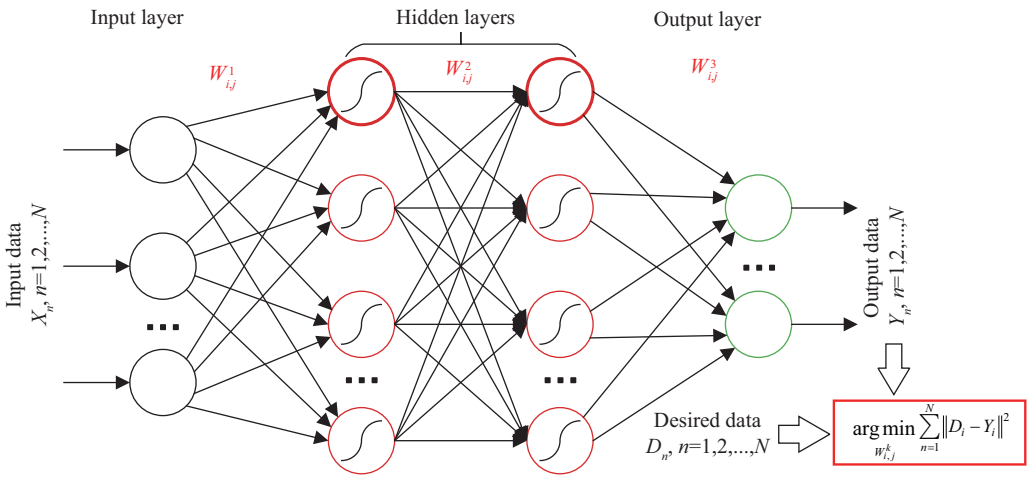

Figure 1 (Color online) Example of supervised learning: learning in deep neural networks.

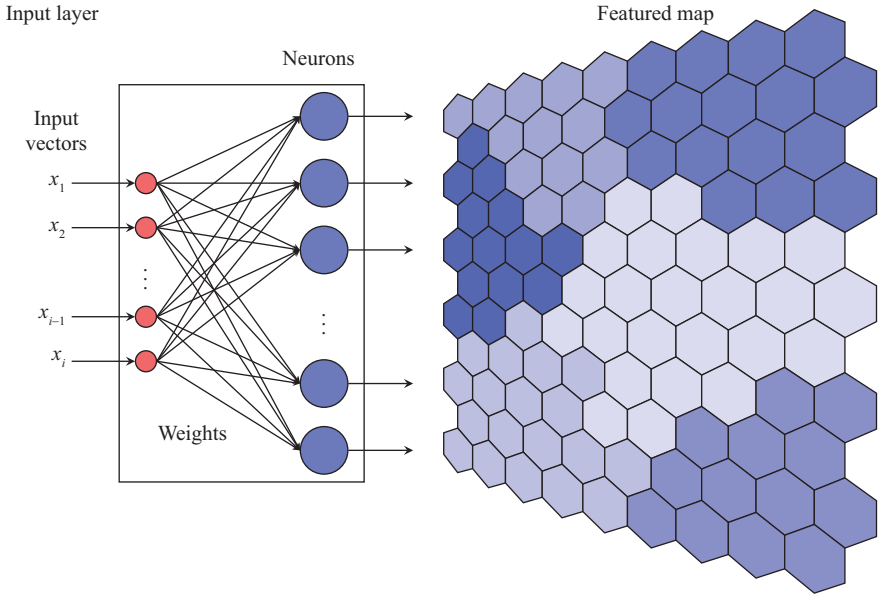

Figure 2 (Color online) Example of unsupervised learning: self organizing map.

and "Environment" and the process is illustrated in Figure 3. The "Agent" will perform certain actions and as a result of this action, the Agent's state will change, leading to either a reward or a penalty. The Agent will then decide the next action based on this result. By iterating the action and reward/penalty process, an Agent learns the Environment.

Popular learning methods in AI learning problems include:

• **Backpropagation (BP).** Backpropagation is a method used in ANNs that are in the category of gradient descent [15]. Backpropagation iteratively calculates gradients of the defined loss function with respect to the weights in the ANN to finally make the output of the ANN close to the known training label. The dynamic learning rate of optimization and acceleration in BP learning are introduced in the author's previous works [16,17]. In [18], the local minima of the BP surface is discussed. Recently, BP has been commonly used to train deep neural networks (DNN), which are neural networks with more than one hidden layer. For example, convolutional neural network (CNN) is a class of feed-forward DNN with multiple hidden layers including a convolutional layer, pooling layer, fully-connected layer, and ReLU layer. CNN can be efficiently trained with the BP method, especially in the fields of image and voice recognition.

• **Q-learning.** The Q-learning algorithm is also referred to as Bellman's algorithm [19], which is a

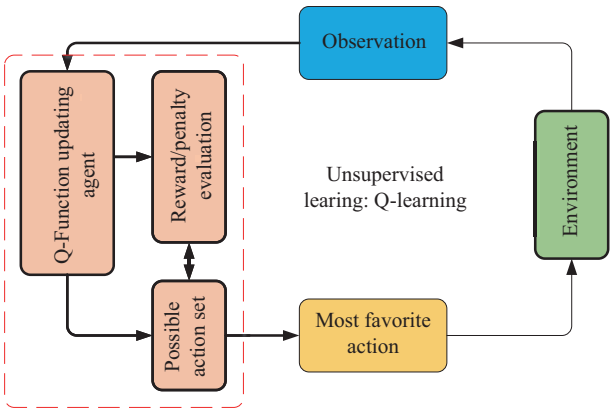

Figure 3 (Color online) Example of reinforcement learning: Q-learning.

classical algorithm for reinforcement learning. In this algorithm, a function (Q function) is defined to evaluate the actions of the "Agent" based on the current "Environment" and outputs the result in the form of an award or penalty. At a certain step, all the possible actions of the "Agent" will be evaluated by the Q function, and the action with the maximum award in the current "Environment" will be selected as the next step and will be actually executed. Bellman proposed a Bellman equation that is a recursive expression that relates the Q functions of consecutive time steps. The Bellman equation basically allows us to iteratively update and approximate the Q function through online temporal difference learning. More details can be found in [20].

