ISSN: 2688-9323

A Survey and Discussion of Artificial Intelligence for Mobile Communication Network

Sheng Liu[a], Jiaxin Fan, Hanhui Li

Wuhan Maritime Communications Research Institute, Wuhan 430205, China

[a]sliuee@foxmail.com

Abstract

The continued growth in node scale, data traffic and service types pose severe challenges to the mobile communication network in terms of service quality and adaptability. Traditional tools and strategies have been difficult to handle the complex network planning, control management, fault location and troubleshooting. The introduction of artificial intelligence (AI) brings new ideas to solve those problems, with a continuous self-evolution capability. This paper establishes an architecture of AI-enabled mobile communication network, and investigates the application of AI from specific aspects such as network slicing, edge computing and mobility management. In addition, the current challenges in AI deployment for mobile communication network are put forward and discussed.

Keywords

Artificial Intelligence (AI), mobile communication network, architecture.

1. Introduction

The rapid development of the mobile communication network in the past two decades has brought dramatical changes to working and living. With the rise of smart phones and the mobile internet, new services continue to emerge, ranging from initial voice and text messages, to the social media and short videos, and now to the immersive media and ultra-high-definition videos [1]. In addition to satisfying the connection between people, the machine-centric Internet of Things is also booming. Hundreds of millions of terminals will access the network, keeping connection and exchanging information with each other. Diversified user terminal access methods and network structures pose unprecedented challenges to the coverage efficiency, resource allocation, and user scheduling of the mobile communication network.

With the sustaining innovations of algorithms and improvement of the hardware performance, artificial intelligence (AI) technologies have been widely used in computer vision, natural speech recognition, e-commerce and other fields, and believed to lead the new industrial revolution. Various branches of the industry are beginning to be strongly combined with AI, to make corresponding predictions and decisions by learning the hidden laws from the data. Along with the expansion of the application scope, AI has gradually penetrated into the field of communications. The advantages of AI technologies in terms of complex features, time-series curve fitting, association relationship reasoning, and optimized decision management can enable the network to have the ability of self-learning and self-updating [2, 3].

In this article, we first proposed an architecture of AI-enabled mobile communication network. Then, AI applications in specific aspects of mobile communication network are investigated and discussed. Finally, several existing challenges for the deployment of AI are presented, which pointed out the further research directions.

ISSN: 2688-9323

2. Architecture of AI-Enabled Mobile Communication Network

The mobile communication network has now developed into a system with a high level of automation. But it is still established on the basis of a rule set derived from a systematic analysis of prior domain knowledge and experience. To realize the future "autonomous network", AI technologies must be closely integrated, to build a network on a data-driven model. With the continuous evolution of the communication network, AI will be more extensively and deeply integrated into the construction of the system, for reducing human error, improving network performance, achieving agile deployment of services, providing a better user experience, and accelerate digital transformation of society.

Figure 1 shows the architecture of an AI-enabled mobile communication system. Ai is configured and distributed in different parts of the system, so that each component becomes an intelligent body, giving continuous vitality to the autonomous network.

