Deep Reinforcement Learning Based Opportunistic Routing for Cognitive Relay Networks

Jintaek Oh, Hyunjoon Suh, Shinhyeok Kang, and Taewon Hwang
School of Electrical and Electronic Engineering
Yonsei University
Seoul, Republic of Korea
Emails: jtoh55@yonsei.ac.kr, hyunjoon.suh@gmail.com, kaci0526@gmail.com, twhwang@yonsei.ac.kr

***Abstract*—In this paper, we consider a cognitive radio (CR) relay network where a secondary relay network and a primary network coexist and share the spectrum. We propose a deep reinforcement learning (DRL) based opportunistic routing (OR) scheme for the secondary relay network to maximize the packet reception probability at a destination via multi-hop relay under the QoS constraint of the primary network. We model the routing problem of the secondary relay network with the QoS constraint of the primary network as a constrained Markov decision process (CMDP). To solve the CMDP, we use a Lagrangian relaxation and obtain an unconstrained MDP. We use deep Q-learning (DQL) to solve the problem. Based on Lagrangian relaxation and DQL, the proposed DRL-based OR scheme finds optimal routing decisions. Simulation results show that the proposed DRL-based OR scheme can improve the packet reception probability of the secondary relay network while satisfying the QoS constraint.**

***Index Terms*—Opportunistic routing, cognitive radio, constrained Markov decision process, deep reinforcement learning**

I. Introduction

The cognitive radio (CR) is a key technology to solve the problem of spectrum scarcity by allowing secondary users (SUs) to exploit the under-utilized licensed spectrum of primary users (PUs). With the emergence of CR network applications such as CR sensor networks, multi-hop routing is becoming essential for wide area coverage. Traditional multi-hop routing schemes first determine a routing path and forward data via the determined path. On the other hand, in opportunistic routing (OR) [1], each node broadcasts a packet and chooses the next forwarder node among the nodes which have actually received the packet. In this way, OR exploits the receive diversity and thus, achieves better performance than traditional routing. An OR scheme for CR multi-hop relay networks has been studied in [2].

Recently, reinforcement learning (RL) based routing schemes have been proposed for CR relay networks. Traditional routing schemes that uses model-based optimization require a lot of prior knowledge, e.g., channel gains among nodes in CR relay networks and PU occupancy probability to obtain the optimal solution. However, in RL an agent tries to find its best policy that maximizes its long-term reward without prior knowledge about the environment. Therefore, there have been many works that study RL-based algorithms for CR and/or relay networks [3]–[6].

In this paper, we propose an deep RL (DRL)-based OR scheme for CR relay networks to maximize the packet reception probability of the secondary relay network while meeting the QoS of the primary network. We model the routing problem as a constrained Markov decision process (CMDP) and use a Lagrangian relaxation to obtain an unconstrained MDP. We solve the unconstrained MDP by using deep Q-learning (DQL) and optimize the dual variable via primal-dual algorithm. The proposed DRL-based OR scheme can be implemented in a decentralized manner because it requires each relay node only to know the Q-functions of its adjacent relay nodes for updating its Q-function. Therefore, it can reduce the signaling overhead among the relay nodes compared with centralized schemes.

II. System Model

As illustrated in Fig. 1, we consider a cognitive relay network where a secondary relay network and a primary network coexist and they share the common spectrum. The primary network consists of a primary-transmitter (PU-TX) and a primary-receiver (PU-RX). The secondary network consists of a source (Src), a destination (Dst), and SU relay nodes, denoted as R_i, $i = 1, \cdots, I$, where I is the number of SU relay nodes. The secondary nodes perform spectrum sensing and determine whether to transmit a packet or not while guaranteeing the QoS of the primary network. The CR relay network operates in a synchronized time-slotted frame structure with a slot duration T. We model the activity of PU-TX in each slot as an independent and identically distributed alternating busy (PU-TX is active) and idle (PU-TX is inactive) process. In each slot, SU nodes perform spectrum sensing and one-hop packet forwarding.

The packet forwarding of secondary relay network is based on OR. In OR, each relay node selects its neighboring nodes, puts them in its forwarder list, and then prioritizes them. The current forwarding node broadcasts its packet to the neighboring nodes and the packet is opportunistically received by some of the nodes in the forwarder list. Then, the next forwarding node is chosen as the one that has the highest

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00253249) and Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0020535, The Competency Development Program for Industry Specialist).

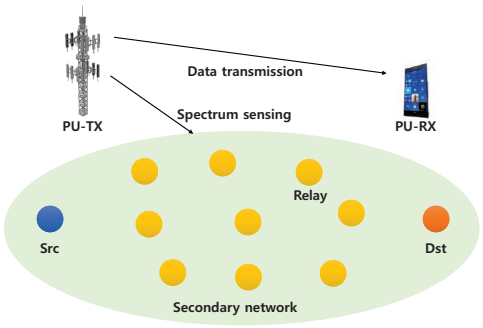

Fig. 1. System model

priority among them. This process is repeated until the packet is forwarded to Dst or dropped after a time limit.