AI techniques of machine learning methods that incorporate the common data science algorithms (e.g., linear models, decision tree, k-means clustering) have been implemented for various commercial use. On the other hand, deep learning methods (e.g., DNN, CNN, reinforcement learning) have received increasing attention in the recent decade, resulting in major breakthroughs in fields like cognitive technology. Meanwhile, new branches in deep learning, like meta-learning, are developing rapidly based on the new concept of "learning to learn". For example, in [21], a model-agnostic meta-learning (MAML) method was proposed. This method does not make any assumptions on the form of the model and requires no extra parameters for meta-learning. Therefore, the method can be applied to various fields including classification, regression, and reinforcement learning. More recent developments in AI techniques are summarized in [22–24]. These new technologies extend the possible applications of deep learning to more complicated problems in various scenarios. This helps bring new opportunities into the applications of AI in 5G.

2 Research directions for AI in 5G

As a universal intelligent problem-solving technique, AI can be broadly applied in the design, configuration, and optimization of 5G networks. Specifically, AI is relevant to three main technical problems in 5G:

• **Combinatorial optimization.** One typical example of the combinatorial optimization problem in 5G NR includes network resource allocation. In a resource-limited network, an optimized scheme must be considered for the allocation of resources to different users who share the network such that the utilization of the resource achieves maximum efficiency. As an application of the HetNet architecture in 5G NR with features like network virtualization, network slicing, and self-organizing networks (SON), network resource allocation problems are growing in complexity and require more effective solutions.

• **Detection.** The design of the communication receiver is an example of the detection problem. An optimized receiver can recover the transmitted messages based on the received signals, achieving minimized detection error rate. Detection will be challenging in 5G within the massive MIMO framework.

• **Estimation.** The typical example is the channel estimation problem. 5G requires accurate estima-

tion of the channel state information (CSI) to achieve communications in spatially-correlated channels of massive MIMO. The popular approach includes the training sequence (or pilot sequence), where a known signal is transmitted and the CSI is estimated using the combined knowledge of the transmitted and received signals.

Many studies have been conducted related to the application of AI in 5G such as those in [25–33]. However, because of the limitations in communication systems and AI, some of the applications may be restricted. First, after years of research and testing, conventional methods have demonstrated their abilities to handle the communication systems. A complete framework with conventional techniques has been developed, which is effective, mature, and easy to implement in real-world scenarios. Second, the capacity of a communication system is constrained by certain upper bounds (e.g., the Shannon limit), and some of the well-designed methods can reach near-optimal performance, suffering negligible loss with respect to the capacity bound. For example, in [34] a transmitter optimization method for MIMO was proposed based on an iterative water-filling algorithm, which closely achieves near-Shannon performance in the general jointly correlated MIMO channels. This type of method will not be over-performed by even the most-advanced AI techniques. Moreover, there are still obstacles to the application of AI learning in real-world problems because of the convergence issues involved in training. Careful checks should be performed to ensure that optimal performance can be "learned" with AI in every specific problem in the communication system. Finally, AI algorithms are usually characterized by large computational complexity, which makes them less competitive compared to conventional methods if the performance improvement is minor.

With all these limitations, nonetheless, AI still demonstrates great potential and prospects in communication systems of the 5G era. As introduced above, 5G includes complicated configuration issues and evolving service requirements, which results in new problems that are hard to model, solve, or implement within the current conventional framework. Therefore, new opportunities and challenges are presented by 5G for AI techniques. For all the challenging problems associated with 5G, typical and potential research directions to which AI can make promising contributions must be identified, evaluated, and investigated.

In this paper, we summarize the potential application directions for AI in 5G under four main categories: (1) problems difficult to model; (2) problems difficult to solve; (3) uniform implementation; (4) joint optimization and detection. In the following analysis and examples, we will observe that for problems in (1) and (2), conventional methods are barely effective and AI techniques are expected to be promising; on the other hand, for (3) and (4), the potential of AI is problem-dependent compared to conventional methods, and careful investigation must be performed to identify if AI is beneficial.

• **Problems difficult to model.** Network optimization problems in communication systems are generally a type of technical problems that are hard to model. Problems include typical issues such as network coverage, interference, and neighboring cell selection and handover. Current solutions mostly depend on the experience of engineers. For 5G NR scenarios, these problems are more challenging because of the complicated network structures and large number of KPIs. Applications of new features such as massive MIMO beamforming [35] are associated with high-dimensional optimization parameters and the optimization problem itself can be difficult to model. In addition, 5G NR involves multiple KPIs, including peak data rate, spectral efficiency, latency, connection density, quality of experience (QoE), and so on. These KPIs must be jointly optimized even if some of them contradict each other [11]. In these situations, an overall optimization model cannot be achieved using conventional methods and AI techniques are expected to be able to handle the KPIs.