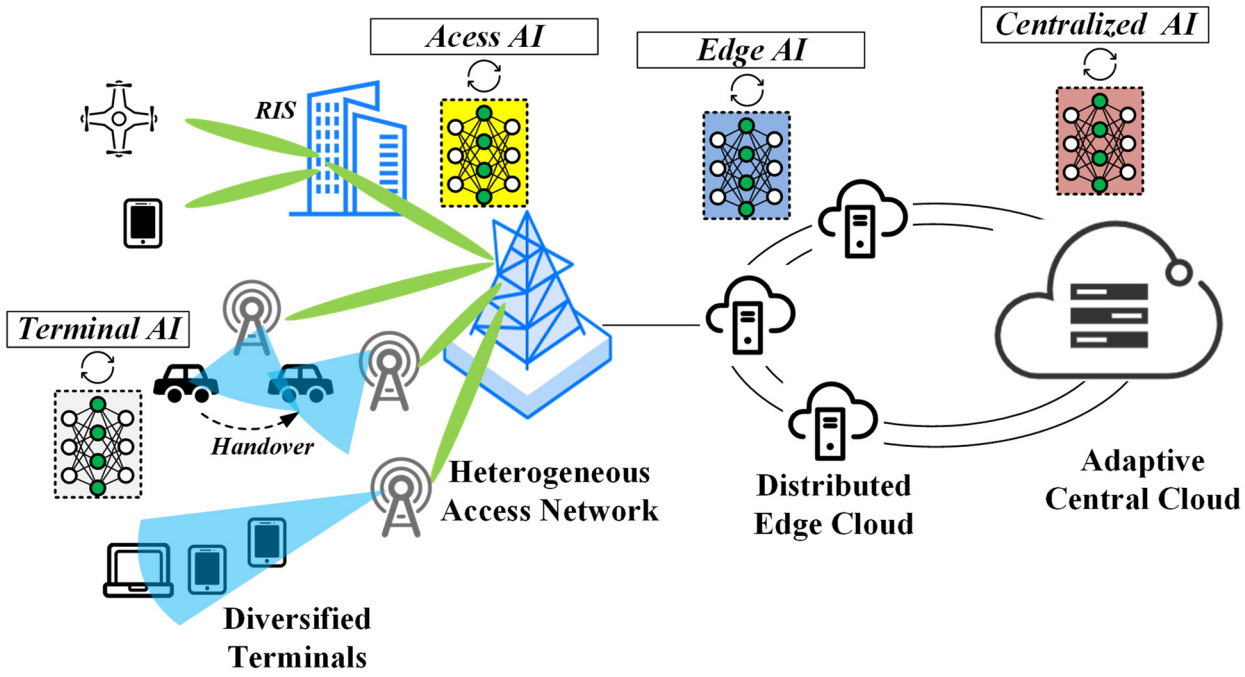

Fig.1 Architecture of AI-enabled mobile communication network

2.1. Central AI

By deploying AI strategies and applications in the core network, central AI can perform analysis of intelligent slice load, network function load, network performance, and abnormal user behavior for different scenarios. The deployment planning, operation and maintenance of the network can therefore be more efficient.

2.2. Edge AI

Edge AI is configured on the edge of the network, which can provide computing and storage resources required for AI algorithms. The introduction of Edge AI can ensure faster business response, shorten the launch time of new services, and enhance the matching rate of the customer demand. At the same time, through edge-cloud collaboration, model training updates can be upgraded from months to hours, which can remarkably improve the efficiency of the operational management.

2.3. Access AI

Access AI embeds an AI accelerator on the base station side to perform user traffic management and spectrum resource allocation. At the same time, widely used wireless technologies such as

ISSN: 2688-9323

carrier aggregation, massive MIMO and beamforming can be combined with AI algorithms to enhance the network performance in coverage, capacity and handover, improve positioning accuracy and reduce interference.

2.4. Terminal AI

Terminal AI is closer to users for deployment. By completing data processing on the end side, terminal AI can better perceive its own operating environment and reduce reliance on the cloud. It can provide users with application modes with faster response, richer content and more intelligent cognition. Meanwhile, the privacy of users will be better protected.

3. Typical Applications

3.1. Network Slicing

Network slicing is an on-demand networking technique that allows operators to separate multiple virtual end-to-end networks on unified infrastructures. Each network slice carries out logic isolation for various types of applications from the wireless access network, bearer network and core network. With the increase of the heterogeneous and dynamic characteristics in future mobile communication systems, statistical models are more difficult to deal with the slice and network performance optimization. [4] incorporated AI into a network-slicing based architecture, to exploit the slice service level agreement (SLA) monitoring data and the slice resource utilization data to facilitate slice deployment, slice adaption, and slice update. Compared to conventional systems controlled by human-driven approaches, the introduction of network slicing has greatly increased the complexity of the network management process, due to a large number of independent logical network instances. [5] proposed a framework that integrates AI into different key functions of slice management, which shows high gain in scheduling of slice traffic, resource allocation, and admission control of new slices.

