논문 21-46-02-07 The Journal of Korean Institute of Communications and Information Sciences '21-02 Vol.46 No.02
https://doi.org/10.7840/kics.2021.46.2.246

# RF 충전 후방산란 인지 무선 네트워크에서 심층 강화학습 기반 모드 최적화 기법

오 선 애•, 권 민 혜*, 김 진 영**, 신 요 안°

# Deep Reinforcement Learning for Mode Optimization in RF-Powered Backscatter Cognitive Radio Networks

Shanai Wu•, Minhae Kwon*, Jin Young Kim**, Yoan Shin°

요 약

RF 충전 후방산란 인지 무선 (Cognitive Radio) 네트워크에서 2차 송신단말 (Secondary Transmitter; ST)은 점유된 1차 채널에 접근하여 에너지를 수집하거나 주변 후방산란을 수행하여 정보를 전송할 수 있으며, 주사용자가 사용하지 않고 비어 있는 1차 채널에 접근하면 수집한 에너지를 사용하여 Active 모드로 데이터를 전송하게 된다. 랜덤하게 변화하는 1차 채널의 점유 상태에 적합한 동작 모드를 수행하면서 전송 성능을 최대화하기 위해, 본 논문에서는 ST가 임의 상태에서 최적의 동작 모드를 학습하는 방안을 제안한다. 또한 에너지 수집 모드를 통해 1차 채널의 점유 상태를 관찰하는 방안을 제안하여 부가적인 에너지 소모를 최소화하며, ST가 순차적으로 동작 모드를 결정하는 문제에 접근할 수 있도록 마르코프 결정 과정을 사용하여 제안하는 기법을 수학적으로 정의한다. ST가 1차 채널과 상호작용 하면서 스스로 최적의 정책을 학습할 수 있는 강화학습을 적용하며, 안정적인 학습을 위해 Deep Q-network 알고리즘을 사용한다. 본 논문에서는 Energy Outage 패널티 (Penalty)를 할당하는 방안을 고려하여 ST가 효율적으로 최적의 정책을 학습하도록 하였다. 학습된 인공신경망을 통해 탐욕 정책으로 동작 모드를 선택하여 수행하는 모의실험을 진행하여 ST가 제안 기법을 통해 달성할 수 있는 전송 성능을 비교 및 분석하였다.

**키워드:** RF 에너지 수집, 인지 무선, 주변 후방산란 통신, 마르코프 결정 과정, 심층 강화학습, 심층 Q-네트워크
**Key Words :** RF energy harvesting, cognitive radio, ambient backscatter communication, Markov decision process, deep reinforcement learning, deep Q-network

ABSTRACT

In an RF-powered backscatter cognitive radio network, the secondary transmitter (ST) harvests energy or backscatters information when the primary channel is busy. Alternatively, the ST actively transmits data using the harvested energy when the primary channel becomes idle. It is critical to decide when to harvest, backscatter, and actively transmit to maximize the throughput of the secondary system under unpredictable primary channel states. In this paper, we propose a deep reinforcement learning-based mode optimization scheme in which the ST can learn the optimal policy through rewards obtained by interacting with the primary

※ 본 논문은 2014년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구결과임 (2014R1A5A1011478).
• First Author : Soongsil University, School of Electronic Engineering, sunae0814@soongsil.ac.kr, 학생회원
° Corresponding Author : Soongsil University, School of Electronic Engineering, yashin@ssu.ac.kr, 종신회원
* Soongsil University, School of Electronic Engineering, minhae@ssu.ac.kr, 정회원
** Kwangwoon University, School of Electronics Engineering, jinyoung@kw.ac.kr, 종신회원
논문번호 : 202012-323-A-RN, Received December 23, 2020; Revised December 30, 2020; Accepted December 30, 2020

channel. To be more specific, the ST is required to perform harvesting mode to observe the reward. We formulate the proposed scheme with a Markov decision process and design a deep Q-network (DQN) for mode optimization. To accelerate the training process, we introduce a penalty for energy outage. The achievable throughput was validated through simulations by considering a greedy policy based on the trained DQN model. It was also compared to the ideal case when the complete information about the primary channel is provided.

## Ⅰ. 서 론

주변의 무선 주파수 (Radio Frequency; RF) 신호를 전기 신호로 변환하는 RF 에너지 수집 기술이 센서 노드와 같은 저전력 (Low-power) 단말의 자기 유지 가능한 (Self-sustainable) 에너지 공급 기술로 부상하고 있다. 특히, 주파수 이용 효율 극대화를 위한 동적 스펙트럼 접근 기술 중의 하나인 인지 무선 (Cognitive Radio; CR) 기술과 결합한 RF 충전 CR 네트워크에 대한 연구가 다양한 방식으로 진행되고 있다. 또한 CR 네트워크에서 RF 에너지 수집 기능을 갖는 2차 송신단말이 (Secondary Transmitter; ST) 1차 채널에 접근하여 기회적으로 에너지를 수집하고 데이터를 전송하는 방안이 [1]에서 제안되었다. 따라서 RF 충전 CR 네트워크에서 ST는 1차 사용자 (Primary User; PU)가 사용하고 있는 채널에 접근하여 에너지를 수집하고, 채널이 비어 있는 동안에 수집한 에너지를 사용하여 기존의 무선 전송 방식으로 데이터를 전송할 수 있다. ST는 데이터를 전송하기 위해 Active 전송을 지원할 수 있는 충분한 에너지와 비어 있는 1차 채널에 접근할 수 있는 기회를 동시에 확보해야 한다. 또한 ST의 에너지 수집 효율은 에너지 전송 채널의 상태, 에너지 변환 효율, 에너지 수집 시간 등 요소와 더불어 PU가 사용하고 있는 채널에 접근할 수 있는 기회에 의해 결정된다. 이와 같이 RF 에너지 수집을 에너지 공급원으로 하는 ST의 전송 성능은 1차 채널 점유 상태의 영향을 크게 받게 된다.