• **Problems difficult to solve.** Network resource allocation is a key issue in 5G NR [35, 36], which includes specific issues in inter-cell resource block allocation, orthogonal pilot resource allocation, beamforming resource allocation, massive MIMO user clustering, and resource pool deployment in virtualized networks. Network resource allocation aims to maximize the throughput of the network while balancing the service rate. It is mostly an NP-hard combinatorial optimization problem, and the computational complexity to solve this type of problem increases exponentially as the size of the systems increases. Traditional solutions use static partitioning of the network to reduce the computational cost of a sub-optimal solution. With the assistance of current modern computing technologies, AI will be a new

effective solution to these problems.

• **Uniform implementation.** Conventional methods are designed in a divide-and-conquer manner for some function blocks in 5G NR. For example, the physical layer in 5G NR consists of a series of signal processing blocks such as multiuser MIMO space-time processing, NOMA signal detection and encoding, and decoding for LDPC or polar codes. Researchers have attempted to optimize the algorithms and implementations of each processing module and achieved practical success. However, the efficient and scalable implementation of the entire communication system with guaranteed performance is still lacking. It is noted that AI techniques are supposed to be capable of handling each of the modules [25–33]. This inspires us to further develop a uniform AI-based implementation that works jointly for all the key modules in the 5G NR physical layer [37]. By unifying the modules with AI methods in both algorithm and hardware, the design, configuration, and implementation of the physical layer communications will be simpler, faster, more economical, and more efficient.

• **Joint optimization and detection.** An intuitive idea for applying AI in 5G is to simply substitute the conventional modules of transmitters and receivers by ANN. However, the capacity of the channels is bounded by the Shannon limit and improvements through the use of ANN are limited. In addition, as discussed above, the complexity and the convergence of training should be carefully examined in this area. Compared to this intuitive method, AI demonstrates more potential in a bigger picture of the cross-layer joint optimization problem that cannot be solved efficiently using conventional methods [38]. Typical examples include the joint optimizations for the physical and media access control layers [39], joint source and channel optimizations [40], and the joint optimization for algorithm and hardware implementation [41].

3 Paradigms of AI in 5G

In this section, examples of the application of AI techniques in 5G are presented, which cover four different problems in 5G including: network resource allocation, SON, uniform 5G accelerator, and the optimization of end-to-end physical layer communication.

3.1 AI for SON: automatic root cause analysis

Self-organizing networks (SONs) establish a new concept of network management that provide intelligence in the operation and maintenance of the network. SON has been introduced by 3GPP as a key component of LTE networks. In the 5G era, network densification and dynamic resource allocation will result in new problems for coordination, configuration, and management of the network, thereby resulting in increased demand for improvements in the SON functions. SON modules in mobile networks can be divided into three main categories: self-configuration, self-optimization, and self-healing. The main objectives of SON involve automatically performing network planning, configuration, and optimization without human intervention to reduce the overall complexity, operational expenditure (OPEX), capital expenditure (CAPEX), and man-made faults. Various studies on AI in SON have been summarized in [42–44]. The studies include those on AI applied in automatic base station configuration, new cell and spectrum deployment, coverage and capacity optimization, cell outage detection and compensation, etc., using approaches such as ANN, ant colony optimization, and genetic algorithm.

In this section, we introduce the automatic root cause analysis framework proposed in [45] as an example of AI in SON. The design of the fault identification system in LTE networks faces two main challenges: (1) A substantial number of alarms, KPIs, and configuration parameters can be considered as fault indicators in the system. Meanwhile, most of the symptoms of these indicators are not labeled with fault causes, and are difficult to identify. (2) The system is not automatic and experts are involved in the analysis of each fault cause. With the large amount of high-dimensional data, human intervention is not efficient but expensive. Authors of [45] proposed an AI-based automatic root cause analysis system that combines supervised and unsupervised learning techniques as summarized in the following steps:

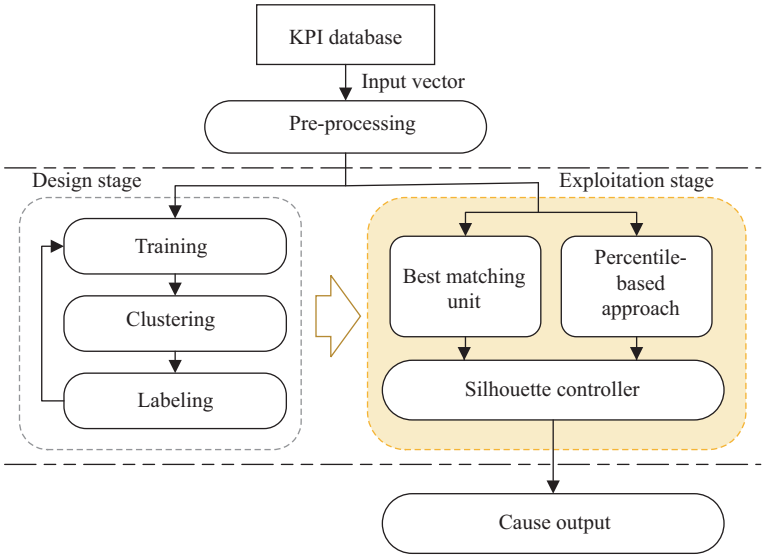

Figure 4 (Color online) Automatic root cause analysis workflow [45].

