Intelligent 5G: When Cellular Networks Meet Artificial Intelligence

Rongpeng Li, Zhifeng Zhao, Xuan Zhou, Guoru Ding, Yan Chen, Zhongyao Wang, and Honggang Zhang

Abstract

5G cellular networks are assumed to be the key enabler and infrastructure provider in the ICT industry, by offering a variety of services with diverse requirements. The standardization of 5G cellular networks is being expedited, which also implies more of the candidate technologies will be adopted. Therefore, it is worthwhile to provide insight into the candidate techniques as a whole and examine the design philosophy behind them. In this article, we try to highlight one of the most fundamental features among the revolutionary techniques in the 5G era, i.e., there emerges initial intelligence in nearly every important aspect of cellular networks, including radio resource management, mobility management, service provisioning management, and so on. However, faced with ever-increasingly complicated configuration issues and blossoming new service requirements, it is still insufficient for 5G cellular networks if it lacks complete AI functionalities. Hence, we further introduce fundamental concepts in AI and discuss the relationship between AI and the candidate techniques in 5G cellular networks. Specifically, we highlight the opportunities and challenges to exploit AI to achieve intelligent 5G networks, and demonstrate the effectiveness of AI to manage and orchestrate cellular network resources. We envision that AI-empowered 5G cellular networks will make the acclaimed ICT enabler a reality.

Introduction

Currently, fourth-generation (4G) cellular networks are being globally deployed to provide all-IP (Internet Protocol) broadband connectivity. Recalling that second-generation (2G) global networks for mobile communications (GSM), debuted in 1991, just started to provide digital voice telephony, and third-generation (3G) cellular networks, launched in 2001, initially provided mobile Internet solutions. It took less than 30 years to successfully transform cellular networks from pure telephony systems to networks that can transport rich multimedia content [1] and have a profound impact on our daily life. Nowadays, the landscape of the information communication technology (ICT) industry is rapidly changing. First, mobile broadband access is expected to have a drastic increase with 1000 times more aggregate throughput [2] and 10 times more at the link level [3] from 2010 to 2020. Second, an increasing number of objects are being digitalized to form the Internet of Things, posing more stringent requirements on latency, battery lifetime, etc. [4]. Therefore, to enhance service provisioning and satisfy the coming diversified requirements, it is necessary to revolutionize the cellular networks with cutting-edge technologies. The standardization of next-generation (5G) cellular networks is being expedited, which also implies more of the candidate technologies will be adopted. This naturally raises questions such as which new technologies might 5G cellular networks possess, and which features will these technologies have in common?

From the very beginning, 5G cellular networks were assumed to be the key enabler and infrastructure provider in the ICT industry, by offering three types of services from enhanced mobile broadband (eMBB) with bandwidth-consuming and throughput-driving requirements to new services such as ultra-reliable low latency service (URLLC) and massive machine-type communications (mMTC). In that regard, though technologies such as densified cells and massive multiple-input multiple-output (MIMO) are essential to boost capacity in the 5G era, it is cost-ineffective to deploy such techniques. Instead, 5G cellular networks mainly revolutionize themselves by initially embracing the intelligence to agilely boost both spectrum efficiency (SE) and energy efficiency (EE). Specifically, 5G cellular networks provide alternative options for radio resource management (RRM), mobility management (MM), management and orchestration (MANO), and service provisioning management (SPM) mechanisms. Hence, it is no longer necessary to build dedicated networks for individual services (e.g., the GSM-Railway communication networks). On the contrary, as depicted in Fig. 1, due to the development of smarter 5G networks, it will be feasible to provide customized end-to-end network slices (NS) [5] to simultaneously satisfy distinct service requirements, such as ultra-low latency in URLLC and ultra-high throughput in eMBB.

There is no doubt that 5G cellular networks will tailor the provisioning mechanisms for different predefined services and pave the way for the application of complete intelligence. However,

Rongpeng Li, Zhifeng Zhao (corresponding author), and Honggang Zhang are with Zhejiang University; Xuan Zhou and Yan Chen are with Huawei Technologies Co., Ltd.; Guoru Ding is with PLA University of Science and Technology; Zhongyao Wang is with the Alibaba Group.

Digital Object Identifier:
10.1109/MWC.2017.1600304WC

1536-1284/17/$25.00 © 2017 IEEE

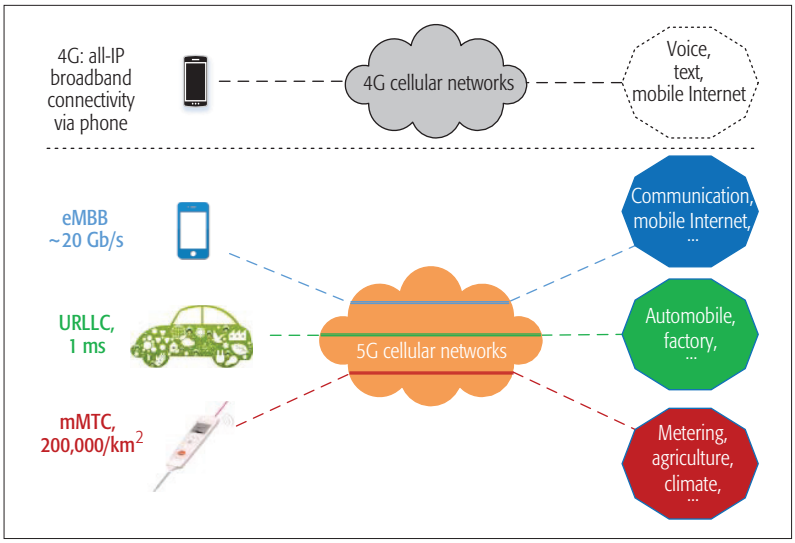

FIGURE 1. 5G cellular networks: a key enabler to all mobile devices across all industries.

it is still challenging and time-consuming for 5G cellular network operators to solve ever-increasingly complicated configuration issues and satisfy evolving service requirements, since 5G cellular networks merely possess more technical options to deal with predefined intelligent problems, rather than gain the ability to interact with the environment (e.g., traffic load, service characteristics). Fortunately, such an interaction falls into the field of artificial intelligence (AI), which is dedicated to empowering machines and systems with intelligence similar to that of humans. Hence, it is promising to apply AI to 5G cellular networks to deal with newly emerging issues.