III. Formulating the Routing Problem as CMDP

In this section, we model the routing problem in the CR relay network as a CMDP. We define the CMDP as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, c_{\text{pen}})$. $\mathcal{P}$ is unknown state transition function and other elements of the CMDP tuple is described as follows.

1) State: The state set can be written as $\mathcal{S} = \mathcal{O} \times \mathcal{M} \times \mathcal{N} \times \mathcal{V}$, where $\mathcal{O} = \{o | o \in \mathbb{R}\}$, $\mathcal{M} = \{0, \cdots, M-1\}$, $\mathcal{N} = \{S, D, 1, 2, \cdots, I\}$, and $\mathcal{V} = \{0, 1\}$ are the sets of observations (sensing results) on the PU states, the slot indices, the secondary nodes, and values of a flag indicating success or failure of packet delivery to Dst in the previous slot, respectively. A state (o, m, n, v) indicates that observation is o, the slot index is m, the forwarder node is SU node n, and the flag value is v. Here, $v = 1$ if a successful packet delivery to Dst has occurred in the previous slot and $v = 0$ otherwise.

2) Action Space: Under state $s = (o, m, n, v)$, the action set $\mathcal{A} = \{a | a \in \{0, 1\}\}$, where $a = 0$ means that forwarder node n in the slot index m does not transmit data, while $a = 1$ means that it transmits data.

3) Reward: The immediate reward received after transition from state $s = (o, m, n, v)$ to state $s' = (o', m', n', v')$ due to an action a is given by

$$\mathcal{R}^a_{ss'} = \begin{cases} 1, & v' = 1, \\ 0, & v' = 0, \end{cases} \tag{1}$$

where $v' = 1$ indicates that a successful packet delivery to Dst has occurred in the transition from s to s'.

4) Cost: The immediate cost received after transition from state $s = (o, m, n, v)$ to state $s' = (o', m', n', v')$ due to an action a is given by

$$c_{\text{pen}}(s, a) = \begin{cases} 1, & \text{with prob. } p^{\text{PU}}_{\text{err}}(s, a) \\ 0, & \text{with prob. } 1 - p^{\text{PU}}_{\text{err}}(s, a) \end{cases},$$

where $p^{\text{PU}}_{\text{err}}(s, a)$ is the packet error probability of the primary network under state s and action a. High transmission power of the secondary network can cause interference to the primary network, which can increase $p^{\text{PU}}_{\text{err}}(s, a)$.

In this paper, we find the optimal routing policy that maximizes the packet reception probability, while ensuring the QoS constraint of the primary network. To formulate this problem, we first define the value function and the cost function of as:

$$V_\pi(s_0) \triangleq \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}^{\pi(S_t)}_{S_t S_{t+1}} \middle| S_0 = s_0 \right]$$

and

$$C_\pi(s_0) \triangleq \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t c_{\text{pen}}(S_t, \pi(S_t)) \middle| S_0 = s_0 \right],$$

respectively, where $\pi : \mathcal{S} \to \mathcal{A}$ is the policy, S_t is the state of the MDP at the tth time step, and γ is discount factor. Then, we consider the following problem:

$$\begin{aligned} \max_{\pi} \ & V_\pi(s_0) \\ & C_\pi(s_0) \le p_{\text{th}} \end{aligned} \tag{2}$$

where p_{th} is the threshold of the error probability.

IV. The Proposed DRL-based OR Scheme

To solve the problem in (2), we use a primal-dual algorithm. From [7], we can obtain the optimal policy of the problem in (2) by solve the following dual problem:

$$\inf_{\lambda \ge 0} \sup_{\pi} \mathcal{L}^\lambda_\pi(s_0),$$

where $\mathcal{L}^\lambda_\pi(s_0) = V_\pi(s_0) - \lambda[C_\pi(s_0) - p_{\text{th}}]$ and λ is Lagrange multiplier. Specifically, we can get the optimal policy using one-dimensional search with respect to $\lambda \ge 0$ and the solution of $\sup_\pi \mathcal{L}^\lambda_\pi(s_0)$, which can be obtained by solving the unconstrained MDP $(\mathcal{S}, \mathcal{A}, \mathcal{R} - \lambda c_{\text{pen}})$. Denote $\pi(\lambda)$ to be the solution of the unconstrained MDP with λ, then we can use a gradient-decent-like algorithm to obtain the optimal λ^* as follows:

$$\lambda_{p+1} = \lambda_p + \theta_p \left[C_{\pi(\lambda_p)}(s_0) - p_{\text{th}} \right], \tag{3}$$

where $\theta_p > 0$ is the updating step size.

