Low-latency resource elements scheduling based on deep reinforcement learning model for UAV video in 5G network

Jilian Jiang1, Yuhe Qiu1 and Yu Su*, Jian Zhou1

1 R&D Department II, China Mobile Chengdu Institute of Research and Development, Chengdu, Sichuan, 610000, China
*Corresponding author’s e-mail: suyu@cmii.chinamobile.com

Abstract. We consider the problem of resource elements allocation in a network environment with multiple users. Previous studies have done a lot of works using traditional methods in terms of bandwidth allocation, which is sufficient to serve for 4G network. However, it cannot be neglected to provide more efficient and intelligent scheduling policies in haste, due to growing demands on high resolution video and image transmission in 5G network. To fit the condition taking resource elements as scheduling unit in 5G network, we proposed deep Q network (DQN) algorithm based on the requirement of low time latency and high resource utilization rate to solve resource elements (RE) scheduling problem. Ultimately, we give out the optimal allocation scheme of resource elements (RE) for four users in fixed condition of time latency and resource utilization rate.

1. Introduction
Recent decades, with higher and higher resolution of video, from 4k to 8k, large amount of data has been transmitted from end to end, especially with the development of high-resolution live broadcast and images returned from UAV[1]. Apart from that, customers have higher requirements for quality of video viewing experience, such as lower time latency and fewer freezes[2]. As a result, a large amount of data flowing into the service function chain could bring a non-negligible transmission latency owing to limited bandwidth even via 5G network.

In 5G network, Ultra Reliable Low Latency Communication technologies including smaller time slot as scheduling unit and high-efficient channel coding technologies such as Polar codes are mainly used to reduce end-to-end aerial latency[3,4]. Moreover, edge computing and network slicing are introduced to further reduce buffer latency[5], especially when it will introduce a non-negligible task queue waiting latency with heavy load on the network[6]. Accordingly, it is necessary to consider efficient bandwidth resource allocation strategies for wireless network scheduling problem, so as to reduce the user-end video latency while effectively reducing the cost of mobile network operators.

Many researches have studied bandwidth scheduling in past decades, however, few literature has studied its application in 5G network, particularly based on the method of deep learning modelling. Previous existing works have done a lot of works using traditional methods from optimization perspective. To attain the maximum usage of bandwidth according to the standard, traffic nature, and service requirement, it is of great importance to improve on the dynamic bandwidth allocation (DBA) algorithm, and subsequently many improved methods are proposed. In [7], it provides a clear review about the dynamic bandwidth allocation (DBA) algorithm and its application in the Ethernet passive optical network (EPON) and gigabit passive optical network (GPON). And quantity of protocol and algorithms, such as modified HSSR (MHSSR), delay variation guaranteed protocol (DVGP), HSSR...
schemes and dynamic hybrid slot-size bandwidth allocation (DHSSBA), are proposed to meet the requirements for real time communication [8]. In recent works, more and more researchers approved of the potential application of deep learning method in bandwidth scheduling in the background of explosive user volume. And a few works introduce deep learning methods to solve bandwidth allocation and scheduling problems. For instance, a supervised learning-based machine learning algorithm are used to optimize bandwidth allocation decisions to achieve low latency, but it relies on dataset extracted from prepared information [9]. And a research applied a reinforcement learning-based solution to facilitate bandwidth allocation, but it didn’t consider the allocation on time slot [10,11].

Specially, it is an innovative policy that scheduling unit is set as 2, 4, or 7 symbol periods instead of a time slot for 5G technology [12]. It guarantees feasibility of scheduling on frequency and symbol periods. However, current researches mostly just focus on scheduling of frequency [10,13] rather than considering both frequency and time. In this paper, we propose DQN-based scheduling and allocation method on the unit of resource elements which is units of resource block for multi-users in 5G network. We construct the network structure of Q network with three fully-connected feedforward hidden layer for DQN and train the neural network using the replay memory which records states, rewards and actions. And based on the deep reinforcement learning method, a scheme of scheduling for resource elements is given.

2. Bandwidth allocation modelling

To stimulate veritable network with multiple users as possible, the number of total users is determined to be $n = 4$ in our work. And used 5G bandwidth is set as the frequency band of Mobile Networks Operator. The minimum scheduling unit is resource element (denoted as RE) rather than resource block (denoted as RB) composed of 14 symbol periods in 5G network. For an initial network environment for bandwidth allocation modelling, a 100 megahertz-wide band of frequencies will be needed. In Figure 1, the 100M RB, in which horizontal axis corresponding to 14 symbol periods and the vertical axis representing total 3276 sub-carriers, is composed of 45864 RE units. When considering the size of data chunk, we suppose that each user transmits a frame of 1028p picture using B-frame compression method whose compression rate reaching 200:1 and its size is calculated using mathematical expression $\frac{1920 \times 1028 \times 12}{280}$.

