Deep Reinforcement Learning for Resource Allocation in 5G Communications

Mau-Luen Tham*, Amjad Iqbal* and Yoong Choon Chang*
*Department of Electrical and Electronic Engineering
Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman (UTAR), Malaysia
E-mail: thamml@utar.edu.my, amjad.iqbal68@1utar.my and ycchang@utar.edu.my

***Abstract*— The rapid growth of data traffic has pushed the mobile telecommunication industry towards the adoption of fifth generation (5G) communications. Cloud radio access network (CRAN), one of the 5G key enabler, facilitates fine-grained management of network resources by separating the remote radio head (RRH) from the baseband unit (BBU) via a high-speed front-haul link. Classical resource allocation (RA) schemes rely on numerical techniques to optimize various performance metrics. Most of these works can be defined as instantaneous since the optimization decisions are derived from the current network state without considering past network states. While utility theory can incorporate long-term optimization effect into these optimization actions, the growing heterogeneity and complexity of network environments has rendered the RA issue intractable. One prospective candidate is reinforcement learning (RL), a dynamic programming framework which solves the RA problems optimally over varying network states. Still, such method cannot handle the highly dimensional state-action spaces in the context of CRAN problems. Driven by the success of machine learning, researchers begin to explore the potential of deep reinforcement learning (DRL) to address the RA problems. In this work, an overview of the major existing DRL approaches in CRAN is presented. We conclude this article by identifying current technical hurdles and potential future research directions.**

***Keywords*—Deep Reinforcement Learning, 5G, Resource Allocation, Cloud RAN**

I. INTRODUCTION

Recent years have witnessed the great evolution in mobile communications, which began in 1980s with the first generation (1G), followed by 2G (1990), 3G (2002), 4G (2010) and the upcoming 5G [1]. The International Telecommunication Union Radiocommunications Standardization Sector (ITU-R) has standardized the ambitious 5G requirements, referred to as International Mobile Telecommunications 2020 (IMT-2020) [2], which encompasses 100 Mb/s user experienced data rate, one-ms latency, mobility up to 500 km/h, and backward compatibility to long term evolution (LTE)/LTE-A. Such design goals stem from the fact that the total mobile data traffic will increase significantly to 69 Exabytes per month in 2022 [3], due to the unprecedented growth of Internet of Things (IoT) devices. Under this premise, it is obvious that telecommunication operators need to take into account the costs from commoditization and quality of service (QoS) for mobile users during the initial phase of 5G deployment.

This work is supported by the Universiti Tunku Abdul Rahman under UTARRF (IPSR/RMC/UTARRF/2017-C2/T08).

In traditional radio access network (RAN) deployment, each base station (BS) is physically attached with a fix number of antennas, which handles baseband processing and radio functions within small coverage. Accommodating higher transmission rates means that a massive number of physical BSs must be installed. This, however, incurs high initial investment, site support, system management, setup and wireless channel interference among users [4]. Cloud RAN (CRAN), a new paradigm for 5G communications, distributes a set of low-power antennas known as remote radio head (RRH) geographically at distinct locations within the coverage area [5-6]. All RRHs are then connected to a centralized control and processing station known as baseband unit (BBU) via a high-speed front-haul link. Consequently, RRHs are able to coordinate with each other and expand the cellular network coverage. This transforms into supportive channel conditions and ultimately excellent QoS for all user equipments.

Inspired by the CRAN advantages, resource allocation (RA) for CRAN has been extensively investigated. Data rate-oriented optimization problems for multi-user CRAN were treated in [7] and [8]. The work in [9] studied the energy efficiency (EE) maximization problem of CRAN subject to individual antenna power constraints. It, however, does not consider any requirement of data service rates, which is important for provisioning heterogeneous multimedia services. In [10], an EE maximization problem has been formulated under the constraints on per-antenna transmission power and proportional data rates among user equipments. The aforementioned RA schemes rely on numerical methods to optimize various performance metrics. Specifically, techniques such as Charnes-Cooper transformation (CCT), Lagrange Dual decomposition, parameterized convex program, and bi-level optimization are utilized to reach optimality in every single time slots.