• **Step 1.** Unsupervised SOM training. SOM as shown in Figure 2 is applied for an initial classification of the high-dimensional KPIs. An SOM is a type of unsupervised neural network capable of acquiring knowledge and learning from a set of unlabeled data. It will process high-dimensional data and reduce the data to a two-dimensional map of neurons that preserves the topological properties of the input data. Therefore, inputs close to each other will be mapped to adjacent neurons. Through this unsupervised process, the high-dimensional KPIs are mapped into a lower-dimensional map that can classify new KPI data by finding the closest neurons.

• **Step 2.** Unsupervised clustering. After SOM training, all the neurons in the SOM system will be clustered into a certain number of groups using an unsupervised algorithm. Because the SOM neurons are already ordered and the difference between the original inputs can be represented by the Euclidean distance between the corresponding neurons, clustering algorithms based on Euclidean distance, e.g., Ward's hierarchical method, can be applied for clustering neurons.

• **Step 3.** Labeling by Experts. After the above two steps, the original high-dimensional data are clustered into several classes. We will finally include experts who will analyze and identify the fault causes of each obtained cluster to ensure that all clusters are labeled.

With the training, clustering, and labeling, an automatic system for network diagnosis is constructed by the workflow shown in Figure 4. A new input of KPIs will first be mapped to a neuron in SOM. Using the label of the cluster to which the neuron belongs to, we can identify the fault and the causes. After obtaining a certain amount of new fault data, we can verify whether the system is right or not and update the system by re-training using the above three steps. Simulation results presented in [45] demonstrate that the proposed root cause analysis system is highly accurate even though it is primarily built using unsupervised techniques.

3.2 AI for resource allocation: OFDMA downlink resource allocation

The OFDM resource block (RB) allocation in 5G NR is more complicated and challenging than ever before because of the support for the three aforementioned generic services. In Figure 5, a typical multi-cell, multi-user downlink resource allocation scenario is illustrated. In this system, the intra-cell interference is eliminated because the RB allocated to different users in the same cell is orthogonal to each other. System interference mainly depends on the inter-cell interference, which makes the RB allocation for users in neighboring cells important. Suppose the throughput of each user can be evaluated based on signal-to-interference ratio (SIR), the target for the optimization of the RB allocation is the maximization of the total system throughput. This is indeed an NP-hard combinatorial optimization problem with nonlinear

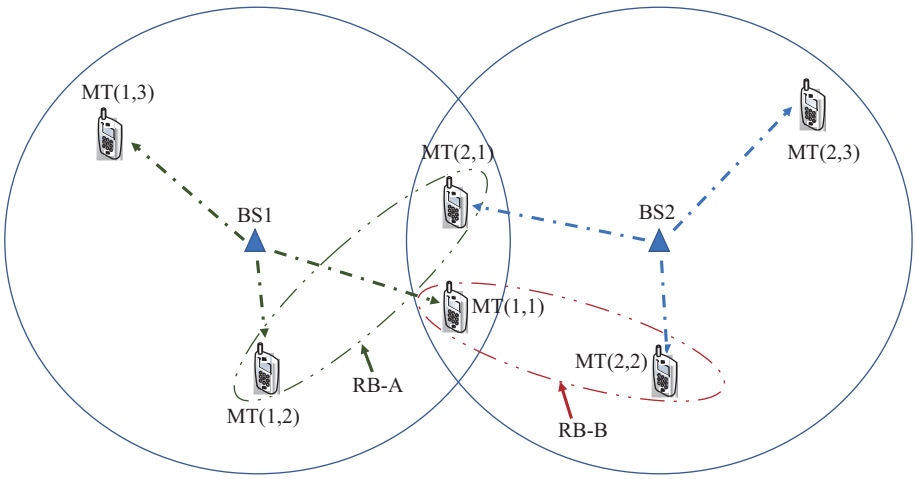

Figure 5 (Color online) Dynamic resource allocation for multi-cell and multi-user systems.

constraints. The complexity of the traditional solution is proportional to the factorial of the number of users in coverage, which is computationally prohibitive.

Q-learning can be applied to this problem. Suppose an "Agent" is in charge of the RB allocation, then the possible "Action" of this "Agent," which is to update the RB for each user, can be selected by the following strategies: (1) Within the same cell, allocate free RB with higher SIR to the users; (2) update the RB allocated to the user with the worst SIR in the current cell continuously to achieve better overall system capacity; (3) for a certain RB, pair or cluster the user with the worst SIR in the current cell with users with the best SIR from the neighboring cell. The first two strategies are intuitive. The third one is applied to avoid allocating the same RB to users in neighboring cells that are located close to the boarder because in such situations, the involved users cannot acquire essential SIR to work appropriately regardless of the transmitting power of the base stations.

After defining all the possible "Actions", the "Agent" will evaluate each of them to select the next "Action" for adjusting the RB allocation that maximizes the overall capacity of the entire system. The Q function is updated according to the Bellman equation [19] at the same time. We iterate through this process until the Q function converges.

These iterations should also be considered with the optimization of the user power allocation. In [46], a framework based on the generalized Nash equilibrium problem (GNEP) is proposed for the optimization of the power control for users from multiple cells that are assigned the same RB. Global quality of service (QoS) constraints are also considered and so Lagrange multipliers are introduced to evaluate the "Actions" and establish the Q function. Refs. [47,48] summarizes the possible applications of AI techniques for 5G resource allocations. In [49], a reinforcement learning-based method is proposed for the new network slicing function in the 5G framework.