RF 충전 CR 네트워크에서 기회적으로 에너지를 수집하고 데이터를 전송하는 ST의 전송 성능을 향상하기 위해 주변 후방산란 통신 (Ambient Backscatter Communication; AmBC) 기술의 적용이 제안되었다 [2]. AmBC 기술은 주변에 존재하는 RF 신호를 반사하여 정보를 전송하는 장거리 통신 기술로서, 전력소모가 상대적으로 적기 때문에 에너지 수집과 결합하여 효율적인 무선전력 통신 네트워크를 구성할 수 있다[3]. 후방산란 송신회로는 소모전력이 약 0.25 $\mu$W인 Reflective 스위치의 하나인 ADG902 RF 스위치를 사용하여 구현할 수 있으며[4,5], RF 신호를 수신하는 단말로부터 7.2 인치 이상 떨어져 있으면 주변 후방산란을 통해 정보를 전송하는 과정에서 간섭이 발생하지 않는다는 연구결과가 보고된 바 있다[4].

본 논문에서는 RF 충전 CR 네트워크에서 ST의 성능을 개선하기 위해 AmBC의 활용을 고려한다. 후방산란 모드로 동작하기 위해 ST는 안테나에서 일정한 양의 에너지가 반사되도록 임피던스를 제어하여 데이터 비트를 전송한다. 반면에 에너지를 수집하기 위해 ST는 안테나에서 수신한 신호를 흡수해야 하며, 효율적으로 에너지를 충전하기 위해 ST는 에너지 수집 모드와 주변 후방산란 모드를 동시에 수행할 수 없게 된다. 또한 랜덤하게 변화하고 예측하기 어려운 1차 채널에 접근하여 점유 상태에 적합한 동작 모드를 수행하면서 얻는 이득을 최대화하기 위해, ST는 계속하여 동작 모드를 결정해야 한다. 따라서 전송 성능을 최대화하기 위해 ST의 상태를 고려하여 동작 모드를 수행하는 방안에 대한 연구가 필요하다.

본 논문에서는 RF 충전 후방산란 CR 네트워크에서 심층 강화학습 기반 모드 최적화 방안을 제안한다. 강화 (Reinforcement)라는 개념은 미국의 영문학자이자 행동심리학자인 버러스 프레더릭 스키너에 의해 제시되었으며, 강화는 동물이 시행착오 (Trial and Error)를 통해 학습하는 방법 중의 하나이다. 스키너의 쥐 실험에서 쥐는 지렛대를 누르는 행동을 했을 때마다 먹이가 나오는 보상을 알게 되었으며, 이전에 배우지 않았지만 직접 시도하면서 행동과 행동의 결과로 나타나는 좋은 보상 사이의 상관관계를 학습하는 것을 확인하였다. 이와 같이 강화학습은 정답이 주어지는 것도 아니고 주어진 데이터를 학습하는 것도 아닌 환경과 상호작용 하면서 “보상”을 통해 스스로 학습하는 것을 의미한다[6]. 따라서 강화학습의 목적은 환경을 탐험하면서 얻는 보상들의 합을 최대화하는 정책을 학습하는 것이며, 강화학습은 순차적으로 행동을 결정해야 하는 문제에 적용된다[6].

본 논문에서 제안하는 모드 최적화 기법은 Learning 구간에서 랜덤하게 변화하는 1차 채널과 상호작용 하면서 받는 보상을 통해 ST가 최적의 정책을 학습하고, Accessing 구간에서 학습된 인공신경망 모

델로부터 탐욕 정책으로 동작 모드를 선택하면서 최대한 많은 데이터 패킷을 전송하는 것을 목표로 한다. 이를 위해 Learning 구간에서 ST가 에너지 수집 모드를 수행하면서 변화하는 자신의 에너지 상태를 관찰하는 방식으로 환경으로부터 받게 될 보상을 스스로 관찰하는 방안을 제안한다. 따라서 ST는 동작 모드에 따른 보상을 통해 임의 상태에서 최적의 정책을 학습할 수 있으며, 강화학습 알고리즘으로 Deep Q-network (DQN)를 사용한다. 또한 주어진 Learning 구간에서 제안 기법을 효율적으로 학습하기 위해 Energy Outage (EO)가 발생하면 패널티를 부여하는 방안을 고려한다.

## Ⅱ. RF 충전 후방산란 CR 네트워크

본 논문에서는 1차 시스템과 2차 시스템이 각각 한 쌍의 송수신단으로 구성된 RF 충전 후방산란 CR 네트워크를 고려하였다. 네트워크 요구사항에 따라 ST는 다양한 방식으로 1차 채널에 접근 가능한데, 1차 시스템과의 충돌을 회피하고자 채널의 점유 상태에 적합한 모드로 동작하는 기회적 스펙트럼 접근 (Opportunistic Spectrum Access; OSA)과 1차 시스템이 허용 가능한 간섭 범위 이하로 ST가 송신전력을 제어하면서 통신하는 동시 스펙트럼 접근 (Concurrent Spectrum Access; CSA)에 대한 연구가 많이 진행되고 있다[7]. OSA와 CSA는 각각 Overlay 및 Underlay 접근 방식으로도 알려져 있다. 본 논문에서는 ST가 점유된 1차 채널에 접근하여 에너지를 수집하거나 주변의 RF 신호를 후방산란하여 데이터를 전송하며, 1차 채널이 비어 있는 동안에 수집한 에너지를 사용하여 Active 모드로 데이터를 전송하는 Overlay 방식으로 동작하는 방안을 고려한다.

### 2.1 1차 채널 모델

1차 시스템에서 PU는 주파수 대역에 접근할 수 있는 권한을 갖고 있는 사용자이며, 따라서 채널의 상태는 PU의 전송 패턴에 따라 변화한다. 본 논문에서는 타임 슬롯 구간이 $T \in \mathbb{R}^+$인 네트워크 모델을 고려하였으며, 여기서 $\mathbb{R}^+$는 양의 실수 집합을 나타낸다. 따라서, 슬롯 $n$에서 1차 채널은 PU에 의해 사용되거나 ($C_n = 1$) PU가 사용하지 않으면 비어 있게 된다 ($C_n = 0$). 이와 같이 1차 채널은 2개의 점유 상태를 갖기 때문에, 슬롯 $n$에서 $p$의 확률로 사용되고, $(1-p)$의 확률로 비어 있도록 모델링할 수 있다[8].