3.2. Edge Computing or Edge AI

Edge computing refers to an open platform close to the source of things or data, that integrates network, computing, storage, and application core capabilities. By execution of computation-intensive and delay-sensitive tasks at the edge servers, the processing latency can be seriously reduced. In fields such as intelligent transportation and industrial Internet of Things, edge computing has outstanding advantages over traditional cloud computing due to its low latency, high reliability, and local security and privacy protection. Combined with AI, edge computing can better play its own advantages, and in turn make AI thrive through richer data and application scenarios. AI can quickly analyze the massive data generated at the edge of the network and provide high-quality decision-making capabilities. Many AI applications related to mobile and the Internet of Things represent a series of practical applications. Those applications are computationally and energy-intensive, sensitive to privacy and latency, which can naturally be consistent with edge computing [6]. The intelligence of edge computing services is not only reflected in the storage and processing of edge data that supports AI, but also in behavior feedback, automatic networking, load balancing, and data-driven network optimization. Therefore, a framework called edge cognitive computing (ECC) is utilized to provide the cognition of users and network environmental information, and the elastic cognitive computing services with higher energy efficiency and user experience [7]. For intelligent transportation scenario, an AI-empowered vehicular network architecture is established for smart vehicular edge computing and caching [2], that can efficiently allocate resources to improve system utility.

ISSN: 2688-9323

3.3. Mobility Management

Mobility management mainly concern the management of mobile terminal location information, security, and service continuity, to optimize the connection between the terminal and the network, and provide guarantees for the application of various network services. With the rapid increase in the number of end users and heterogeneous cell sizes, mobility management has become one of the key challenges for mobile network, especially in future ultra-dense network consisted of numerous small base stations (SBS). The frequently handover procedure of user equipment from one attached BS to another may cause large delay and obvious throughput reduction. Existing strategies lack sufficient accuracy in predicting mobility and may not well support the handover procedure. An intelligent mechanism based on Long Short Term Memory scheme is utilized to predict movement trends, and bring significant benefits for mobile users while guaranteeing the network energy efficiency [8]. When heterogeneous user mobility patterns occurs in the ultra-dense network, frequent handover (HO) process may diminish the capacity gain, and consume more energy. To optimize handover controllers, a framework based on DNN is used to reduce HO rates and ensure the system throughputs [9]. Furthermore, the stochastic reward of the classical online learning model for mobility management is fixed and inaccurately reflect the real scenario of the user movement or environment change. Therefore, a piece-wise stationary online learning algorithm is proposed in [10], to learn the varying throughput distribution and solve the frequent handover problem.

3.4. Channel Coding and Decoding

Channel coding is a process of adding redundancy to the source data to make it adapted to the complex channel transmission environment. While channel decoding is the process of making a decision on the received signal. [11] introduced a deep learning-based polar code construction algorithm by inherently taking the actual decoder and channel into account. This learned code shows better performance over conventional polar code in terms of decoding latency and complexity. The deep learning technique was also able to improve the Belief Propagation (BP) decoding algorithm of the High Density Parity Check Code (HDPC) codes, by training the edges of the Tanner graph that represent the given linear code [12]. In order not to be limited to optimize well known decoding schemes, one-shot decoding of random and structured codes by using deep neural networks was proposed [13], which indicates deep learning-based decoding a promising decoding approach.