In this article, we will try to answer what key technical progress is in 5G cellular networks, why it is crucial to embrace AI in the 5G era, and how AI can contribute to management and orchestration in the 5G era.

The Intelligence in 5G Cellular Networks

AI is the science and engineering of making machines as intelligent as humans, and has long been applied to optimize communication networks in diverse configurations [6]. According to the extent of intelligence, AI could be divided into two levels. The first and basic level of AI is that one machine or entity can provide multiple pre-defined options and respond to the environment in a different yet deterministic manner. For example, as discussed later, 5G will allow granted and grant-free transmission for eMBB and mMTC services, respectively. In other words, the network will intelligently adjust the configuration after detecting different pre-defined service indicators. The second and complete level of AI is that one machine or entity possesses full capability to interact (e.g., sense, mine, predict, and reason) with the environment. More importantly, the machine or entity is able to learn how to make appropriate responses, even when it faces strange scenarios or tasks. In this section, we will highlight how the candidate technologies grant preliminary intelligence (i.e., the basic level of AI) to cellular networks, and transform cellular networks from being network-centric to being user-centric and information-centric with significant SE and EE improvement.

Radio Resource Management

Current 4G cellular networks heavily rely on orthogonal frequency-division multiplexing (OFDM) as the signal bearer and the base of associated access schemes. Since OFDM can be used in both frequency-division duplex (FDD) and time-division duplex (TDD) formats, FDD and TDD 4G cellular networks share a similar frame structure by grouping a static number of subcarriers and symbols into one resource block (RB). Benefiting from the satisfactory subcarrier orthogonality in OFDM, information transmitted in different RBs can be separately decoded at the receivers with limited computational cost. However, it is stubborn to use OFDM to simultaneously satisfy service requirements from different users with various channel conditions, user terminal (UE) capabilities (multiple access support, full duplex mode, feature or smart phones), mobility, frequency bands, and so on. Given that, 5G cellular networks aim to introduce new waveforms and provide softer air interfaces. Specifically, filter-bank multi-carrier (FBMC) and unified-filter multi-carrier (UFMC) are famous candidates for more flexible frame structures and waveforms in the 5G era. As their names imply, FBMC and UFMC both add filters to combat out-of-band leakage across subcarriers and make it unnecessary to strictly synchronize across RBs. Therefore, 5G cellular networks can provide different air interface solutions in different RBs, in which different multiple access schemes, TTI (transmission time interval) parameters, waveforms, and duplex mode, pilot signals, etc., can be well defined [7]. For example, as seen in Fig. 2a, larger bandwidth and symbol length can be applied to eMBB to yield a higher rate, while smaller TTI can be configured for URLLC to shorten response latency.

Similar to the evolution from OFDM to FBMC/UFMC, 5G cellular networks potentially adopt non-orthogonal multiple access (NoMA) schemes such as sparse coding multiple access (SCMA). Such NoMA schemes overlap information from two transmitters in the same radio resource and apply successive interference cancellation (SIC) receivers (or even more computationally-exhaustive maximum-likelihood receivers) to decode the received information. Apparently, NoMA could potentially lead to higher throughput. Moreover, another advantage of NoMA is that it makes possible grant-free transmission in the uplink (UL), if the UE identity and the preamble for grant-free UL transmission are mapped together. Instead of waiting for resource allocation commands as in 4G cellular networks, it is feasible to decode the overlapped information from two UEs at the same resources by using SIC receivers. From Fig. 2b, in spite of the reliability advantage for granted transmission, grant-free transmission in UL could avoid the cumbersome signaling procedures and save latency for small packets at a trivial performance loss. Comparatively, 5G cellular networks eventually have one alternative option, which is quite suitable for mMTC service.