### 2.2 2차 시스템

RF 충전 후방산란 CR 네트워크에서 ST는 점유된 1차 채널에 접근하여 에너지를 수집하거나 주변에 존재하는 RF 신호를 후방산란하여 데이터를 전송하며, 1차 채널이 비어 있으면 수집한 에너지를 사용하여 Active 모드로 데이터를 전송할 수 있게 된다. 타임 슬롯의 시작점에서 확률 $\lambda$로 발생하는 데이터 패킷은 용량이 $Q$인 데이터 저장장치에 보관되었다가 ST가 전송 기회를 확보하면 순차적으로 SR한테로 전달되며, 발생된 데이터 패킷이 저장장치의 용량을 초과하면 오래된 데이터부터 손실하게 된다. ST는 패킷의 손실을 최소화하고 성공적인 데이터 전송을 위해 1차 채널에 접근하여 점유 상태에 적합한 모드를 결정하여 수행해야 한다.

#### 2.2.1 에너지 수집 모드

ST는 슬롯 $n$에서 1차 채널이 PU에 의해 사용되고 있다고 판단되면, RF 신호를 흡수하여 전기 신호로 변환한 후 다음 타임 슬롯에서부터 데이터를 전송하기 위해 사용할 수 있다. 반면에 후방산란을 수행하기 위해 ST는 안테나에서 일정한 양의 에너지가 반사되도록 임피던스를 제어하여 데이터 비트를 전송하기 때문에, ST는 효율적으로 에너지를 수집하기 위해 수집 모드와 후방산란 모드를 동시에 수행할 수 없다.

RF 에너지 수집기에서의 수신 RF 전력이 감소하면, RF 신호를 전기 신호로 변환하는 효율도 감소하기 때문에 수집 가능한 에너지의 양이 영향을 받게 된다[9]. 또한 RF 공급원의 종류에 따라 송신전력이 상이하여, ST가 수집 가능한 에너지는 RF 공급원의 종류, 에너지 전송 거리 등과 같은 환경적인 요소에 의해 결정된다. 실제로 수 W 수준의 송신 전력으로 안정적으로 신호를 전송하는 TV 또는 FM 라디오 타워와 같은 정적인 RF 에너지 공급원으로부터 최대 수 km 떨어진 곳에서 $\mu$W 수준의 에너지를 수집 가능한 것이 관찰되었으며, 센서 노드와 같은 저전력 단말의 통신을 위해 충분한 에너지를 공급할 수 있게 된다[10,11].

본 논문에서는 ST가 점유된 1차 채널에 접근하여 한 개의 슬롯 $T$ 동안에 평균 $e_h$개의 에너지 유닛 (Unit)을 수집할 수 있다고 가정하였다. 수집한 에너지는 Active 모드로 데이터를 전송하면서 사용하기 위해 용량이 $E$인 배터리에 저장된다.

#### 2.2.2 주변 후방산란 모드

후방산란 송신회로는 다이폴 안테나의 Branch 사이에 트랜지스터로 구성된 스위치를 연결하여 구현할

수 있다. 스위치의 입력은 '1'과 '0'으로 구성된 비트 시퀀스이며, 안테나의 임피던스를 제어하여 많은 양의 에너지를 반사하거나 (Reflective) 적은 양의 에너지를 반사하면서 (Non-reflective) 후방산란 통신을 수행한다. 스위치에 '0'이 입력되면 트랜지스터가 꺼지면서 수신 신호를 흡수하기 때문에 안테나에서 소량의 에너지가 반사되는 반면에, '1'이 입력되면 트랜지스터가 켜지면서 다이폴 안테나의 Branch 사이에서 Short 현상이 발생하여 많은 양의 에너지가 안테나로부터 반사된다.

후방산란 송신회로를 구현하기 위해 Reflective 스위치의 하나인 ADG902 RF 스위치를 사용할 수 있으며, ADG902 RF 스위치의 소모 전력은 약 0.25 $\mu$W이다[4,5]. 또한 ST는 RF 신호를 흡수하면서 데이터 비트 '0'을 전송할 수 있기 때문에 본 논문에서는 후방산란 송신회로에서 소모되는 에너지를 고려하지 않으며, ST는 거의 모든 에너지 상태에서 주변 후방산란을 수행할 수 있다고 가정하였다.

최초의 AmBC 프로토타입 (Prototype)을 제시한 참고문헌 [4]에서는 1 MW의 송신전력을 사용하는 TV 타워로부터 약 4 km 떨어진 곳에 위치한 후방산란 송신단말이 1 kbps의 후방산란 전송률로 동작하면, 실내와 실외에서 각각 1.5 피트, 2.5 피트 떨어진 후방산란 수신단말이 $10^{-2}$의 비트오율을 만족할 수 있음을 검증하였다. 후방산란 수신단말은 전력소모를 최소화하기 위해 Averaging 기법으로 후방산란 신호를 복조하였으며, 이 때 0.54 $\mu$W의 전력을 소모하는 것으로 확인되었다[4].

후방산란 전송거리를 확장하고 후방산란 전송률을 개선하기 위해 수신단에서의 저전력 코딩 기법과 다중 안테나 기법이 제안되었다[12]. 539 MHz TV 신호를 1 kbps의 전송률로 후방산란하는 경우에 수신단에서 저전력 코딩 기법을 사용하여 최대 24 m 떨어진 곳에서 후방산란되는 신호를 수신 가능하며, 수신단에서 다중 안테나 기법을 사용하는 경우에 1 Mbps의 후방산란 전송률로 약 2 m의 전송 거리를 달성할 수 있음을 검증하였다[12]. 또한 저전력 코딩 기법과 다중 안테나 기법을 사용하는 수신단에서 각각 8.9 $\mu$W와 422 $\mu$W의 전력을 소모하는 것으로 확인되었다.