Most of the abovementioned RA works can be defined as *instantaneous* since the optimization decisions are derived from the current network state without considering past network states. This may lead to suboptimal results from the perspective of long-term network performance. For instance, pursuing instantaneous energy efficiency may yield to unnecessary turning ON / OFF of RRHs, which is associated with enormous power and timing overheads [11]. Such issue exhibits same flavor with the well-known *ping-pong* effect in

5G handover scenario [12]. Classical work in [13] has demonstrated the possibility of incorporating long-term optimization effect into these RA solutions via utility theory. However, the growing complexity and heterogeneity of network environments has rendered the RA solution intractable.

Reinforcement Learning (RL), a dynamic programming method, has been regarded as one of the promising candidates in optimizing long-term utility of resource allocation [14]. The popularity arises from the fact that several networking problems can be modeled as Markov Decision Processes (MDPs), where RL becomes relevant. RL is about training an agent which interacts with its state-changing environment. The RL agent repeatedly interacts with the environment and collects reward as evaluation. At each time slot, the RL agent chooses an action, which will affect the environment state. Q-learning, one of the widely used RL algorithms, learns a policy which tells the RL agent to select actions that potentially maximizes the expected value of the total cumulative reward, starting from the current state. The work in [15] has adopted Q-learning optimization for ON or OFF policy of BSs with the goal of maximizing EE while meeting user rate demands. Specifically, a Q-table containing rewards for all feasible state-action pairs is constructed. The challenge lies in the convergence of Q-table when the state-action space grows. Consequently, a range of specific CRAN problems poses severe memory and computational challenges to Q-learning based problems.

A major step forward from constructing Q-table of all feasible state-action pairs exhaustively is the adoption of neural networks for estimating Q-table value, which can be denoted as deep Q learning (DQN) [16]. Broadly speaking, deep reinforcement learning (DRL) indicates a set of algorithms that estimates value functions (DQN) or policy functions (policy gradient method) via deep neural networks (DNN). In this way, DRL enables RL to scale to CRAN RA problems that were earlier deemed infeasible. This is aligned with the trend that recent success of machine learning has drawn unparalleled research interest in integrating versatile machine intelligence into 5G [17]. In this paper, we present an overview of the DRL-based RA research in CRAN and pinpoint some open research issues.

The rest of the paper is organized as follows. Section II describes the basic principles of DRL-based CRAN. Section III presents an overview of the major approaches to DRL-based RA in CRAN. Section IV features some interesting open research issues and concludes the paper.

II. DRL-BASED CRAN

Fig. 1 depicts a typical DRL-based CRAN scenario, which consists of which consists of a set of R RRHs $\mathcal{R} = \{1,2,\dots,R\}$, a set of U user equipments (UEs) $\mathcal{U} = \{1,2,\dots U\}$ and a set of B BBUs $\mathcal{U} = \{1,2,\dots B\}$ Most of the signal processing is done at the cloud BBU pool , which is connected to all RRHs via optical fiber or other lossless wired connection. The BBU pool acts a DRL agent, which continuously interacts with the environment, takes an action

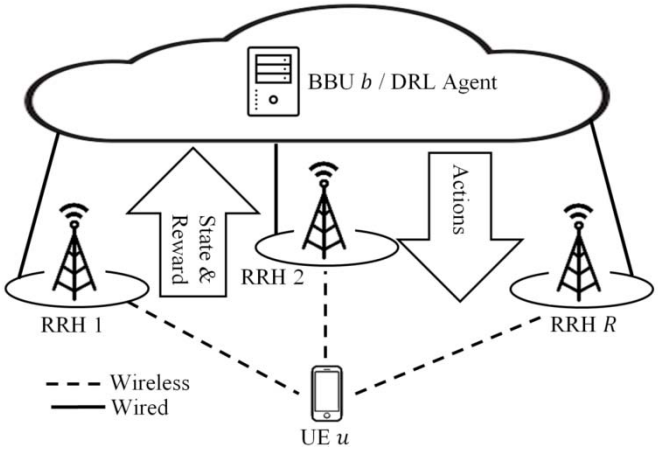

Fig. 1. DRL-based CRAN scenario.

and calculates reward based on past executed actions. Specifically, in each time slot t, based on the current state, BBU pool will make an action a_t in order to maximize the desired objective function. Different BBUs are associated with different computing capabilities and therefore computing task offloading among BBUs may be necessary.