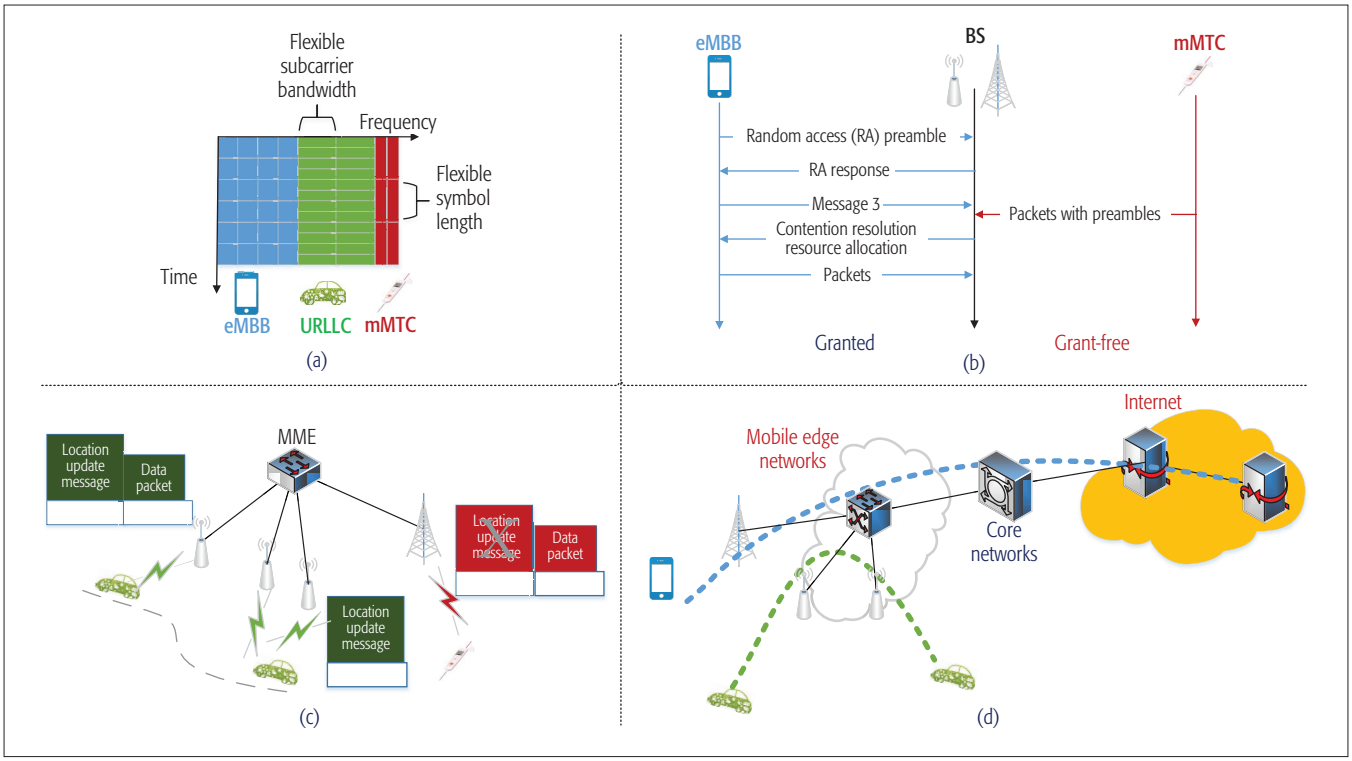

FIGURE 2. Candidate technologies for intelligent cellular networks: a) flexible bandwidth and symbol length enabled by FBMC and UFMC; b) granted and grant-free transmission enabled by NoMA; c) flexible mobility management schemes; d) dynamic service provisioning stack.

Mobility Management

In 4G cellular networks, there exist two states to manage UEs' location awareness to the core network (e.g., evolved packet system (EPS)). All EPS mobility management (EMM)-connected UEs should periodically report their locations, so as to guarantee the session continuum and information reachability. Definitively, it is resource-consuming to treat all UEs the same. Instead, some UEs at static positions (e.g., UEs for mMTC metering services) should only need to report at the very beginning of network attachment. Taking account of practical considerations, 5G cellular networks introduce multiple-tier mobility management states to make the mobility management mechanism more flexible. For example, for mMTC UEs possessing characteristics such as immobility, cost-sensitivity, and stringent requirements on energy-efficiency, 5G cellular networks will wait for the communication request from UEs and reactively start the data transmission (Fig. 2c). Meanwhile, 5G cellular networks also tailor mobility management for some vertical industries, based on regional characteristics. In other words, once UEs enter a specific region, they could be granted higher-level support (e.g., dual connectivity) for mobility management and thus update their locations in a more proactive manner.

Management and Orchestration

Recently, the industry has witnessed the increasing maturity of software-defined networks (SDN) [8]. In particular, some well known operators such as AT&T, China Mobile, Telefonica, and vendors such as Cisco and Huawei, have co-established the Open Networking Lab (http://onlab.us/) to bring openness and innovation in SDN to the Internet, and initiated the project called Central Office Re-architected as a Datacenter (CORD). CORD has successfully completed the virtualization of existing hardware devices such as CPE (customer premises equipment), OLT (optical line transmission) and BNG (broadband network gateway), and produced software counterparts (e.g., applications running on open network operating system (ONOS)) on top of commodity hardware. Moreover, CORD has provided a framework on which these software elements (plus any other cloud services an operator may want to run) can be plugged into, leading to a coherent end-to-end system. Therefore, operators of 5G cellular networks might borrow the concept of CORD and deploy selected functionalities according to their own demands. Meanwhile, in order to orchestrate services from different vendors, 5G network operators can leverage a more centralized SDN controller and adopt various means such as exposing the same infrastructure-level interfaces or using common cloud operating systems (e.g., ONOS) to shield differences between multi-vendor hardware servers in a distributed deployment manner [5].

Service Provisioning Management

In addition to softer air interface enabled by FBMC or UFMC, it is also expected in 5G to intelligently program the forwarding route of one service by leveraging the application interfaces (APIs) in SDN and have a more flexible service provisioning stack. With the evolution of SDN and network function virtualization (NFV), 5G cellular networks have advocated a revolutionary concept called network slicing (NS) [9]. Instead of building dedicated networks for different services, NS allows operators to intelligently create

Cellular networks have alternative options in the 5G era for access and service provisioning mechanisms and thus gain the foundation to apply preliminary intelligence. However, 5G cellular networks are still lagging behind what is actually required in practice.

customized network pipes to provide optimized solutions for different services that require diverse functionalities, performance metrics, and isolation criteria. Moreover, session management in 5G will be able to adapt to UE attributes and service requirements, by adjusting configurations such as session categories, anchoring points, and service continuum capabilities. Specifically, mobile edge has potentially evolved to replace its forwarding-only functionality to an area equipped with storage, memories, and computational power capabilities [3]. Therefore, according to practical requirements, UEs could select anchoring and forwarding points between the anchoring point (e.g., serving gateway) in core networks (CNs) and the mobile edge networks. For example, in Fig. 2(d), services with stringent requirements on mobility and service continuum could shift their anchoring points to the edge networks with closer proximity. Moreover, in 4G, device-to-device (D2D) communication merely supports proximity services and public safety communications. But network-assisted direct communication between vehicles and UEs comes to a reality, and the vehicle-to-vehicle infrastructure (V2X) services are becoming a hot topic to better accommodate the URLLC services of vertical industries (e.g., automobiles).

Thanks to the huge advance in signal processing capabilities evolved as Moore's Law, 5G cellular networks can take advantage of advanced yet computation-consuming technologies in almost every aspect spanning from the physical layer to the network architecture. Therefore, 5G cellular networks are able to provide alternative options for different scenarios, exhibit some preliminary intelligence, and satisfy the minimal requirements to adopt complete AI.