이와 같이 후방산란 전송률은 후방산란 수신회로에 의해 결정되는 것을 알 수 있으며, Closed-form으로 도출하기 어렵게 된다. 따라서 본 논문에서는 ST가 실제로 점유된 채널에 접근하여 후방산란을 수행할 수 있는 성능을 주로 고려하였다. 즉, ST는 주변에 RF 신호가 존재하면 후방산란을 성공적으로 수행하여 정보를 전송할 수 있으며, ST가 후방산란을 통해 한 개의 슬롯 동안에 $d_b$개의 데이터 패킷을 전송할 수 있다고 가정하였다.

#### 2.2.3 Active 전송 모드

ST는 1차 채널이 비어 있다고 판단되면 기존의 Active 무선 전송 방식으로 데이터를 전송할 수 있게 된다. 추가적인 에너지 공급원이 없는 경우에, ST는 RF 에너지 수집 모드를 통해 Active 전송 모드를 지원하기 위한 충분한 에너지를 확보해야만 한다.

참고문헌 [11]에 따르면 960 kW의 송신전력을 사용하는 TV 타워로부터 4.1 km 떨어진 곳에서 평균 60 $\mu$W의 에너지를 수집하였으며, 수집한 에너지를 사용하여 Active 전송 모드로 데이터를 전송하면서 평균 5 kbps의 전송률을 달성하였음을 검증하였다. 또한 참고문헌 [13]은 915 MHz의 RF 신호로부터 에너지를 수집하여 1 Mbps의 전송률로 신호를 전송하면서 98 $\mu$W의 전력을 소모하는 경우에 6 m 떨어진 수신단에서 $10^{-3}$ 비트오율을 얻을 수 있음을 검증하였다.

따라서 Active 전송 성능은 주변의 RF 신호로부터 수집한 에너지를 사용하여 지원할 수 있는 Active 송신전력의 영향을 받게 되는 것을 확인할 수 있다. 또한 ST가 일정 크기의 송신전력을 유지하면서 데이터를 전송하는 동시에 1차 시스템과의 충돌이 발생하지 않으면 SR은 페이딩된 신호를 높은 확률로 복호할 수 있게 된다. 따라서 본 논문에서는 ST가 목표 전송률을 만족시키기 위한 송신전력 조건하에 비어 있는 1차 채널에 접근하여 한 개의 슬롯 동안에 Active 전송 모드로 동작하면서 $d_a$개의 데이터 패킷을 전송할 수 있으며, 이 때 $e_a$개의 에너지 유닛을 사용한다고 가정하였다.

## Ⅲ. 제안하는 강화학습 기반 모드 최적화

그림 1의 제안하는 심층 강화학습 기반 모드 최적화 예시에서 도시한 것과 같이, 제안 기법은 Learning 구간에서 랜덤하게 변화하는 1차 채널과 상호작용 하면서 ST가 받게 될 누적 보상을 최대화하는 동작 모드를 수행하며 최적의 정책을 학습한다. Accessing 구간에서는 학습된 최적의 정책에 따라 행동하면서 최대한 많은 데이터 패킷을 전송하는 것을 목표로 한다. 강화학습을 통해 ST는 예측이 어려운 1차 채널에 대한 정보가 없어도, 상호작용을 통해 받게 되는 보상으로부터 임의 상태에서 최적의 동작 모드를 학습할 수

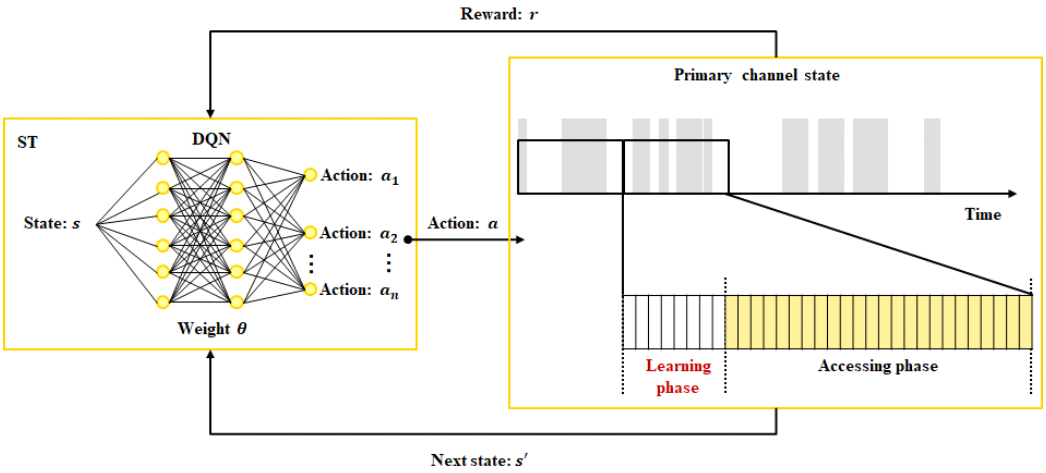


그림 1. 제안하는 심층 강화학습 기반 모드 최적화 예시
Fig. 1. Example of the proposed deep reinforcement learning-based mode optimization

있게 된다. ST는 1차 채널의 점유 상태에 적합한 동작 모드를 수행한 경우에만 보상을 받게 되며, 본 논문에서는 ST가 어떤 동작 모드를 선택해야 할지를 고려하여 앞으로 받게 될 보상을 스스로 관찰하는 방안을 제안한다. ST는 에너지 수집 모드를 수행하면서 변화하는 자신의 에너지 상태를 관찰하여 주변에 RF 신호가 존재하는지를 판단할 수 있다. 따라서 그림 2의 슬롯 구조에서 도시한 것과 같이, 수집 모드를 수행한 후 에너지 상태에 변화가 발생하면 주변 후방산란 모드로 동작하고, 에너지 상태가 변하지 않으면 Active 모드로 데이터를 전송할 수 있게 된다.