Without loss of generality, let us consider that $B = 1$ and each RRH and UE has one single antenna. Correspondingly, the signal-to-interference-plus-noise ratio (SINR) at UE u can be formulated as [18]:

$$SINR_u = \frac{|h_u^H w_u|^2}{\sum_{v \neq u} |h_u^H w_u|^2 + \sigma^2}, u \in U \tag{1}$$

where $h_u = [h_{1u}, h_{2u}, \dots . h_{RU}]^T$ and each component h_{ru} represents the channel state information (CSI) from RRH r to UE u; likewise $w_u = [w_{1u}, w_{2u}, \dots . w_{Ru}]^T$ and each component w_{ru} stands for the beamforming weight from RRH r to UE u. σ^2 stands for the additive white Gaussian noise (AWGN). The Shannon capacity for user u R_u can be calculated by

$$R_u = Blog_2\left(1 + \frac{SINR_u}{\Gamma}\right), u \in U \tag{2}$$

where B and Γ denotes the transmission bandwidth and realistic capacity gap, respectively. Regarding the power modeling for each RRH, we adopt the model as in [17]:

$$P_r = \begin{cases} P_{r,active} + \frac{1}{\eta} P_{r,trans} & ; \ r \in \mathcal{M} \\ P_{r,sleep} & ; \ r \in \aleph \end{cases} \tag{3}$$

where $P_{r,trans}$ denotes the transmission power of RRH r, adhering to $P_{r,trans} = \sum_{r \in \mathcal{M}} \sum_{u \in U} |w_{r,u}|^2$. η is a constant representing the efficiency of power amplifier. $P_{r,active}$ denotes the power usage of the active RRH, which is critical in sustaining the essential activity of the RRH. The RRH will turn into sleep status whenever it is not chosen for data transmission. Still, it will consume power of $P_{r,sleep}$. $\mathcal{M} \subseteq \mathcal{R}$

and $\aleph \subseteq \mathcal{R}$ stand for the sets of active and inactive RRHs, respectively. Besides that, we take into account the transition power (from active / sleep to sleep / active status). Υ equals the set of mode-transition RRHs in the existing time slot t, which is controlled by the BBU. Armed with the above framework, we can define state and action spaces in the subsequent section.

III. DRL-BASED RA IN CRAN

Generally speaking, DRL consists of two phases namely offline DNN construction phase and online deep Q learning phase [16]. DNN is adopted to estimate the correlation between each state-action match ($\boldsymbol{s}$,$\boldsymbol{a}$) and its value function $Q(\boldsymbol{s}, \boldsymbol{a})$, which is the expected cumulative reward when the environment commences at state s and pursues action $\boldsymbol{a}$. $Q(\boldsymbol{s}, \boldsymbol{a})$ can be formulated as:

$$Q(\boldsymbol{s}, \boldsymbol{a}) = E\left[\sum_{t=0}^{\infty} \mu^t r_t(\boldsymbol{s_t}, \boldsymbol{a_t}) \middle| \boldsymbol{s_0} = \boldsymbol{s}, \boldsymbol{a_0} = \boldsymbol{a}\right] \quad (4)$$

where r_t represents the reward achieved in time slot t, and $\mu \in (0,1]$ is the discount factor which indicates the adjustment between the prompt and future rewards. (4) lies at the heart of most DRL-based RA schemes where different system assumptions, objective functions and optimization variables dictate the specific definitions of $\boldsymbol{s}$, $\boldsymbol{a}$ and r_t. Table I summarizes the existing related works.