3.3 AI for baseband signal processing: Uniform 5G accelerator

The baseband signal processing in 5G consists of a series of signal processing blocks, including massive MIMO detection, NOMA detection, and decoding for polar codes. The increased number of baseband blocks results in more hardware area and varied implementation structures. However, we notice that the belief propagation algorithm based on factor graphs can be applied to all the blocks as proposed in [50–54]. For each specific block, the frameworks are kept unchanged and we only have to adapt the symbol set and constraints of the variables to the certain function. Therefore, a uniform accelerator for the baseband can be designed based on the belief propagation algorithms with configurable variables.

However, the performance of belief propagation is limited in some baseband blocks in certain scenarios. Here, AI can be a possible solution to such problems. By improving the belief propagation methods with the AI techniques, an AI-based uniform accelerator can be constructed. The AI-aided belief propagation algorithms can be designed with the following two methods:

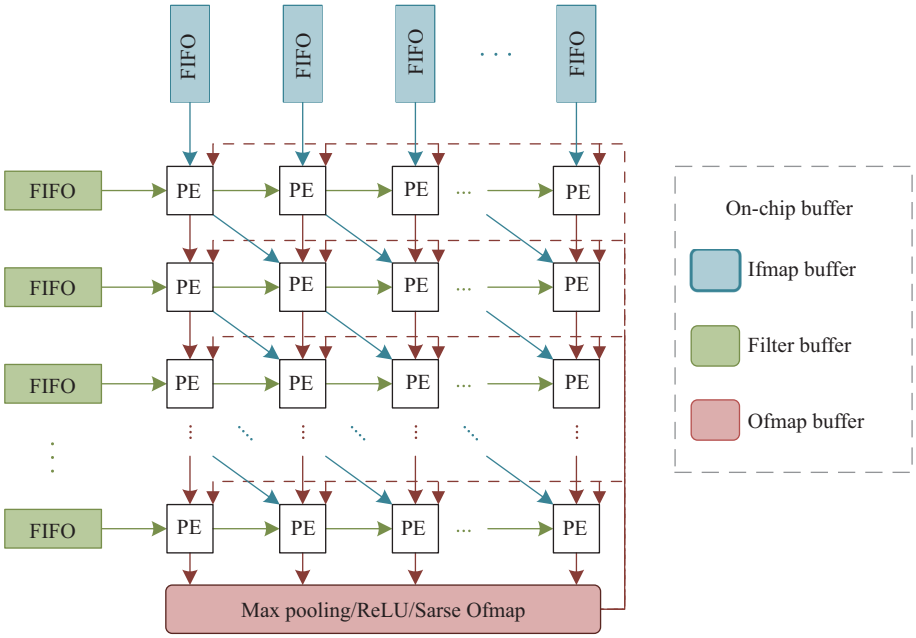

Figure 6 (Color online) Systolic array hardware structure for neural networks [28].

• **DNN-aided belief propagation.** (1) Unfold the factor graph of belief propagation by duplicating the iterations to form a DNN; (2) Train the DNN by supervised training. Applications of this method in the baseband include the DNN-based polar codes decoder proposed in [30] and the DNN-aided MMO detector proposed in [55].

• **Belief propagation-based CNN.** (1) Map each node in the factor graph of belief propagation to one pixel in a picture, in which connected nodes should be mapped into neighboring pixels; (2) Train the CNN using the obtained pictures. This method is utilized in the BP-CNN channel decoder proposed in [56].

The neural networks are highly self-adaptive and reliable. By applying DNN and CNN in the baseband, we can achieve performance enhancements as long as a uniform hardware implementation framework exists. Actually, the core operation of CNN is the convolution, while the core operation of DNN is the multiplication of the two-dimensional matrices. We notice that the systolic architecture can realize both operations. Figure 6 illustrates a reconfigurable systolic architecture designed for accelerating convolutional neural network [28]. It can be seen that the systolic architecture is regular and scalable, which supports different CNNs and DNNs. This motivates us to explore the possibilities of reusing the same hardware architecture to realize both 5G and DL algorithms.

The authors of [31] indicate that in a system formed with the channel encoder, the channel, the channel equalizer, and the decoder (as shown in Figure 7), the equalizer and the decoder can be implemented with a CNN and a DNN, respectively. The associated AI accelerator can be jointly realized by two strategies: (1) The uniform architecture. The overall receiver can be folded into one uniform processor to save the hardware area. This processor first works as a CNN-based equalizer with the input signals from the channel. The output of the CNN will be saved at this point. The processor will then function as a DNN-based decoder, for which the saved output from the CNN will serve as the input. The decoding results will be the final output. (2) The cascade architecture. Two processors will be cascaded directly to construct the receiver, one being the CNN-based equalizer while the other being the DNN-based decoder. This architecture has more hardware components but achieves a higher throughput rate.

Overall, the AI-based uniform accelerator is more flexible for the hardware implementation, and can therefore achieve various system requirements.

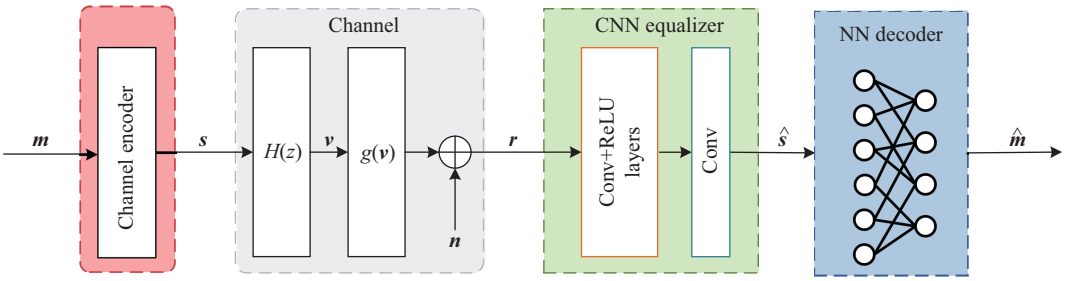

Figure 7 (Color online) Architecture of a receiver including neural network equalizer and decoder [31].