ST는 스스로 변화를 감지하면서 임의의 상태에서 최적의 정책을 학습하기 위해 타임 슬롯마다 동작 모드를 결정해야 하며, ST가 이와 같이 순차적으로 동작 모드를 결정하는 문제에 접근할 수 있도록 하기 위해 마르코프 결정 과정 (Markov Decision Process; MDP)를 통해 제안 기법을 수학적으로 정의해야 한다. MDP를 구성하는 요소들로는 상태 공간 (State Space), 행동 공간 (Action Space), 보상 (Reward), 상태전이확률 (State Transition Probability), 감가율

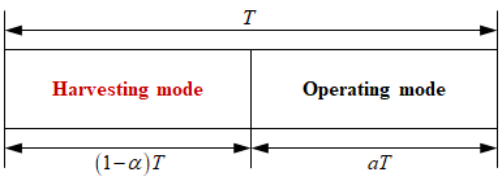


그림 2. 제안하는 수집 모드 기반 관찰 기법의 슬롯 구조
Fig. 2. Slot structure for the proposed harvesting-based reward observation

(Discount Factor) 등이 있다. 본 논문에서는 ST가 동작하면서 관찰할 수 있는 자신의 상태가 MDP의 상태 정보가 되며, 1차 채널에 접근하여 수행할 수 있는 동작 모드가 행동이 된다. 또한 1차 채널의 점유 상태에 적합한 Active 전송과 후방산란을 통해 데이터를 전송하면서 보상을 받을 수 있으며, ST의 상태는 1차 채널이 점유될 확률과 패킷 발생 확률 등에 따라 확률적으로 변화한다. 마지막으로 감가율을 적용하여 ST가 미래에 데이터를 전송하면서 받게 될 보상의 가치를 감가시켜 고려할 수 있게 된다.

### 3.1 제안 기법의 MDP 문제 정의

#### 3.1.1 상태 공간

제안 기법은 한 개의 타임 슬롯을 슬롯 구간이 $T/2$로 동일한 두 개의 서브 슬롯으로 나누어 동작한다. ST는 첫 번째 서브 슬롯 동안에 에너지 수집 모드를 수행하면서 에너지 상태의 변화를 관찰하며, 두 번째 서브 슬롯에서 선택할 동작 모드에 따른 보상을 관찰한다. 즉, ST는 첫 번째 서브 슬롯 동안에 발생하는 에너지 상태의 변화 $\Delta e_n$에 따라, 다음과 같이 상태 값 $o_n$을 결정하게 된다.

$$o_n = \begin{cases} 1, & \Delta e_n \geq e_{th} \\ 0, & \Delta e_n < e_{th} \end{cases}, \tag{1}$$

여기서 $e_{th}$는 ST가 한 개의 서브 슬롯 동안에 수집 가능한 평균 에너지를 의미한다. 본 논문에서는 TV

타워와 같이 일정 크기의 송신전력으로 안정적으로 신호를 전송하는 정적인 RF 공급원을 가정하였기 때문에 $e_{th} \approx 0$이다. 즉, 에너지 상태에 변화가 없으면 1차 채널이 비어 있다고 판단하고, 잔여 에너지가 증가하면 1차 채널이 사용되고 있다고 판단한다. 따라서 ST의 상태 공간을 다음과 같이 정의할 수 있다.

$$S = \{o_n, d_n, e_n\}, \tag{2}$$

여기서 $0 \le d_n \le Q$와 $0 \le e_n \le E$는 각각 슬롯 $n$에서 데이터 저장장치 내의 패킷 개수와 배터리의 잔여 에너지 유닛을 의미한다. 즉, ST는 보상을 관찰한 값과 생성된 데이터 패킷의 수 및 잔여 에너지 상태를 고려하여 동작 모드를 선택하게 된다.

#### 3.1.2 행동 공간

ST는 첫 번째 서브 슬롯에서 1차 채널의 점유 상태를 관찰하기 위해 에너지 수집 모드를 수행하고, 두 번째 서브 슬롯에서 상태에 적합한 동작 모드를 선택하여 수행할 수 있다. 따라서 ST가 선택할 수 있는 동작 모드는 최종적으로 다음과 같게 된다.

$$a = \begin{cases} 0, & \{H, I\} \\ 1, & \{H, H\} \\ 2, & \{H, B\} \\ 3, & \{H, A\} \end{cases}, \tag{3}$$

여기서 $I$, $H$, $B$, $A$는 각각 ST가 행동을 취하지 않을 Idle 모드, 에너지 수집 모드, 주변 후방산란 모드, Active 전송 모드를 나타낸다. 즉, ST는 에너지 수집 모드를 통해 Active 전송에 따른 보상을 관찰하여도 잔여 에너지가 부족하거나 저장된 데이터 패킷이 $d_a$보다 적은 경우에 Idle 모드로 동작하게 되며, 전송해야 하는 데이터 패킷이 $d_b$보다 적으면 Idle 모드를 선택하거나 잔여 에너지가 부족하면 두 번째 서브 슬롯에서 계속하여 에너지 수집 모드를 수행할 수 있다. 따라서 제안 기법에 따른 ST의 행동 공간은 다음과 같이 정의할 수 있다.

$$A_s = \begin{cases} \{H, I\}, & o = 0, \{e < e_a \text{ or } d < d_a\} \\ & o = 1, \{e = E \text{ or } d < d_b\} \\ \{H, H\}, & o = 1, e < E, d < d_b \\ \{H, B\}, & o = 1, d \ge d_b \\ \{H, A\}, & o = 0, e \ge e_a, d \ge d_a \end{cases}. \tag{4}$$