To solve the unconstrained MDP with λ, we use DQL algorithm in [8]. A centralized implementation of the DRL-based OR scheme requires the central unit to collect global information, i.e., rewards from all the SU nodes, which results in a huge signalling overhead. Therefore, we implement the DRL-based OR scheme in a decentralized manner by utilizing the property of updating Q-function: Each SU node n needs to keep only its Q-function and updates its Q-function by using Q-functions of its adjacent SU nodes.

V. Simulations

In this section, we present the simulation results on the performance of the proposed DRL-based OR scheme. The bandwidth of CR relay network is 10 MHz and the transmission power of the PU-TX is 23 dBm. The active probability of PU-TX is 0.3. The SU relays are uniformly deployed in a grid between Src and Dst. To implement DQL, we use a deep Q-network that has two hidden layers. The activation function is chosen as rectified linear unit (Relu) and the learning rate

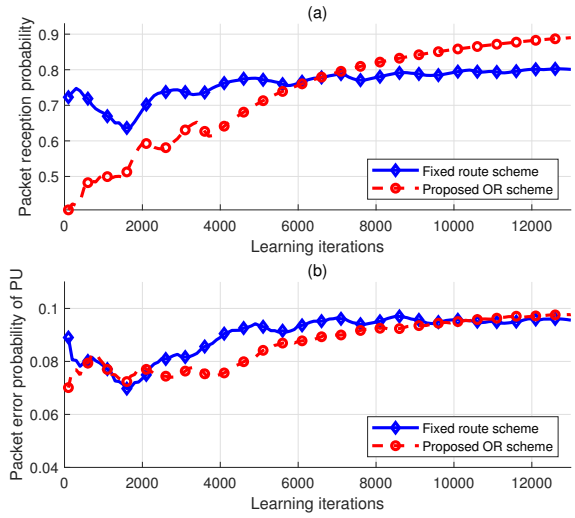

Fig. 2. (a) Packet reception probability, (b) Packet error probability of PU

is set to $\alpha = 0.0005$. Fig. 2-(a) shows the packet reception probabilities of the proposed DRL-based OR scheme and a fixed route scheme versus learning iterations, where the fixed-route scheme first determines the route from Src to Dst before forwarding packets and then controls the packet transmissions of SU nodes by using DRL. From the above results, the packet reception probability of the proposed scheme is higher than that of the fixed-route scheme. Fig. 2-(b) shows the packet error probabilities of PU for the proposed DRL-based OR scheme and fixed route scheme. We set the threshold of the error probability for PU $p_{\text{th}} = 0.1$. Both the proposed DRL-based OR scheme and the fixed route scheme satisfy the QoS constraint of PU.

VI. Conclusions

In this paper, we have proposed the DRL-based OR scheme for CR relay networks to maximize the packet reception probability while guaranteeing the QoS of the primary network. We have modeled the routing problem of the secondary relay network under the QoS constraint of the primary network as a CMDP. We have obtained optimal decisions of the CMDP by employing a Lagrangian relaxation and DQL. The proposed DRL-based OR scheme can significantly improve the reliability and spectral efficiency of CR relay networks by combining receive diversity of OR and spectrum sharing of CR.

References

[1] S. Biswas and R. Morris, "ExOR: Opportunistic multi-hop routing for wireless networks," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 133–144, Aug. 2005.

[2] Y. Liu, L. X. Cai, and X. S. Shen, "Spectrum-Aware Opportunistic Routing in Multi-Hop Cognitive Radio Networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 10, pp. 1958–1968, Nov. 2012.

[3] M. Levorato, S. Firouzabadi, and A. Goldsmith, "A reinforcement learning optimization framework for cognitive interference networks," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Monticello, IL: IEEE, Sep. 2011, pp. 1633–1640.

[4] A. A. Bhorkar, M. Naghshvar, T. Javidi, and B. D. Rao, "An adaptive opportunistic routing scheme for wireless ad-hoc networks," in *2009 IEEE International Symposium on Information Theory*, Jun. 2009, pp. 2838–2842.

[5] Y. Saleem, K. A. Yau, H. Mohamad, N. Ramli, M. H. Rehmani, and Q. Ni, "Clustering and Reinforcement-Learning-Based Routing for Cognitive Radio Networks," *IEEE Wireless Communications*, vol. 24, no. 4, pp. 146–151, Aug. 2017.

[6] A. Paul and S. P. Maity, "Outage Analysis in Cognitive Radio Networks With Energy Harvesting and Q-Routing," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6755–6765, Jun. 2020.

[7] E. Altman, *CONSTRAINED MARKOV DECISION PROCESSES*. Boca Raton, FL, USA: CRC Press, 1999.

[8] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.