In [19], a RL-based offloading strategy has been proposed to choose the RRH and the offloading rate based on the existing battery level, the past data rate to each RRH and the estimated amount of the harvested energy. The authors further accelerate the learning speed based on convolutional neural network (CNN) which compresses the state space. In [20], a double DQN based strategic computation offloading scheme has been designed for ultra-dense sliced RAN. Furthermore, the double DQN is coupled with a Q-function decomposition approach. In [21], a DQN method which uses a DNN to predict the action-value function of Q-learning has been devised to manage the computational resource allocation and offloading decision.

Similar work can be found in [22], where the authors considered both the error probability of decoding and violation probability of delay in order to support low latency communications. In [23], a stepwise RA algorithm that minimizes the total power consumption of CRAN has been proposed. It relies on the combination of DQN and convex optimization to select which RRHs to turn ON and to allocate transmission power among these active RRHs. Such low-complexity algorithm, however, may yield infeasible solution if the number of active RRHs is too low. Furthermore, similar to [19-22], the training process is not in a self-supervised learning mode. The authors in [24] have addressed this issue by proposing Monte Carlo Tree Search (MCTS) algorithm. In MCTS, beginning from a root state, it will mimic routes into the future in order to attain a favorable action by calculating the reward value. Besides that, the work in [24] has improved the traditional DNN by separating the last DNN layers to build a sub neural network for accommodating higher action dimension.

IV. RESEARCH DIRECTIONS AND OPEN ISSUES

We pinpoint issues that remain worthy for further investigation as well as future research.

- The prime challenge discovered in all R&D efforts is the difficulty in searching optimality for the DRL-based problem. A significant portion stems from training process involving the large state-action dimension. Therefore, an effective DRL-based RA should be able to shrink the state-action space by using transfer learning. In this way, it can constantly absorb the features of newcomers and lessens random explorations at the early stage [17]. Stepwise design could be another efficient way to scale down complexity while approaching the optimal system performance. As demonstrated in [23], the continuous action space of dynamic power allocation has been effectively shifted from the MDP to convex optimization.

Table I. Comparison of Existing DRL Based RA Algorithms

Work	[19]	[20]	[21]	[22]	[23]	[24]
Learning Algorithm	CNN+Q-learning	Q-function decomposition + double DQN	DNN+Q-learning	DNN+Q-learning	DNN+Q-learning	MCTS+MLT
Objective Function	Latency	Energy	Sum Cost of Delay and Energy	Task Success Rate	Power	Latency & Energy
Action Space $\boldsymbol{a}$	Offloading Rate & Computation Resource	Offloading Rate & Computation Resource	Binary Offloading & Computation Resource	Computation Resource	Binary On/Off & Power adaptation	Communication Resource, Offloading Rate & Computation Resource
State Space $\boldsymbol{s}$	Battery Level, Renewable Energy generated in a Time Slot & Number of Potential Transmission Rates corresponding to Each Edge Device	Task Queue State, Energy Queue State & Channel Qualities between UEs and RRHs	Computing Capability	Waiting Time of the Tasks to be processed at the Head of Buffers, Queue Length of the Buffers & CSI	User Demand Rate On/OFF of RRHs	Computing Capability, Radio Bandwidth Resource State & Task Request State

- Another issue seldom discussed in most of existing DRL-based RA works is signaling overhead. From the implementation viewpoint, incorporating the signaling overhead into RA problem will be beneficial. The signaling overhead is tightly connected with the accuracy of channel estimation. That is, when fast fading happens, more signaling will be exchanged so that DRL agent can keep up with the CSI. It is still unclear how much performance degradation must be sacrificed when imperfect CSI occurs. Therefore, an effective DRL-based RA should be able to record and preserve historical observations, enabling the DRL agent to excecute accurate CSI prediction, given partial observations. Recurrent neural network (RNN) such as long short-term memory (LSTM) could be one of the promising solutions.

REFERENCES

[1] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu and H. Wei, "5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64-71, June 2017.

[2] J.-C. Guey et al., "On 5G Radio Access Architecture and Technology", *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 2-5, Oct. 2015.