Artificial Intelligence for Cellular Networks

Cellular networks have alternative options in the 5G era for access and service provisioning mechanisms and thus gain the foundation to apply preliminary intelligence. However, 5G cellular networks are still lagging behind what is actually required in practice. First, the number of configurable parameters in a typical 4G node has increased to 1500 from 500 in a 2G node and 1000 in a 3G node [4]. If this trend continues, a typical 5G node is expected to have 2000 or more parameters. Therefore, it is critical to enhance intelligence in the 5G era to realize the self-organizing features (e.g., self-configuration, self-optimization, and self-healing). Second, the service types (e.g., eMBB, URLLC, mMTC) defined in the 5G era are static. However, new types of services continually evolve, and the pattern in existing services frequently changes as well. In this case, 5G cellular networks still lack functionalities to automatically recognize a new type of service, infer the appropriate provisioning mechanism, and establish the required network slice. Third, 5G cellular networks heavily depend on a centralized network architecture in SDN, and still lack the agility and robustness under the scenario of ever-increasing heterogeneous and complicated cellular networks. To self-organize parameters that become significantly larger, auto-build the network slices for emerging services, and gain sufficient flexibility for network maintenance, it is essential for cellular networks to observe environment variations, learn uncertainties, plan response actions, and configure the networks properly. Coincidentally, AI mainly solves how to learn the variations, classify the issues, forecast future challenges, and find potential solutions, by interacting with the environment. Therefore, cellular networks could leverage the concept of cognitive radio [10] and interact with the environment using AI, so as to fully accelerate the evolution and enter into a brand-new intelligent 5G era.

AI has evolved to multi-disciplinary techniques such as machine learning, optimization theory, game theory, control theory, and meta-heuristics [11]. Among them, machine learning belongs to one of the most important subfields in AI. Usually, depending on the nature of the learning objects and signals to a learning system, machine learning is typically classified into three broad categories:

Supervised Learning: A supervised learning agent will be fed with example inputs and their desired outputs, and aims to determine a general rule that nicely maps inputs to outputs. Supervised learning has been widely applied to solve channel estimation issues in cellular networks. For example, assume that there exists a wireless channel h, the receiver tries to exploit the transmit preamble s and the received signal $y = hs + n_0$ (with n_0 denoting the noise) to estimate h. For such a supervised learning problem, it is common to use probabilistic models to characterize the transition probability $\mathcal{P}(y|s)$ from s to y and take advantage of the well known Bayes learning methods to obtain the results. The well known Kalman filtering and particle filter methods also play a very important role in optimizing cellular networks.

Unsupervised Learning: Compared to the aforementioned supervised learning, the input information for unsupervised learning does not possess *priori* labels. Therefore, the unsupervised learning agent has to depend on its own capability to find the embedded structure or pattern in its input. Usually, unsupervised learning aims to discover hidden patterns and find the suitable representation in the input data. In the field of AI, unsupervised learning is applied to estimate the hidden layer parameters in neural networks and plays an important role in deep learning methods. Meanwhile, unsupervised learning may be the most widely applied AI category in cellular networks. For example, principal component analysis (PCA) and singular value decomposition (SVD) methods have been used to manipulate the receiving matrix of massive MIMO to reduce the computational complexity. Moreover, 5G NoMA receivers also adopt some factor graph-based methods such as expectation-maximization and message-passing algorithms to achieve lower bit error rate. On the other hand, some classifiers such as the K-means Algorithm are also useful to detect network anomalies.

Reinforcement Learning: Inspired by both control theory and behaviorist psychology, the reinforcement learning agent could obtain its goal by interacting with a dynamic environment. However, the agent does not have explicit knowledge of whether it has come close to its goal. Instead, the agent should take actions in an environment so as to maximize the cumulative reward in a

Modules	Examples	Algorithms	Comments
Sensing	Detection of network anomalies or events by multiple-entry data from hybrid sources	Logistic Regression (LR) Support Vector Machine (SVM) Hidden Markov Model (HMM)	Hypothesis test plays an important role in this aspect. But different algorithms have specific scenarios. Compared to SVM, LR is more suitable for sensing scenarios with a heavy number of property combinations and stringent accuracy requirements. On the other hand, HMM is also applicable for sensing if we try to compute the state's probability and regard a comparably larger probability as the occurrence of anomalies or events.
Mining	Classifying services according to the required provisioning mechanisms (e.g., bandwidth, error rate, latency)	Supervised learning: • Gradient Boosting Decision Tree (GBDT) Unsupervised learning: • Spectral Clustering • One-class SVM • Replicator Neural Networks (RNN)	Supervised learning heavily relies on the labeling quality of data samples, while unsupervised learning depends on the accuracy or suitability of parameter (e.g., threshold) settings.
Prediction	Forecasting the trend of UE mobility or the traffic volume of different services	Kalman Filtering (KL) Auto-Regressive Moving Average (ARMA) Auto-Regressive Integrated Moving Average (ARIMA) Deep Learning (DL): • Recurrent Neural Networks (RNN) • Long-Short Term Memory (LSTM) Compress Sensing (CS)	KL/ARMA/ARIMA could well follow the variations of a one-time sequence, but fail to capture the characteristics behind this sequence. On the other hand, DL algorithms like RNN and LSTM have the capability to find the embedded characteristics and leverage the long-time dependency in the sequence. Meanwhile, CS is a dedicated tool to investigate the universal sparsity in mobile traffic series and resources (e.g., BSs).
Reasoning	Configuration of a series of parameters to better adapt services.	Dynamic Programming (DP) • Branch-and-Bound Method • Primal-and-Dual Method Reinforcement Learning (RL) • Actor-critic Method • Q-Learning Method Transfer Learning (TL)	DP, which might belong to a generalized sense of AI, is generally exploited to solve the Bellman equation, based on complete knowledge of the considered environment. In contrast, RL approximates the optimal solution of the Bellman equation without knowing the environment *a priori*, by iteratively updating its policy or value function. Besides, a combination of RL and TL could yield superior results.