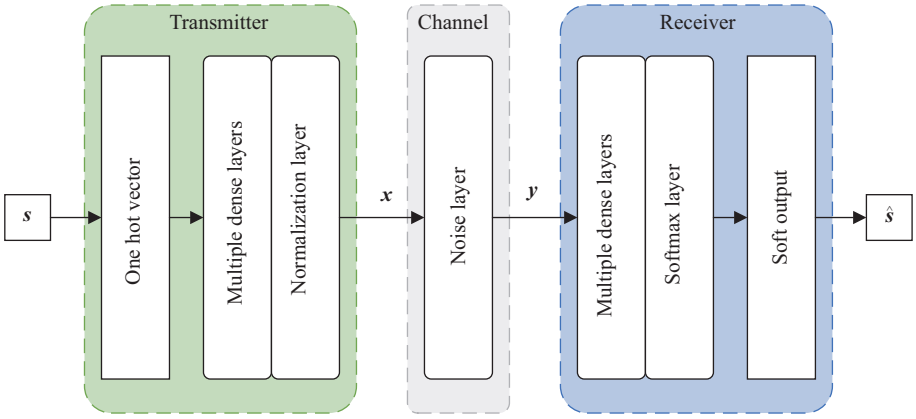

Figure 8 (Color online) A simple autoencoder for an end-to-end communication system [58].

3.4 AI for physical layer: DNN-based end-to-end communication

As mentioned above, AI, especially DNN, has been applied to different function blocks in the physical layer, e.g., modulation recognition [57], polar codes decoder [30], and MIMO detection [56]. For the joint optimization of two or more blocks, AI algorithms also achieve success, e.g., the aforementioned joint optimization of channel equalizer and channel decoder proposed in [31]. Refs. [58, 59] both provide a comprehensive summary of AI in the physical layer. However, optimizing each of the blocks individually does not guarantee the optimization for the entire physical layer communication [58]. From the viewpoint of the entire end-to-end communication system, the intuitive connection of different AI modules may result in extra computational cost for both training and online tasks. Therefore, a joint optimization method for the end-to-end system is needed.

Ref. [59] proposes an autoencoder-based end-to-end system optimization method, in which the communication system is recast as an end-to-end reconstruction optimization problem, and a novel concept, the autoencoder, is introduced to serve as a simplified representation of the system. The autoencoder is a type of ANN. It aims at learning a representation (encoding) for a set of data in an unsupervised manner, which will be capable of reconstructing compressed inputs at the output layer. In the proposed approach in [59], the end-to-end system is simply represented by three blocks, the transmitter, the receiver, and the channel. The transmitter and the receiver are both represented as fully connected DNNs. A normalization layer is connected to the transmitter to guarantee the energy constraints, whereas a softmax activation layer is placed before the receiver to output soft decisions for the received information. The AWGN channel in between is represented as a simple noise layer with a certain variance. The resulting autoencoder has a structure as shown in Figure 8. This autoencoder is trained based on bit error rate (BER) or block error rate (BLER). After training, the autoencoder will be able to reconstruct the transmitted signals based on the received signals.

The autoencoder is a novel concept that is different from all traditional, conventional methods. DNNs are utilized to represent the entire end-to-end system without considering the specific models in each traditional function block. Therefore, in scenarios that are too complicated to model, The autoencoder

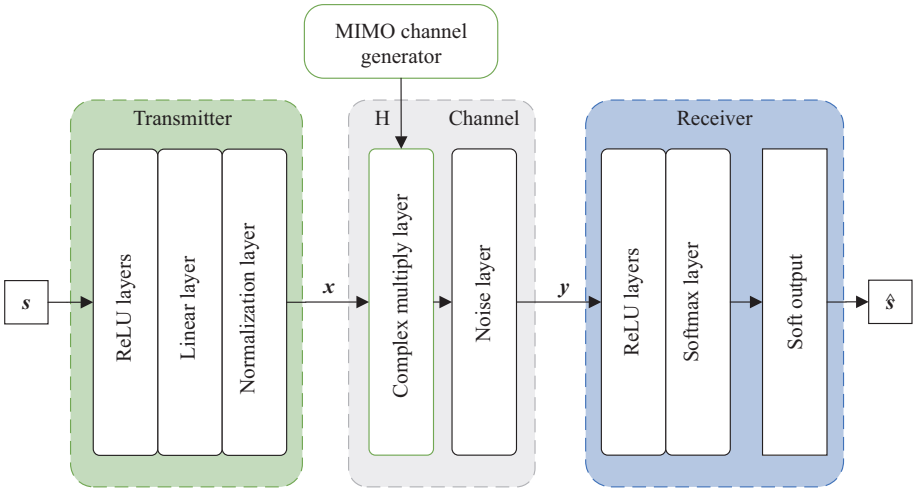

Figure 9 (Color online) A general MIMO channel autoencoder architecture [58].

can be an appropriate solution to "learn" these scenarios and optimize the performance. This scheme is extended to a multi-user scenario with interfering channels in [59] and is further extended to MIMO in [60] by adding a module for the channel matrix as shown in Figure 9. Simulation results presented in [59, 60] illustrate that the autoencoder can "learn" different scenarios with various CSI and numerous antennas and achieves enhanced BER performance.