3.5. Channel Estimation

Channel estimation refers to the process of describing the influence of the physical channel on the input signal, which has a serious impact on the demodulation result of the received signal. For DOA estimation and channel estimation of massive MIMO systems, a framework has been proposed that integrates the deep learning technique into the systems by leveraging the spatial structure [14]. [15] presents a combination of the channel estimation and multi-user detection found on sparse based k-nearest neighbor algorithm and cat swarm optimization algorithm, which shows better performance in terms of pilot patterns with respect to BER and MSE. Most existing channel estimation algorithms require channel model-based pre-training, which may not converge in optimum on real channel scenarios. To overcome this limitation, an online channel estimation and equalization scheme has been proposed for OFDM systems [16]. The computational complexity is also significantly reduced by using a single hidden layer feedforward network.

3.6. RF Front Ends and Antenna Design

For optimizing the system performance of future mobile communication systems, a novel tunable radio frequency (RF) frontend that combines advances in AI and ML is proposed [17]. The power of cognition is added to the RF front ends to optimally control commonly used RF

ISSN: 2688-9323

components. [18] introduced an integrated machine learning and coordinated beamforming strategy that attains high data rate gains in high-mobility large-array scenarios. For satellite-terrestrial networks in future mobile systems, an AI-based self-learning (ASL) network framework is proposed to realize the pointing and tracking for ground mobile stations and terminals [3]. Machine learning methods, such as Gaussian process regression, SVM and ANNs, have also been widely applied to accelerate the entire antenna design procedure. The core idea is to build a model to predict the designated characteristics at the new points in the design space using the training set generated at the sampled points [19].

3.7. End-to-end Optimization

Those above AI applications for the physical layer focus on the individual processing tasks such as modulation, channel estimation, antenna design, etc. While the End-to-end performance of the communication system based on AI should also be evaluated. [20] represent the end-to end communications system as one deep neural network (NN) that can be trained as an autoencoder, and extend this ideal to multiple transmitters and receivers. For the end-to-end system based on DNN, the channel state information is constantly changing with time and location, resulting that the gradient cannot be reversed and the neural network model cannot be trained. [21] utilize stochastic perturbation techniques to train an end-to-end communication system, without relying on explicit channel models.

3.8. Spectrum Allocation

Spectrum resources are the basis for realizing large-capacity bandwidth services. In the constantly crowded spectrum environment, spectrum allocation aims to improve the spectral efficiency, coordinate the distribution and deal with the growth of wireless data traffic. A variety of evolutionary algorithms such as genetic algorithm, particle swarm optimization and ant colony optimization have been widely used in dynamic spectrum allocation [22-24]. [25] analyzed the characteristics and application scenarios of various reinforcement learning algorithms for dynamic spectrum allocation. To deal with the problem of cooperative spectrum sensing, an framework called deep cooperative sensing is proposed, which trains the sensing samples using CNN instead of explicit mathematical modeling [26]. Spectrum scarcity problem is also urgent for application prospects of the internet of things (IoT), due to the demand of connecting the things widely. [27] proposed a deep recurrent neural network (RNN) based algorithm to provide optimal resource allocation results for the NOMA heterogeneous IoT.

4. Challenges in AI Deployment

4.1. Suitable Data for AI

Mobile communication systems are generating large amounts of data at any time, but most of these data will soon disappear. At the same time, mobile communication data cannot be directly used in the AI framework and needs to be converted, such as from one-dimensional to multi-dimensional, from complex to real, and from fragmented to structured. Those factors lead to the fact that data that is truly suitable for AI processing is still very scarce.

4.2. Realtime Performance

The mobile communication system is hierarchical. The closer to the physical layer of communication, the higher the requirement for real-time performance, and the closer to the upper layer, the higher the tolerance for delay. Therefore, for different layers, it is necessary to fully evaluate the application method of AI, whether it is online learning or offline learning, complex calculations or simplified calculations, otherwise it will not be able to meet the requirements of low-latency scenarios.

ISSN: 2688-9323

4.3. Computing Capability and Power Consumption

AI algorithms, especially deep learning, usually have high computational complexity and require huge memory space to store data, which poses challenges to computing power. The equipment and hardware located in the convergence center or cloud computing center are usually capable, but for some edge computing nodes or mobile terminals, they will face computing power problems. At the same time, AI computing will bring huge challenges to energy consumption. Especially in the next-generation network, base stations or terminals located in space, air, and water face the problem of energy supply, so some strategies are needed to meet the energy management needs of AI applications.