TABLE 1. Typical AI algorithms to enhance cellular networks.

	4G	5G	AI modules				Intelligent 5G
			Sensing	Mining	Prediction	Reasoning	
Services	MBB	eMBB/mMTC/URLLC					Service-aware
RRM	Granted	Granted or grant-free Flexible bandwidth Flexible symbol length	✓	✓	×	✓	UE-specific on-demand
MM	Unified	On-demand	✓	×	✓	✓	Location tracking/awareness
MANO	Simple	Operator-tailored	✓	✓	✓	✓	Enhanced self-organizing and trouble-shooting capability
SPM	Unified	End-to-end NS	×	✓	✓	✓	Network slice auto-instantiation

TABLE 2. The evolution toward intelligent 5G.

Markov decision process (MDP). Therefore, reinforcement learning demonstrates strong pattern recognition ability. Researchers in the field of cognitive radio usually model the dynamic transition of spectrum availability as a Markov chain, and extensively apply reinforcement learning methods (e.g., Q-learning and the actor-critic method [12, 13]) to make the decision whether or not it is suitable for secondary transmission in one primary licensed spectrum, in terms of least interference to the primary spectrum.

Table 1 summarizes what typical AI algorithms could solve. Apparently, AI can be used to enhance the response of cellular networks to stimuli by learning key network parameters. For example, AI makes it possible to sense in a timely manner the variations in network traffic, resource utilization, user demand, and possible threats, and further makes it possible to smartly coordinate UEs, base stations (BSs), and network entities. Table 2 illustrates functionalities upon which intelligent cellular networks may be built.

Figure 3 illustrates a possible AI-empowered 5G cellular network architecture, in which an AI controller will act as an application on top of ONOS or an independent network entity, and communicate with RAN, CN, or global SDN controllers using open interfaces. Specifically, the AI center will read service-level agreements (e.g., requirements on rate, coverage, failure duration, redundancy, etc.), UE-level information (e.g., receiver category, battery limitation), network-level information (e.g., spectrum, number of serving subscribers, QoS (quality of service), key performance indicators of network functions, scheduled maintenance period, etc.), and infrastructure-level information (e.g., server type, CPU, memory, storage, network standard) from the SDN control-

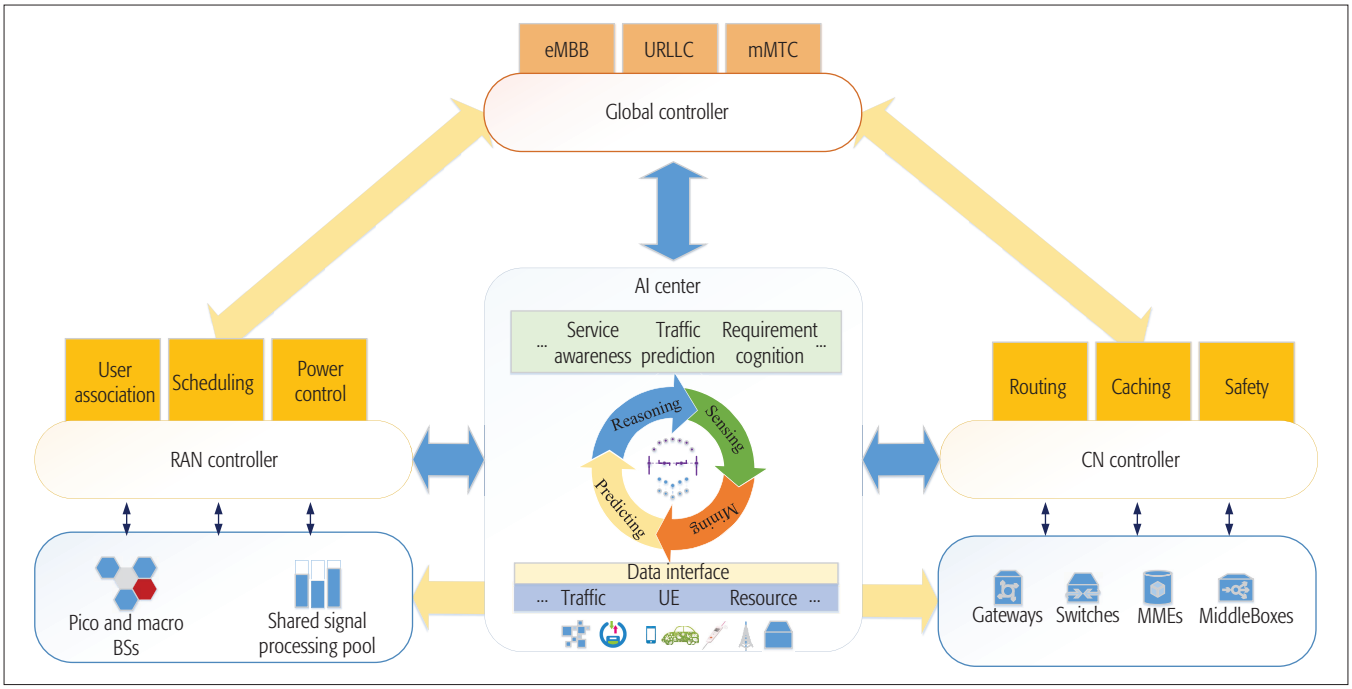

FIGURE 3. 5G cellular networks enabled by AI.

lers, so as to get in touch with cellular network data such as traffic information, UEs, and network resources. Afterward, the AI center will utilize its embedded modules (e.g., sensing, mining, prediction, and reasoning) to process the obtained information, and feedback learning results, which may include traffic characteristic analysis reports (e.g., service provisioning suggestion), UE-specific controlling information (e.g., serving priority, bandwidth allocation, mobility tracking command), and network configuration notification (e.g., parameter adjustment, access method, network error alert), to the SDN controllers. For example, AI leverages the sensing module to track the location of UEs and uses the predicting module to forecast the mobility trend based on the historical moving pattern. Afterward, it takes advantage of the reasoning module and proactively notifies the UEs to update the location record, so as to prepare handover resources and save signaling cost of mobility management.