4 Conclusion

5G promises significant breakthroughs in traditional mobile communication systems. While enhancing the service capability of traditional mobile networks, it further evolves to support the applications of IoT in various fields including business, manufacturing, health care, and transportation. Therefore, 5G will serve as the basic technology for future IoT technologies that connect and operate entire societies. Aiming to support differentiated applications with a uniform technical framework, 5G is facing enormous challenges. With the revival and rapid developments in recent years, AI is rising to these challenges. AI is a potential solution to the problems associated with the 5G era, which will lead to revolutionary concepts and capabilities in communication systems.

Many studies have already been conducted for applying AI in 5G. In this paper, instead of reviewing all existing literature, we focus on clarifying promising research directions with the greatest potential. Through additional efforts in these research directions, 5G is anticipated to achieve significantly better performance and more convenient implementations compared to traditional communication systems. With the inspiring research paradigms introduced in this paper, we look forward to the remarkable applications of AI in 5G in the near future.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61501116, 61521061).

References

1 You X H, Pan Z W, Gao X Q, et al. The 5G mobile communication: the development trends and its emerging key techniques (in Chinese). Sci Sin Inform, 2014, 44: 551–563
2 Li L M, Wang D M, Niu X K, et al. mmWave communications for 5G: implementation challenges and advances. Sci China Inf Sci, 2018, 61: 021301
3 Wang C X, Wu S B, Bai L, et al. Recent advances and future challenges for massive MIMO channel measurements and models. Sci China Inf Sci, 2016, 59: 021301
4 Zhang J H, Tang P, Tian L, et al. 6–100 GHz research progress and challenges from a channel perspective for fifth generation (5G) and future wireless communication. Sci China Inf Sci, 2017, 60: 080301

5 Tao X F, Han Y, Xu X D, et al. Recent advances and future challenges for mobile network virtualization. Sci China Inf Sci, 2017, 60: 040301
6 3GPP. Way forward on the overall 5G-NR eMBB. Workplan RP-170741. 2017. ftp://ftp.3gpp.org/TSG RAN/TSG RAN/TSGR 75/Docs/RP-170741.zip
7 3GPP. Study on new radio access technology: radio access architecture and interfaces (release 14). TR38.801, v14.0. 2017. http://www.3gpp.org/ftp/Specs/archive/38 series/38.801/38801-e00.zip
8 ITU-R. Minimum requirements related to technical performance for IMT2020 radio interface(s). Report ITU-R M.2410-0. 2017. https://www.itu.int/pub/R-REP-M.2410-2017
9 3GPP. LTE Enhancements and 5G Normative Work. Release-15. 2018. http://www.3gpp.org/release-15
10 You X H, Wang D M, Sheng B, et al. Cooperative distributed antenna systems for mobile communications. IEEE Wirel Commun, 2010, 17: 35–43
11 Yang W J, Wang M, Zhang J J, et al. Narrowband wireless access for low-power massive internet of things: a bandwidth perspective. IEEE Wirel Commun, 2017, 24: 138–145
12 ITU-T. LS/o on the results of the 1st meeting of the ITU-T focus group on ma- chine learning for future networks including 5G (FG ML5G). FG ML5G-0-004. 2018. http://www.3gpp.org/ftp/tsg sa/WG1 Serv/TSGS1 82 Dubrovnik/Docs/S1-181271.zip
13 3GPP. 5G system network data analytics services stage 3. TS 29.520 (CT3). 2018. http://www.etsi.org/deliver/etsi ts/129500 129599/129520/15.00.00 60/ts 129520v150000p.pdf
14 Whitley D. A genetic algorithm tutorial. Stat Comput, 1994, 4: 65–85
15 Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw, 2015, 61: 85–117
16 You X H, Chen G A, Cheng S X. Dynamic learning rate optimization of the backpropagation algorithm. IEEE Trans Neural Netw, 1995, 6: 669–677
17 You X H. Can backpropagation error surface not have local minima. IEEE Trans Neural Netw, 1992, 3: 1019–1021
18 Yu X H, Chen G A. Efficient backpropagation learning using optimal learning rate and momentum. Neural Netw, 1997, 10: 517–527
19 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey. J Artif Intell Res, 1996, 4: 237–285
20 Watkins C J C H, Dayan P. Q-learning. Mach Learn, 1992, 8: 279–292
21 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. 2017. ArXiv: 1703.03400
22 Wu J X, Gao B B, Wei X S, et al. Resource-constrained deep learning: challenges and practices. Sci Sin Inform, 2018, 48: 501–510
23 Zhou Z H. Machine learning: recent progress in China and beyond. China Sci Rev, 2018, 5: 20
24 Zhong Y X. Artificial intelligence: concept, approach and opportunity. Chin Sci Bull, 2017, 62: 2473
25 Gatherer A. Machine learning modems: how ML will change how we specify and design next generation communication systems. IEEE ComSoc Tech News, 2018. https://www.comsoc.org/ctn/machine-learning-modems-how-ml-will-change-how-we-specify-and-design-next-generation
26 Yang C, Xu W H, Zhang Z C, et al. A channel-blind detection for SCMA based on image processing techniques. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2018. 1–5
27 Zhang C, Xu W H. Neural networks: efficient implementations and applications. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 1029–1032
28 Xu W H, You X H, Zhang C. Efficient deep convolutional neural networks accelerator without multiplication and retraining. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018. 1–5
29 Xu W H, Wang Z F, You X H, et al. Efficient fast convolution architectures for convolutional neural network. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 904–907
30 Xu W H, Wu Z Z, Ueng Y L, et al. Improved polar decoder based on deep learning. In: Proceedings of IEEE International Workshop on Signal Processing Systems (SiPS), 2017. 1–6
31 Xu W H, Zhong Z W, Be'ery Y, et al. Joint neural network equalizer and decoder. In: Proceedings of IEEE International Symposium on Wireless Communication Systems (ISWCS), 2018. 1–6
32 Xu W H, Be'ery Y, You X H, et al. Polar decoding on sparse graphs with deep learning. In: Proceedings of Asilomar Conference on Signals, Systems, and Computers (Asilomar), 2018. 1–6
33 Xu W H, You X H, Zhang C. Using Fermat number transform to accelerate convolutional neural network. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 1033–1036
34 Gao X Q, Jiang B, Li X, et al. Statistical eigenmode transmission over jointly correlated MIMO channels. IEEE Trans Inform Theor, 2009, 55: 3735–3750
35 Wang D M, Zhang Y, Wei H, et al. An overview of transmission theory and techniques of large-scale antenna systems for 5G wireless communications. Sci China Inf Sci, 2016, 59: 081301
36 Gesbert D, Hanly S, Huang H, et al. Multi-cell MIMO cooperative networks: a new look at interference. IEEE J Sel Areas Commun, 2010, 28: 1380–1408
37 Jing S S, Yu A L, Liang X, et al. Uniform belief propagation processor for massive MIMO detection and GF (2^n) LDPC decoding. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2017. 961–964
38 Gandhi V S, Maheswaran B. A cross layer design for performance enhancements in LTE-A system. In: Proceedings of IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016. 905–909