4.4. Adaptability of AI Model

With the introduction of new base stations on platforms such as street lamps, UAVs, and satellites, the mobile communication network is moving towards a heterogeneous network. The differentiation of the network environment becomes more and more obvious. At the same time, devices or terminals that access the network may have different QoS requirements. Those may lead to an invalidation, by applying a single AI model to complex and dynamic networks and ensure robustness. Thus, the adaptability and differentiated deployment of AI models is a critical concern.

4.5. Security Risk

The in-depth application of AI in network slicing, edge computing, and physical layer will bring new security threats and risks to a certain extent, and put forward higher requirements for data protection, security protection, and operational deployment. Just as image spoofing can cause problems in autonomous driving, the implantation of malicious data or the tampering of algorithm models will also make autonomous networks very risky. Therefore, people have to face new forms of security in the future AI-driven mobile communication systems.

5. Conclusion

This paper gives a survey and discussion of AI for mobile communication network. Four types of AI components, named central AI, edge AI, access AI, terminal AI are deployed in different locations of the network, to provide a powerful engine for different network functions, and finally constitutes a complete AI architecture. Furthermore, several specific examples of AI frameworks and algorithms for mobile communication network are presented, showing significant improvements or enhancements in network slicing, mobility management, end-to-end optimization, etc. Although AI will bring an evolution to mobile communication network, some challenges must to be faced and overcome, such as data source, delay, and power consumption. Only when those problems are solved, the efficiency of AI can be fully exerted in the mobile communication network, so as to provide more flexible and convenient services.

References

[1] J. Nightingale, P. Salva-Garcia, J. M. A. Calero and Q. Wang. 5G-QoE: QoE Modelling for Ultra-HD Video Streaming in 5G Networks, IEEE Transactions on Broadcasting, vol. 64 (2018), p. 621-634.

[2] Y. Dai, D. Xu, S. Maharjan, et al. Artificial Intelligence Empowered Edge Computing and Caching for Internet of Vehicles. IEEE Wireless Communications, vol. 26 (2019), p. 12-18.

[3] Q. Liu, J. Yang, C. Zhuang, A. Barnawi and B. A Alzahrani. Artificial Intelligence Based Mobile Tracking and Antenna Pointing in Satellite-Terrestrial Network, IEEE Access, vol. 7 (2019), p. 177497-177503.

[4] X. Shen, J. Gao, et al. AI-Assisted Network-Slicing Based Next-Generation Wireless Networks," IEEE Open Journal of Vehicular Technology, vol. 1 (2020), p. 45-66.

[5] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs and X. Costa-Perez. Network Slicing Meets Artificial Intelligence: An AI-Based Framework for Slice Management, IEEE Communications Magazine, vol. 58 (2020), p. 32-38.

[6] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo and J. Zhang. Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing, Proceedings of the IEEE, vol. 107 (2019), p. 1738-1762.

[7] M. Chen, W. Li, G. Fortino, et al. A Dynamic Service Migration Mechanism in Edge Cognitive Computing. ACM Transactions on Internet Technology, vol. 19 (2019), p. 1-15.

[8] C. Wang, Z. Zhao, Q. Sun and H. Zhang. Deep Learning-Based Intelligent Dual Connectivity for Mobility Management in Dense Network, 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), 2018, p. 1-5.

[9] Z. Wang, L. Li, Y. Xu, et al. Handover Control in Wireless Systems via Asynchronous Multi-User Deep Reinforcement Learning, IEEE Internet of Things Journal, vol. 5 (2018), p. 4296-4307.

[10] Y. Zhou, C. Shen, M. van der Schaar. A non-stationary online learning approach to mobility management. IEEE Transactions on Wireless Communications, vol. 18 (2019), p. 1434-1446.