To evaluate reward scores of actions from one state to another, the QoS is measured in term of three indexes including time latency, used rate of resource block and used fairness, modelled and calculated as follows. For one user, we could obtain its transmitted latency (denoted as RTT) using the expression (1), in which the size of data chunk is denoted as $D$, and used RE for the $i$th user is denoted as $u_i$. And the used rate of resource block is computed using the expression (2), where $rRE$ means the
rest number of resource elements and $tRE$ means total number of resource elements, as fairness index is calculated by expression (3) to avoid selfish and greedy occupation.

$$RTT_i = \frac{Hi}{Di}, i = 1, \cdots, A$$  \hspace{2cm} (1)
$$ur = 1 - \frac{tRE}{RE}$$  \hspace{2cm} (2)
$$F = \frac{\sum_{i=1}^{A} (RTT_i - RTT)^2}{A}$$  \hspace{2cm} (3)

To meet the QoS requirement, each user ought to adaptively adjust its occupation of resource block based on unified allocation policy with restrictive conditions shown in (4)-(5) with each user selecting resource elements in continuous locations to come into being a rectangle block as possible, where $T$ means total number of resource elements in a resource block.

$$re_i = S^{(\frac{tRE}{n})}_k, k = 1, \cdots, n, i = 1, \cdots, n$$  \hspace{2cm} (4)
$$\sum_{i=1}^{n} re_i \leq T$$  \hspace{2cm} (5)

where $S(\cdot)$ is a stochastic operation which selects an integer in a set of $\{0,1,\cdots,\frac{tRE}{n}\}$ at random.

$$RB(k) = \{re_1(k), \cdots, re_n(k)\}, n = 4$$  \hspace{2cm} (6)

3. A reinforcement learning method for bandwidth resource scheduling

This application proposal provides a new idea of bandwidth resource scheduling, which reasonably allocates and schedules the limited bandwidth resources by combining with the reinforcement learning method. The deep reinforcement learning method, whose principle is to learn the reward of an action from one state to the next state, is applied to allocating bandwidth in defined network environment, including user request, resource limitation, etc. So as to minimize the obtained latency and maximize the utilization of bandwidth resources in the network, in this paper we resort to the deep Q-network (DQN) proposed in [14]. Its training procedure is shown in Algorithm 1 and its overall framework is shown in Figure 2.

All users have to take an action where numbers and locations of resource elements for each user are determined by allocation policy at each time frame. This bandwidth resource scheduling process is essentially a Markov decision process (MDP). The agent makes decision depending on current state $s(k)$ and chooses any action $a(k)$ from a set of actions $A$. The state $s(k)$ is determined by occupancy of resource block, and that is $s(k) = RB(k)$. The action transforms current state $s(k)$ to the next state $s(k + 1)$ and correspondingly obtains a reward $r(k)$ which can be defined as

$$\begin{align*}
r(k) &= 10 \text{ if } RTT_i < t \text{ and } UR > \eta \text{ and } F \leq \epsilon \\
&= 0 \text{ otherwise}
\end{align*}$$  \hspace{2cm} (7)

The core principle of DQN is derived from Q-learning. Q-learning method is to learn a policy $\pi$ by solve the optimization problem that maximizes a discounted reward which is defined as

$$V^\pi(s(k)) = \sum_{i=k}^{T} y^{i-k} r(i)$$  \hspace{2cm} (8)

where $y$ is the discount coefficient and $T$ denotes the terminal time frame reaching the goal state. Then this optimization problem turns into how to get an optimal policy $\pi^*$.?

$$\pi^* = \arg \max_{\pi} V^\pi(s) \forall s$$  \hspace{2cm} (9)

However, mostly in practice it has several intermediate states from start to outcome, thus it is difficult to obtain directly given reward value from one state to the next. Thereby, Q-value is introduced to compute reward in intermediate steps via the iteration rule

$$Q(s,a) = r(s,a) + y \max_{a} Q(s',a)$$  \hspace{2cm} (10)
Unlike the conventional Q-learning method using given action-value table, DQN proposes a deep neural network $Q(s, a, \theta)$ to evaluate Q value for an action from one state to next state. It stores the transfer samples $(s(k), a(k), r(k), s(k + 1))$ obtained by the agent interacting with the environment at each timestep to the replay memory. When it stuffs D with sufficient transition samples, we could randomly take out some samples (ie.256 samples as a minibatch in this paper) to train Q network in training procedure, so as to disrupting correlation among samples. Then the procedure is to copy parameters and network structure of Q network to target $Q'$ network. Q network could be trained by adjusting parameters $\theta$ by minimize the loss between Q network and target $Q'$ network

$$L(\theta) = \frac{1}{|D_k|} \sum_{i \in D_k} (Q(i) - Q(s(i), a(i); \theta))^2$$ (11)

where the $\Omega_k$ denotes the number of samples in a minibatch. And the target Q network is given by the Bellman equation

$$Q'(i) = r(i) + \gamma \max_{a} Q(s(i + 1), a; \theta_0), \forall i \in \Omega_k$$ (12)
4. Experimental results