39 Kuen J, Kong X F, Wang G, et al. DelugeNets: deep networks with efficient and flexible cross-layer information inflows. In: Proceedings of IEEE International Conference on Computer Vision Workshop (ICCVW), 2017. 958–966
40 Farsad N, Rao M, Goldsmith A. Deep learning for joint source-channel coding of text. 2018. ArXiv: 1802.06832
41 Xu X W, Ding Y K, Hu S X, et al. Scaling for edge inference of deep neural networks. Nat Electron, 2018, 1: 216–222
42 Wang X F, Li X H, Leung V C M. Artificial intelligence-based techniques for emerging heterogeneous network: state of the arts, opportunities, and challenges. IEEE Access, 2015, 3: 1379–1391
43 Klaine P V, Imran M A, Onireti O, et al. A survey of machine learning techniques applied to self-organizing cellular networks. IEEE Commun Surv Tut, 2017, 19: 2392–2431
44 Pèrez-Romero J, Sallent O, Ferrús R, et al. Knowledge-based 5G radio access network planning and optimization. In: Proceedings of IEEE International Symposium on Wireless Communication Systems (ISWCS), 2016. 359–365
45 Gómez-Andrades A, Munoz P, Serrano I, et al. Automatic root cause analysis for LTE networks based on unsupervised techniques. IEEE Trans Veh Technol, 2016, 65: 2369–2386
46 Wang J H, Guan W, Huang Y M, et al. Distributed optimization of hierarchical small cell networks: a GNEP framework. IEEE J Sel Areas Commun, 2017, 35: 249–264
47 Bogale T E, Wang X, Le L B. Machine intelligence techniques for next-generation context-aware wireless networks. 2018. ArXiv: 1801.04223
48 Li R, Zhao Z, Zhou X, et al. Intelligent 5G: when cellular networks meet artificial intelligence. IEEE Wirel Commun, 2017, 24: 175–183
49 Zhao Z, Li R, Sun Q, et al. Deep reinforcement learning for network slicing. 2018. ArXiv: 1805.06591
50 Ren Y R, Zhang C, Liu X, et al. Efficient early termination schemes for belief-propagation decoding of polar codes. In: Proceedings of IEEE International Conference on ASIC (ASICON), 2015. 1–4
51 Fossorier M P C, Mihaljevic M, Imai H. Reduced complexity iterative decoding of low-density parity check codes based on belief propagation. IEEE Trans Commun, 1999, 47: 673–680
52 Yang J M, Song W Q, Zhang S Q, et al. Low-complexity belief propagation detection for correlated large-scale MIMO systems. J Sign Process Syst, 2018, 90: 585–599
53 Liu L, Yuen C, Guan Y L, et al. Gaussian message passing iterative detection for MIMO-NOMA systems with massive access. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), 2016. 1–6
54 Yang J M, Zhang C, Zhou H Y, et al. Pipelined belief propagation polar decoders. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2016. 413–416
55 Tan X S, Xu W H, Be'ery Y, et al. Improving massive MIMO belief propagation detector with deep neural network. 2018. ArXiv: 1804.01002
56 Liang F, Shen C, Wu F. An iterative BP-CNN architecture for channel decoding. IEEE J Sel Top Signal Process, 2018, 12: 144–159
57 Lv X Z, Wei P, Xiao X C. Automatic identification of digital modulation signals using high order cumulants. Electronic Warfare, 2004, 6: 1
58 Wang T Q, Wen C K, Wang H Q, et al. Deep learning for wireless physical layer: opportunities and challenges. China Commun, 2017, 14: 92–111
59 O'Shea T, Hoydis J. An introduction to deep learning for the physical layer. IEEE Trans Cogn Commun Netw, 2017, 3: 563–575
60 O'Shea T J, Erpek T, Clancy T C. Deep learning based MIMO communications. 2017. ArXiv: 1707.07980