#### 3.1.3 보상

ST는 후방산란 모드와 Active 전송 모드를 통해 데이터를 전송하면서 즉각적인 보상을 받게 된다. 따라서 ST가 동작 모드를 선택하였을 때의 상태에 따라 받게 될 보상을 다음과 같이 정의할 수 있다.

$$R(s,a) = \begin{cases} d_b, & a = 2, o = 1, d \ge d_b \\ d_a, & a = 3, o = 0, e \ge e_a, d \ge d_a \\ 0, & otherwise \end{cases}. \tag{5}$$

이와 같이 ST는 상태에 적합한 동작 모드를 수행하면서 양의 보상을 받고, 상태에 적합하지 않은 동작 모드를 선택하면 불필요한 에너지 소모와 패킷 손실이 발생할 수 있기 때문에 음의 보상을 받으면서 최적의 정책을 학습할 수 있게 된다. 또한 ST는 타임 슬롯마다 동작 모드를 결정해야 하며 행동을 결정하는 시점인 현재에 가까운 보상일수록 더 큰 가치를 가지게 된다. 따라서 0과 1 사이의 값을 가지는 감가율 $\gamma$를 적용하여 추후에 데이터를 전송하면서 받게 될 보상의 가치를 감가시켜 고려할 수 있다.

### 3.2 DQN 알고리즘

본 논문에서는 (4)와 같이 주어지는 제안 기법의 최적 정책을 학습하기 위해 DQN 알고리즘을 사용한다. 대부분 강화학습 알고리즘의 기본적인 아이디어는 상태 $s$에서 행동 $a$가 얼마나 좋은지를 나타내는 $Q(s,a)$ 값을 예측하여 행동에 따른 보상과 감가된 미래의 보상을 최대화하는 행동을 선택하는 것이며, 이는 벨만 (Bellman) 최적 방정식에 기반한 것이다. DQN은 Off Policy인 Q-learning과 인공신경망을 함께 사용하는 방식이며, 탐험하면서 $Q(s,a)$ 값을 업데이트하는 Q-learning과 달리 DQN은 $Q(s,a)$ 값을 근사하여 출력하는 인공신경망의 가중치를 경사하강법 (Gradient Descent)으로 업데이트한다[14]. 또한 DQN은 Q-learning과 마찬가지로 학습하는 정책은 탐욕 (Greedy) 정책을 따른다. 직접 시도하면서 경험으로부터 학습하는 강화학습에서 환경을 충분히 탐험하는 것은 학습 결과에 직접적인 영향을 주는 중요한 문제이며, 따라서 DQN은 $\epsilon$-탐욕 정책으로 행동을 결정한다. 또한 DQN은 경험 리플레이 (Experience Replay)가 가능하고 목표 인경신경망을 사용하여 학습하는 특징을 갖고 있다.

#### 3.2.1 경험 리플레이

DQN은 환경에서 탐험하면서 얻은 $(s, a, r, s')$ 샘

플을 리플레이 메모리 (Replay Memory)에 저장하였다가 학습에 사용하는 경험 리플레이가 가능한 특징을 갖고 있다. 여기서 $s$, $a$, $r$, $s'$은 각각 현재의 상태, 현재 상태에서의 행동, 행동에 따른 보상 및 행동에 따라 전이하게 될 다음 상태를 나타낸다. DQN은 리플레이 메모리에서 샘플을 배치로 랜덤하게 추출하여 인공신경망을 업데이트하기 때문에 샘플들간의 상관관계가 학습에 주는 영향을 완화하였으며, 한 개의 샘플로 학습하는 것이 아니라 배치 (Batch)로 학습하기 때문에 타임 스텝마다 업데이트하여도 안정적인 학습이 가능하다. 본 논문에서는 ST가 타임 슬롯마다 동작 모드를 결정하고, DQN은 타임 스텝마다 가중치가 업데이트되기 때문에 타임 슬롯과 타임 스텝이 동일한 의미를 가진다.

#### 3.2.2 행동 정책

강화학습에서는 정확하지 않을 수 있는 예측한 $Q(s,a)$값에 따라 탐욕적으로 행동하는 것보다, $\epsilon$-탐욕 정책을 따르면서 환경에서 직접 시도하면서 탐험하는 것이 더 큰 의미를 갖는다. 임의 상태에서 $m$개의 행동을 선택할 수 있다고 가정한 경우에 DQN은 $\epsilon$의 확률로 랜덤 정책에 따라 행동을 결정하며, $(1-\epsilon)$의 확률로 탐욕 정책을 따른다.

$$\pi(a|s)=\begin{cases}\frac{\epsilon}{m}+1-\epsilon, & a^*=\mathrm{argmax}_{a\in A}Q(s,a)\\ \frac{\epsilon}{m}, & \mathrm{Random}\ \ a\in A\end{cases}. \quad (6)$$

따라서 Q-learning과 마찬가지로 환경에서 탐험하면서 최적의 정책을 학습할 수 있으며, DQN은 학습을 반복할수록 $\epsilon$의 값을 제어하여 탐험의 확률을 감소시킬 수 있다.

#### 3.2.3 학습 정책

DQN은 인공신경망을 업데이트하기 위해 오류함수 (Loss Function)로 다음과 같이 주어지는 평균제곱오차 (Mean Squared Error; MSE)를 사용한다.

$$L_i(\theta_i)=\left(r+\gamma\max_{a'}Q(s',a';\theta_{i-1})-Q(s,a;\theta_i)\right)^2, \quad (7)$$

여기서 $\theta_i$는 $i$번째 타임 스텝에서의 인공신경망의 가중치를 나타낸다. DQN은 정답이 주어지지 않기 때문에 $r+\gamma\max_{a'}Q(s',a';\theta_{i-1})$가 정답의 역할을 한다. $Q(s,a;\theta_i)$는 학습하는 인공신경망이 예측한 값을 나타내며, DQN은 평균제곱오차를 최소화하는 방향으로 인공신경망의 가중치를 업데이트한다.