[11] M. Ebada, S. Cammerer, A. Elkelesh and S. t. Brink. Deep Learning-Based Polar Code Design." 57th Annual Allerton Conference on Communication, Control, and Computing, 2019, p. 177-183.

[12] E. Nachmani, Y. Be'ery, D. Burshtein. Learning to decode linear codes using deep learning, Proc. 54th Annu. Allerton Conf. Commun. Control Comput. (Allerton), 2016, p. 341-346.

[13] T. Gruber, S. Cammerer, J. Hoydis, and S. ten Brink. On Deep Learning-based Channel Decoding, Proc. Information Sciences and Systems (CISS), 2017, p. 1-6.

[14] H. Huang, J. Yang, H. Huang, Y. Song, G. Gui. Deep Learning for Super-Resolution Channel Estimation and DOA Estimation Based Massive MIMO System, IEEE Transactions on Vehicular Technology, vol. 67 (2018), p. 8549-8560.

[15] S. N. Motade and A. V. Kulkarni. Channel Estimation and Data Detection Using Machine Learning for MIMO 5G Communication Systems in Fading Channel, Technologies, vol. 6 (2018), p. 72.

[16] J. Liu, K. Mei, X. Zhang, D. Ma, and J. Wei. Online Extreme Learning Machine-Based Channel Estimation and Equalization for OFDM Systems, IEEE Communications Letters, vol. 23 (2019), p. 1276-1279.

[17] T. Cousik, R. Shafin, Z. Zhou, K. Kleine, J. Reed, and L. Liu. CogRF: A new frontier for machine learning and artificial intelligence for 6G RF systems, arXiv preprint, 2019, arXiv:1909.06862.

[18] A. Alkhateeb, S. Alex, P. Varkey, et al. Deep learning coordinated beamforming for highly-mobile millimeter wave systems. IEEE Access, vol. 6 (2018), p. 37328-37348.

[19] Q. Wu, Y. Cao, H. Wang, et al. Machine-learning-assisted optimization and its application to antenna designs: Opportunities and challenges, China Communications, vol. 17 (2020), p. 152-164.

[20] T. O'Shea, J. Hoydis. An introduction to deep learning for the physical layer. IEEE Transactions on Cognitive Communications and Networking, vol. 3 (2017), p. 563-575.

[21] V. Raj, S. Kalyani. Backpropagating through the air: Deep learning at physical layer without channel models. IEEE Communications Letters, vol. 22 (2018), p. 2278-2281.

[22] M. Liu, H. Zhang, R. Fan, et al. The GA solution of dynamic Spectrum allocation in cognitive radio based on collaboration and fairness, 2011 Third Pacific-Asia Conference on Circuits, Communications and System (PACCS), 2011, p. 1-4.

[23] P. M. Pradhan, G. Panda. Comparative performance analysis of evolutionary algorithm based parameter optimization in cognitive radio engine: A survey, Ad Hoc Networks, vol. 17 (2014), p. 129-146.

[24] M. Yang, and J. P. An. An Ant Colony Optimization Algorithm for Spectrum Assignment in Cognitive Radio Networks, Journal of Electronics & Information Technology, vol. 33 (2011), p. 2306-2311.

[25] Y. Wang, Z. Ye, P. Wan, et al. A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks, Artificial Intelligence Review, vol. 51 (2019), p. 493-506.

ISSN: 2688-9323

[26] W. Lee, M. Kim, and D. H. Cho. Deep Cooperative Sensing: Cooperative Spectrum Sensing Based on Convolutional Neural Networks, IEEE Transactions on Vehicular Technology, vol. 68 (2019), p. 3005-3009.

[27] M. Liu, T. Song, and G. Gui. Deep Cognitive Perspective: Resource Allocation for NOMA-Based Heterogeneous IoT With Imperfect SIC, IEEE Internet of Things Journal, vol. 6 (2019), p. 2885-2894.