[3] J. Wu, "Green wireless communications: from concept to reality [industry perspectives]," *IEEE Wireless Commun.*, vol. 19, pp. 4–5, Aug. 2012.

[4] X. Wang, "C-RAN: The Road Towards Green RAN," *China Commun. J.*, Jun 2010.

[5] NGMN Alliance 5G White Paper, Mar. 2015, [online] Available: https://www.ngmn.org/5g-white-paper/5g-white-paper.html.

[6] P. Rost et al., "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68-76, May 2014.

[7] V. D. Papoutsis and S. A. Kotsopoulos, "Chunk-based resource allocation in distributed MISO-OFDMA systems with fairness guarantee," *IEEE Commun. Lett.*, vol. 15, no. 4, pp. 377–379, Apr. 2011.

[8] C. He, B. Sheng, P. Zhu, and X. You, "Energy efficiency and spectral efficiency tradeoff in downlink distributed antenna systems," *IEEE Wireless Commun. Lett.*, vol. 1, no. 3, pp. 153–156, Jun. 2012.

[9] C. He, B. Sheng, P. Zhu, X. You, and G. Y. Li, "Energy-and spectralefficiency tradeoff for distributed antenna systems with proportional fairness*," IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 894–902, May 2013.

[10] M.-L. Tham, S. F. Chien, D. W. Holtby, S. Alimov, "Energy-efficient power allocation for distributed antenna systems with proportional fairness", *IEEE Trans. Green Commun. Netw.*, vol. 1, no. 2, pp. 145-157, Jun. 2017.

[11] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs", in *Proc. IEEE ICC,* pp. 1-6, May 2017.

[12] M. Tayyab, X. Gelabert and R. Jäntti, "A Survey on Handover Management: From LTE to NR," *IEEE Access*, vol. 7, pp. 118907-118930, 2019.

[13] C. M. Yen, C. J. Chang, and L. C. Wang, "A Utility-Based TMCR Scheduling Scheme for Downlink Multiuser MIMO-OFDMA Systems", *IEEE Trans. Veh. Technol.*, vol. 59, no. 8, pp. 4105-4115, 2010.

[14] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, Cambridge, MA:MIT Press, 1998.

[15] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Distributed Q-Learning for Energy Harvesting Heterogeneous Networks", *IEEE ICC 2015 workshop on Green Communications and Networks with Energy Harvesting Smart Grids and Renewable Energies*, 2015.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, et al. Human-level control through deep reinforcement learning, *Nature*, Vol. 518, No. 7540, 2015, pp. 529–533.

[17] C. Zhang, P. Patras, H. Haddadi, "Deep learning in mobile and wireless networking: A survey", 2018, [online] Available: https://arxiv.org/abs/1803.04311.

[18] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks", *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, 2016.

[19] M. Min, L. Xiao, Y. Chen et al., "Learning-based computation offloading for iot devices with energy harvesting," *IEEE Trans. Veh. Technol*., vol. 68, no. 2, pp. 1930–1941, Feb 2019.

[20] X. Chen, H. Zhang, C. Wu et al., "Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning," *IEEE Internet Things J*., vol. 6, no. 3, pp. 4005-4018, June 2019.

[21] J. Li, H. Gao, T. Lv et al., "Deep reinforcement learning based computation offloading and resource allocation for mec," in *Proc. IEEE WCNC*, pp. 1–6, April 2018.

[22] T. Yang, Y. Hu, M. C. Gursoy et al., "Deep reinforcement learning based resource allocation in low latency edge computing networks," in *Proc. ISWCS*, pp. 1–5, Aug 2018.

[23] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs", in *Proc. IEEE ICC,* pp. 1-6, May 2017.

[24] J. Chen, S. Chen, Q. Wang, B. Cao, G. Feng and J. Hu, "iRAF: A Deep Reinforcement Learning Approach for Collaborative Mobile Edge Computing IoT Networks," *IEEE Internet Things J*, vol. 6, no. 4, pp. 7011-7024, Aug. 2019.