#### 3.2.4 목표 인공신경망

DQN은 예측한 값을 목표로 학습하기 때문에 타임 스텝마다 정답이 계속하여 변화하면서 학습에 영향을 주는 것을 방지하기 위해, 가중치 $\theta^-$를 갖는 목표 인공신경망을 따로 구현하여 정답을 제공한다.

$$L_i(\theta_i)=\left(r+\gamma\max_{a'}Q(s',a';\theta_{i-1}^-)-Q(s,a;\theta_i)\right)^2. \quad (8)$$

목표 인공신경망이 제공하는 정답을 통해 학습하는 인공신경망의 가중치를 타임 스텝마다 업데이트하며, 일정 시간 간격으로 목표 인공신경망을 학습하는 인공신경망의 가중치로 업데이트한다.

## Ⅳ. 모의실험 결과

DQN이 제안 기법을 학습하고 학습된 인공신경망으로부터 탐욕 정책으로 행동을 결정하면서 ST가 달성 가능한 전송 성능을 검증하기 위한 모의실험을 진행하였다. 모의실험을 위해 데이터 저장장치의 사이즈와 배터리의 용량을 모두 10으로 고려하였으며, 한 개의 슬롯 동안에 RF 신호로부터 수집 가능한 평균 에너지 $e_h$와 Active 모드로 동작하면서 소모하는 에너지 $e_a$를 모두 1로 고려하였다.

본 논문에서는 ST가 주변 후방산란 모드와 Active 전송 모드를 수행하면서 한 개의 동작 주기 동안에 전송 가능한 데이터 패킷을 설정하기 위해, 앞서 언급한 참고문헌 [4]와 [11]에서 제시한 프로토타입을 통해 얻은 후방산란 전송률과 Active 전송률을 참고하였다. 이와 같은 기존의 연구결과들은 프로토타입을 구현하기 위해 고려한 물리적인 요소들에 따라 후방산란 전송률과 Active 전송률이 상이하지만, 에너지 소모가 상대적으로 많은 Active 전송이 동일 시간 동안에 더 많은 데이터 패킷을 전송할 수 있음을 확인할 수 있다. 따라서 본 논문에서는 ST가 주변 후방산란 모드와 Active 전송을 모드를 통해 각각 1개와 2개의 데이터 패킷을 전송할 수 있다고 가정하였다.

### 4.1 DQN 학습 결과 및 검증

인공신경망을 학습시키기 위해 500개의 타임 스텝이 반복되는 길이가 유한한 에피소드를 고려하였으며, DQN을 설계하기 위한 하이퍼파라미터는 표 1에서 정리하였다. 심층 강화학습을 구현하기 위해 2개의 은

표 1. DQN의 하이퍼파라미터
Table 1. DQN hyperparameters

| Hyperparameter | Value |
|---|---|
| Number of hidden layers | 2 |
| Number of hidden nodes | 24 for each hidden layer |
| Activation function | "ReLU" for input and hidden layers, "linear" for output layer |
| Optimization | Adam optimizer |
| Learning rate, $\alpha$ | 0.0001 |
| Discount factor, $\gamma$ | 0.99 |
| Epsilon, $\epsilon$ | 1 → 0.01 |
| Batch size | 64 |

닉층을 사용하였으며, 각 은닉층을 24개의 노드들로 구성하였다. 입력층과 은닉층은 모두 활성화함수로 ReLU를 사용하였으며, 상태에서 선택 가능한 행동에 대한 $Q(s,a)$값을 출력하기 위해 출력층에서는 선형 함수를 활성화함수로 사용하였다. DQN은 오류함수인 MSE를 최소화도록 가중치를 업데이트하기 때문에 경사하강법으로 동작하는 Adam Optimizer를 사용하였다. 이 때 학습율 $\alpha$는 0.0001로 설정하였다. 감가율 $\gamma$는 0.99로 설정하여 ST가 미래에 받게 될 보상을 큰 비중으로 고려하였다. 학습을 수행하는 초기 단계에 환경을 충분히 탐험할 수 있도록 $\epsilon$은 1로 설정하였으며 DQN이 학습을 진행할수록 $\epsilon$이 0.01까지 감소되도록 구현하였다. DQN은 타임 스텝마다 학습하며, $(s,a,r,s')$ 샘플 한 개로 학습하는 것이 아니라 리플레이 메모리에서 샘플을 배치로 랜덤하게 추출하여 학습하며 배치 사이즈는 64로 설정하였다. ST가 탐험하면서 얻은 샘플들은 리플레이 메모리에 저장하였으며, 메모리 용량이 부족하면 오래된 샘플들부터 삭제된다. 또한 목표 인공신경망은 에피소드가 종료되는 시점마다 학습을 수행하는 인공신경망의 가중치로 업데이트하였다.

본 논문에서와 같이 랜덤하게 변화하는 1차 채널에서 최적의 정책을 학습하고, 학습된 인공신경망으로부터 동작 모드를 선택하면서 일정한 시간 동안에 최대한 많은 데이터 패킷을 전송해야 하는 경우에 학습의 정확도뿐만 아니라 효율적인 학습 방안이 필요하다. 또한 Learning 구간이 길어지면 1차 채널의 점유 패턴이 변화하여 ST가 최적의 정책을 다시 학습해야 하는 문제가 발생할 수도 있다. 따라서 본 논문에서는 사전에 설정한 최소의 경험을 시도한 이후 Energy Outage (EO)가 발생하면 패널티를 할당하는 방안을 고려하였다. 즉, 데이터 패킷이 발생하지 않은 관계로 전송해야 하는 데이터가 없는 경우와 데이터 저장장치의 용량이 부족하여 패킷 손실이 불가피한 경우를 제외하고 Active 전송이 최적의 정책인 아닌 상태에서 선택되면 EO가 발생하게 되며, EO가 발생하면 보상에 비해 상대적으로 큰 패널티를 할당하여 행동의 나쁜 정도를 학습하도록 하였다.