Based on our proposed DQN-based bandwidth algorithm, we can carry out experiments and obtain experimental results. In our experiment, it is necessary to determine parameters and indexes defined in bandwidth allocation modelling and DQN training. To satisfy the requirements of low time latency and high resource utilization rate as possible, the limitation on used rate of resource block, time latency and fairness index are set to $t = 90\%$, $\eta = 200\text{ms}$, $\alpha = 0.5$. In our experiments, the deep neural network (DQN) used to approximate the action-value consists of three fully-connected feedforward hidden layers, and the number of neurons in the three hidden layer are 256, 256, and 512, respectively. And the ReLU (rectified linear unite) function are applied as activation functions of hidden layers in Q network and
target Q network, and tanh function for the last hidden layer, shown in Figure 2. The training procedure runs when the replay memory is full of 400 transfer samples. The observation stage includes 300 steps, while the exploration step is set to 100000. In observation stage, select an action from the prepared set of actions randomly. Then, in exploration stage, select an action with smaller probability than a certain value $\varepsilon_k$. The value of $\varepsilon_k$ is updated based on $\varepsilon_k = 0.8 \times (1 - \frac{k}{K})$ where $k$ denotes the $k$th step and $K$ denotes total number of training steps.

Based on above constraint and modelling parameters, we construct the structure of DQN with Python program and train the Q network using the constantly produced transfer samples. We record loss one time each 1000 iterations which is computed by expression (11). To illustrate the effect of resource block allocation, we present the scheme of ultimate allocation policy in Figure 3.

5. Conclusion
To sum up, our research can be summarized as follows.

- Our given bandwidth scheduling policy is completely different from the slot scheduling strategy of 4G LTE, the allocation and scheduling based on resource elements at symbol level is realized.
- Otherwise, a deep reinforcement learning algorithm based on limitation of latency, occupancy and fairness is proposed to allocate user resource elements, including bandwidth demand and time slots, which will bring more efficient communication in the scene of video and image transmission from UAVs.
- Apart from that, according to deep learning to predict user bandwidth demand and network load, further our work use reinforcement learning for resource scheduling to make scheduling more automatic, which makes the scheduling process of the whole system more intelligent and fast.

Also, the future work could be undertaken further by combining big data and deep learning methods. We could explore further in bandwidth allocation and scheduling policies using deep learning methods based on user and channel states, which can design simultaneous scheduling of different transmission time interval and different frequency domain resources for different users and different channel states. Accordingly, the bandwidth allocation strategy can be judged according to the user's behavior habits and network channel state. The scheduling could be more diversified and DQN-based bandwidth scheduling will become more flexible, which is suitable for the needs of various types of users.

Acknowledgments
This work receives financial support from China Mobile Chengdu Institute of Research and Development. At last, I gratefully acknowledge the help of my colleagues Qianqian Cheng and Shuijie Wang.

References
[1] Westphal, C., Melodia, T., Zhu, W., & Timmerer, C. (2016). Guest editorial video distribution over future internet. IEEE Journal on Selected Areas in Communications, 34(8): 2061-2062.
[2] Qi wang. (2020) Research on Video Transmission and Resource Scheduling Technology Based on Machine Learning. Chengdu: Jienan Chen.
[3] Won, J. W. , & Ahn, J. M. . (2020) 3GPP URRLC patent analysis. Sciedirect. ICT Express.
[4] Cao, H., Hu, Y., & Yang, L. (2020). Towards intelligent virtual resource allocation in UAVs-assisted 5G networks. Computer Networks, 107660.
[5] Borylo, P., Tornatore, M., Jaglarz, P., Shahriar, N., Cholda, P., & Boutaba, R. (2020). Latency and energy-aware provisioning of network slices in cloud networks. Computer Communications.
[6] Yue Liu. (2018) Research on the resource allocation strategy on low latency scenario. Beijing: Yuehong Gao.
[7] Thangappan, T., Therese, B., Adi, S., & Gadda, S. S. (2020). Review on Dynamic Bandwidth Allocation of GPON and EPON. Journal of Electronic Science and Technology, 100044.

[8] Morshed, M. S., Hossen, M., & Rahman, M. M. (2019). Dynamic hybrid slot-size bandwidth allocation algorithm for reducing packet delay and jitter variation of real time traffic in EPON. Optik, 183, 523-533.

[9] Ruan, L., Dias, I., & Wong, E. (2019, May). Machine Intelligence in Supervising Bandwidth Allocation for Low-latency Communications. In 2019 IEEE 20th International Conference on High Performance Switching and Routing (HPSR) (pp. 1-6).

[10] Ruan, L., Dias, M. P. I., & Wong, E. (2020). Towards self-adaptive bandwidth allocation for low-latency communications with reinforcement learning. Optical Switching and Networking, 100567.

[11] Li, X., Fang, J., Cheng, W., Duan, H., Chen, Z., & Li, H. (2018). Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach. IEEE access, 6, 25463-25473.

[12] Alfadhli, Y., Chen, Y. W., Liu, S., Shen, S., Yao, S., Guidotti, D., ... & Chang, G. K. (2019). Latency performance analysis of low layers function split for URLLC applications in 5G networks. Computer Networks, 162, 106865.

[13] PAN, Su; CHEN, Yuqing. A bandwidth allocation and energy-optimal transmission rate scheduling scheme in multi-services wireless networks. AEU-International Journal of Electronics and Communications, 2018, 95: 97-106.

[14] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. nature, 518(7540), 529-533.