On the other hand, 5G cellular networks can maintain the normal working status under the condition of potential damages (e.g., hacking) to the AI center. Meanwhile, the AI center could (semi-)periodically exchange information with the SDN controllers in normal states, while it starts emergent responses to schedule the minimum required resources, once the conventional SDN controllers encounter malfunctions. Therefore, compared to the complete centralized architecture in conventional networks, the AI center and the SDN controllers virtually constitute a multi-tier decision-making system, thus being able to improve the network robustness.

Opportunities

In addition to the benefits to the RRM, MM, MANO, and SPM, AI could further contribute to solving the following issues.

Overloading of Cellular Network Data: Cellular networks generate vast volumes of records by provisioning different services and types of UEs under various channels, network entity configurations, and energy consumption conditions. In particular, AI could exploit cellular network data to forecast potential events and predict traffic volume and help to pre-allocate network resources. Meanwhile, AI provides a unified means to mine the relevancy in such abundant data and helps build a more concrete mapping from service requirement to network configuration. Furthermore, AI could generate some operating reports to describe and summarize the subscriber and network experience statistics, which is relevant when setting billing and market policies.

Inter-Networking of Heterogeneous Cellular Networks: Currently, operators have deployed heterogeneous BSs in the 4G era, including pico-cells (providing high capacity), micro-cells (providing wide coverage for eMBB) and macro-cells (supplying even wider coverage for signaling and mMTC services). AI could analyze the requirements of one emerging service and contribute to the selection of the most appropriate access point to accommodate such a service, in terms of SE, EE, or other more complicated criteria. For example, AI could generate UE-specific policies to make some UEs attach to pico BSs for larger throughput while letting some UEs connect to macro BSs to maintain fundamental information exchange.

Difficulties in an Operator Supporting Subsystem: Usually, cellular networks merely rely on thresholds to monitor network anomalies. Therefore, operator engineers have to be vigilant enough to systematic alerts and read user guides to cope with unexpected network conditions. But AI could use cellular network data to derive common network traffic patterns. Therefore, when networks experience traffic with unfamiliar patterns, AI can start troubleshooting at the very beginning. Similarly, networks could take advantage of AI to

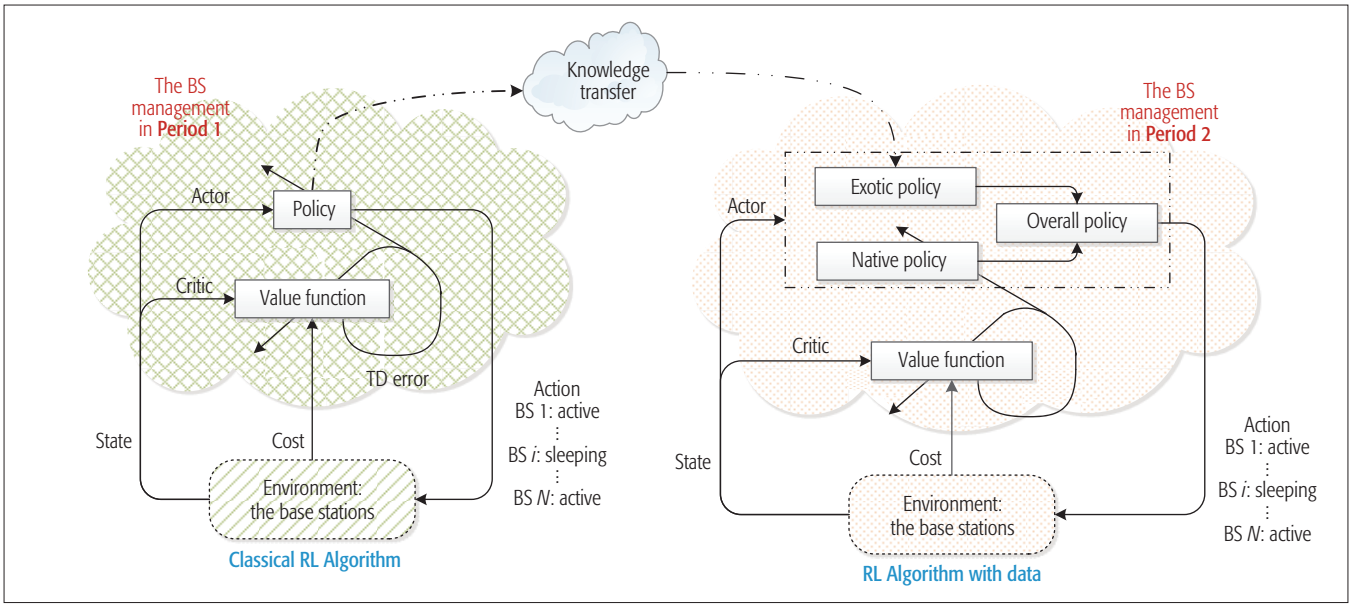

FIGURE 4. The AI application framework for reinforcement learning-based greener cellular networks.

shield against potential safety threats, once AI perceives activity and address anomalies.

Challenges in Integrating RANs and CNs: Usually, the management of RANs and CNs is isolated, thus being not scale enough for network evolution. As mobile edge computing becomes more common, AI grants the controller more power and capability to jointly schedule wireless and wired resources, choose the appropriate content distribution and caching server (e.g., edge and core server), and provide more unified protection against possible network threats.

Challenges

In spite of the apparent opportunities, there are also challenges to apply AI to cellular networks. First, in the 5G era, network data is a double-edged sword. It definitively provides precious opportunities for AI to analyze trends and recognize patterns. However, it is also difficult to derive a simple model or pattern that perfectly matches the data. Therefore, the derived results, which possibly consist of lots of parameters, are very difficult to read and lose value for practical application. Second, in order to save and process cellular network data in a timely manner, a significant amount of storage and computational resources are needed, and there might be threats to information security. Also, it usually is necessary to collect data in a centralized manner before applying most AI algorithms. These factors inevitably add to the computational capability of network entities and BSs and put a huge burden on the practical cost of products.

Use Case: Traffic-Aware Greener Cellular Networks

In this section, we demonstrate how to take advantage of AI to enhance the MANO, so as to build greener cellular networks [12]. It is well known that over 80 percent of power consumption takes place in RANs, especially the BS, since the present BS deployment is on the basis of peak traffic loads and generally stays active irrespective of the huge variations in traffic load. Therefore, benefiting from cloud pooling of baseband resources, an SDN controller [8] can be leveraged to sense traffic variations and adjust the working status of under-utilized BSs, thus improving energy efficiency. Meanwhile, an AI center plays a crucial role in learning traffic variations and adjusting BS switching policy. Here, we briefly talk about two AI schemes to design traffic-aware greener cellular networks, and show how AI could effectively solve this problem.

The most intuitive approach is to first forecast traffic loads in the near future and then adjust the status of BSs, so as to satisfy the predicted traffic loads but incur minimal energy consumption. For traffic prediction, we can resort to the prediction module of AI. For example, our previous work [14] modifies the popular ARMA algorithm by incorporating traffic sparsity in both the temporal and spatial domains and demonstrates the prediction error for aggregate traffic records collected from China Mobile could be as low as 15 percent, in terms of normalized root mean square error. Furthermore, we also predict the service-level traffic with more bursty property in [15] and demonstrate appealing accuracy as well, by deriving the traffic model (e.g., α-stable models) from realistic traffic records and utilizing a stable model-based compressive sensing algorithm.[1] On the other hand, in order to determine the appropriate BS switching policy, we take advantage of the branch-and-bound algorithm, one kind of dynamic programming method, to solve the formulated optimization theory [14] and show that along with practical traffic variations, significant energy savings could still be expected. In particular, when traffic hits to the lowest point in the early morning (from 6 AM to 8 AM), the AI-induced BS switching policy merely costs 55 percent of the energy that would be consumed if we emply no energy saving scheme.

In order to attain the BS switching solution, it is also viable to merge the prediction and reasoning

[1] It can be regarded as a success to combine the AI sensing and prediction modules.

Based on the sensing, prediction, and reasoning modules, AI could contribute to designing traffic-aware greener cellular networks. Meanwhile, such a design represents the typical methodology to exploit AI and reflects the effectiveness of AI on future cellular networks.

modules together, by applying a Markov chain to model possible traffic load variations and making use of the actor-critic algorithm in Fig. 4 [13], a reinforcement learning (RL) approach [12]. Specifically, the AI center would first estimate traffic load variations based on the on-line experience. Afterward, it can select one of the possible BS switching operations under the estimated circumstance and then decrease or increase the probability of the same action to be later selected according to the feedback cost information from SDN controllers, by updating the policy and value function. After repeating the actions and gradually knowing the corresponding costs, the AI center would know how to switch the BSs for one specific traffic load profile. Meanwhile, as cellular network traffic exhibits strong self-similarity, the AI center could exchange the learned results in different periods and optimize the strategy in a faster manner. Our numerical results in a practical BS deployment scenario with simulated traffic traces also show that without knowing traffic variations beforehand, this RL method [12] could still converge quickly and approach the energy saving performance of a solution with perfectly predicted traffic knowledge.

In a word, based on the sensing, prediction, and reasoning modules, AI could contribute to designing traffic-aware greener cellular networks. Meanwhile, such a design represents the typical methodology to exploit AI and reflects the effectiveness of AI on future cellular networks.

Conclusion

In this article, we have suggested applying AI to cellular networks. We first discussed the initial intelligence emerging in nearly all aspects of 5G cellular networks, including radio resource management, mobility management, general management and orchestration, and service provisioning management. Following such intelligence, we argued it is still essential to bring more AI functionalities to 5G cellular networks by envisioning several prospective opportunities and listing some potential challenges. Finally, we provided a use case on how to obtain greener 5G cellular networks and demonstrated the thrilling effectiveness of AI. We could boldly argue that AI empowered 5G cellular networks will successfully enter the central stage of a digitalized world.

Acknowledgment

This article is supported by the National Postdoctoral Program for Innovative Talents of China (No. BX201600133), the Program for Zhejiang Leading Team of Science and Technology Innovation (No. 2013TD20), the Zhejiang Provincial Technology Plan of China (No. 2015C01075), the National Natural Science Foundation of China (No. 61501510), and the Natural Science Foundation of Jiangsu Province (Grant No. BK20150717).

References

[1] C. X. Wang *et al.*, "Cellular Architecture and Key Technologies for 5G Wireless Communication Networks," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 122–30.
[2] J. G. Andrews *et al.*, "What Will 5G Be?" arXiv:1405.2957 [cs, math], May 2014; available: http://arxiv.org/abs/1405.2957.
[3] D. Soldani and A. Manzalini, "Horizon 2020 and Beyond: On the 5G Operating System for a True Digital Society," *IEEE Vehic. Tech. Mag.*, vol. 10, no. 1, Mar. 2015, pp. 32–42.
[4] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to Empower SON with Big Data for Enabling 5G," *IEEE Network*, vol. 28, no. 6, Nov. 2014, pp. 27–33.
[5] X. Zhou *et al.*, "Network Slicing as a Service: Enable Industries Own Software-Defined Cellular Networks," *IEEE Commun. Mag.*, vol. 54, no. 7, Jul. 2016, pp. 146–53.
[6] X. Wang, X. Li, and V. C. M. Leung, "Artificial Intelligence-Based Techniques for Emerging Heterogeneous Network: State of the Arts, Opportunities, and Challenges," *IEEE Access*, vol. 3, 2015, pp. 1379–91.
[7] C. L. I *et al.*, "New Paradigm of 5G Wireless Internet," *IEEE JSAC*, vol. 34, no. 3, Mar. 2016, pp. 474–82.
[8] R. Li *et al.*, "The Prediction Analysis of Cellular Radio Access Network Traffic: From Entropy Theory to Networking Practice," *IEEE Commun. Mag.*, vol. 52, no. 6, June 2014, pp. 238–44.
[9] 3GPP, "Study on architecture for next generation system (TR 23.799)," 2016; available: http: //www.3gpp.org/DynaReport/23799.htm.
[10] J. Mitola and G. Q. Maguire, "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Commun.*, vol. 6, no. 4, Aug. 1999, pp. 13–18.
[11] J. Qadir *et al.*, "Artificial Intelligence Enabled Networking," *IEEE Access*, vol. 3, 2015, pp. 3079–82.
[12] R. Li *et al.*, "TACT: A Transfer Actor-Critic Learning Framework for Energy Saving in Cellular Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, Apr. 2014, pp. 2000–11.
[13] V. Konda and J. Tsitsiklis, "Actor-Critic Algorithms," *SIAM J. Contr. Optim.*, vol. 42, no. 4, 2000, pp. 1143–66.
[14] R. Li *et al.*, "Energy Savings Scheme in Radio Access Networks via Compressive Sensingbased Traffic Load Prediction," *Trans. Emerg. Telecommun. Tech. (ETT)*, vol. 25, no. 4, Apr. 2014, pp. 468–78.
[15] R. Li *et al.*, "The Learning and Prediction of Application-Level Traffic Data in Cellular Networks," arXiv:1606.04778 [cs], June 2016; available: http://arxiv.org/abs/1606.04778

Biographies

Rongpeng Li received the Ph.D and B.E. from Zhejiang University, Hangzhou, China and Xidian University, Xi'an, China, in June 2015 and June 2010, respectively, both as "excellent graduates." He is now a postdoctoral researcher at Zhejiang University, Hangzhou, China. From August 2015 to September 2016, he was a researcher at the Wireless Communication Laboratory, Huawei Technologies Co. Ltd., Shanghai, China. His research interests focus on applications of artificial intelligence, data-driven network design, and resource allocation of cellular networks (especially full-duplex networks). He was granted by the National Postdoctoral Program for Innovative Talents, which had a grant ratio of 13 percent in 2016.

Zhifeng Zhao (corresponding author) is an associate professor with the College of Information Science and Electronic Engineering, Zhejiang University, China. He received the Ph.D. degree in communication and information systems from the PLA University of Science and Technology, Nanjing, China, in 2002. From September 2002 to December 2004, he was a postdoctoral researcher at Zhejiang University. His research areas include cognitive radio, wireless multi-hop networks (ad hoc, mesh, WSN, etc.), wireless multimedia network and green communications.

Xuan Zhou is a senior architect with the Service Provider Operation Lab (SPO Lab) of Huawei Technologies. He received his Ph.D. in communication and information systems from Zhejiang University, Hangzhou, China. From 2009 to 2014, he worked as a system engineer at China Mobile Zhejiang Company. His research efforts focus on innovative service and network management in 5G, network function virtualization, and software-defined networks. He is also the architect of the world's first 5G end-to-end network slicing demo, which was shown at Mobile World Congress (MWC) 2016 in Barcelona, Spain.

Guoru Ding [S'10, M'14, SM'16] received his B.S. degree (hons.) in electrical engineering from Xidian University, Xi'an, China, in 2008, and his Ph.D. degree (hons.) in communications and information systems from the College of Communications Engineering, Nanjing, China, in 2014. Since 2014, he has been an assistant professor with the College of Communications Engineering. Since April 2015, he has been a postdoctoral research associate at the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His research interests include cognitive radio networks, massive MIMO, machine learning, and big data analytics over wireless networks. He currently serves as a guest editor of the *IEEE Journal on Selected*

Areas in Communications. He was a recipient of the Best Paper Awards from EAI MLICOM 2016, IEEE VTC 2014-Fall, and IEEE WCSP 2009.

YAN CHEN received her B.Sc. and Ph.D. degrees from Zhejiang University in 2004 and 2009, respectively. She was a visiting researcher at the University of Science and Technology (HKUST) from 2008 to 2009. In 2009, she joined Huawei Technologies (Shanghai) Co., Ltd. She was the team leader and project manager of the internal project Green Radio Excellence in Architecture and Technology (GREAT) from 2010 to 2013, during which time she was also the project leader of the umbrella project Green Transmission Technologies (GTT) at the GreenTouch™ Consortium. Since 2013, she has been the technical leader and project manager of the internal 5G air interface design project. Her current research interests are more toward future communication system design to efficiently support the multiplexing of different service scenarios with diversified requirements.

ZHONGYAO WANG received a B.Sc. in mathematics from the School of Mathematical Sciences, Peking University, Beijing, China in 2009. He joined the Deep Algorithm Department of Alibaba Group, Hangzhou, China in 2013, where he currently is a staff engineer. From 2010 to 2013, he worked in ASML (Brion) as a software engineer. He has focused on anti-fraud, prediction, and recommendation in the past few years. He is now responsible for user profile and ID-mapping, which are two key fundamental infrastructures inside Alibaba group.

HONGGANG ZHANG is a professor with Zhejiang University, China, and was the International Chair Professor of Excellence at the Université Européenne de Bretagne and Supélec, France from 2012 to 2014. He is also an honorary visiting professor with the University of York, UK. He served as the chair of the Technical Committee on Cognitive Networks of the IEEE Communications Society from 2011 to 2012. He is currently involved in research on green communications, serving as the series editor for the *IEEE Communications Magazine* series on Green Communications and Computing Networks.