제안된 EO 패널티를 적용한 학습 방안의 효율을 검증하기 위해 500개의 타임 스텝으로 구성된 유한한 에피소드를 500번 반복하면서 학습한 DQN 모델과의 비교를 진행하였다. 결과로부터 제안된 EO 패널티를 적용한 학습을 통해 충분히 학습한 경우에 근접하는 성능을 얻을 수 있도록 DQN이 제안 기법을 효율적으로 학습할 수 있는 것을 확인하였다[15]. 따라서 본 논문에서는 ST가 최적의 정책을 학습하는 Learning 구간을 500개의 타임 스텝을 갖는 유한한 에피소드 200개로 구성하였다.

그림 3은 EO 패널티를 적용하여 DQN을 학습시키는 과정을 도시하였다. ST가 1차 채널에 접근하여 동작 모드를 시도하면서 경험한 샘플에 따라 DQN의 학습 결과가 다르게 나타날 수 있으며, EO 패널티를 적용한 학습 방안은 학습을 위해 비교적 적은 시간을 소모하기 때문에 DQN을 신속하게 다시 학습시킬 수 있다.

그림 4는 타임 스텝이 500인 유한한 에피소드를 500번 반복하면서 EO 패널티를 적용하여 학습한 DQN으로부터 탐욕 정책에 따라 행동하는 ST의 전송 성능을 여러 차례 모의실험을 통해 얻은 결과를 도시한다. 결과로부터 학습된 DQN 모델이 Accessing 구간에서 안정적인 전송 성능을 얻을 수 있는 것을 확인

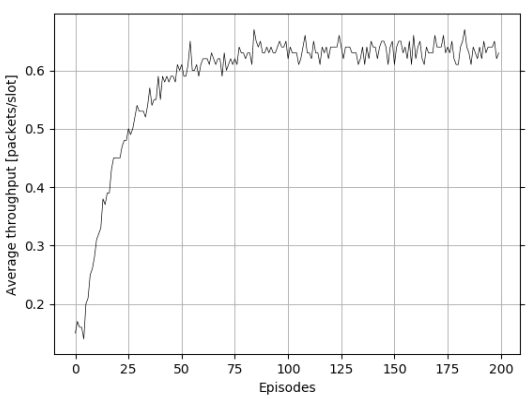


그림 3. Energy Outage 패널티를 적용한 DQN 학습 과정
Fig. 3. DQN training with energy outage penalty

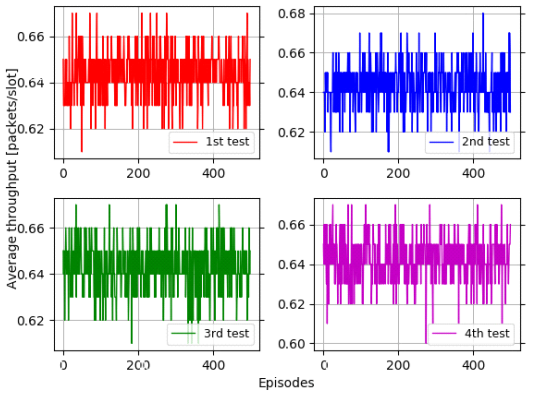


그림 4. 학습된 DQN 모델로부터 탐욕 정책으로 동작하는 ST의 전송 성능 검증
Fig. 4. Average throughputs of ST with greedy policy based on the trained DQN model

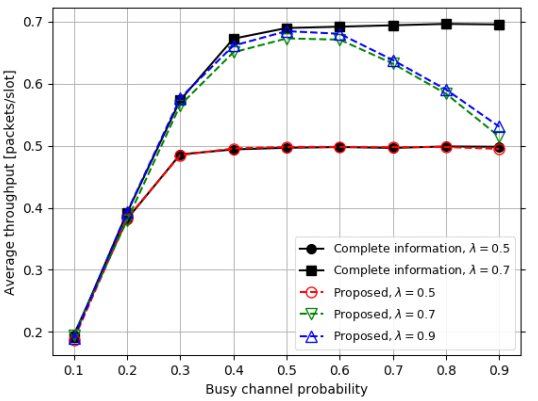


그림 5. 1차 채널이 점유될 확률에 따른 ST의 전송 성능
Fig. 5. Average throughput of ST corresponding to busy channel probability

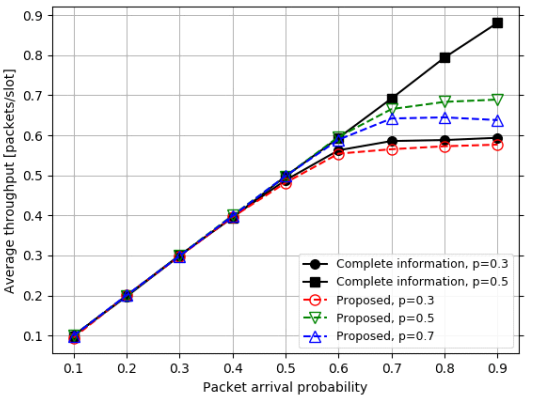


그림 6. 패킷 도착 확률에 따른 ST의 전송 성능
Fig. 6. Average throughput of ST corresponding to packet arrival probability

할 수 있다.

### 4.2 제안 기법의 성능 비교

제안 기법의 성능을 검증하기 위해 1차 채널의 점유 상태에 대한 완벽한 정보를 알고 있는 이상적인 "Complete Information" 경우와의 성능 비교를 진행하였다. Complete Information을 가정한 경우에 1차 채널의 점유 상태에 대한 정보를 획득하기 위해 추가적인 시그널링 (Signalling)과 프로세싱 (Processing) 에너지 소모가 발생할 수 있지만, 본 논문에서는 ST가 제안 기법을 통해 최적에 근접한 Complete Information 대비 달성할 수 있는 전송 성능에 중점을 두어 이와 같은 사항들은 고려하지 않았다.

타임 스텝이 500인 유한한 에피소드를 500번 반복하는 모의실험을 진행하여, EO 패널티를 적용하여 제안 기법을 학습한 DQN 모델로부터 탐욕 정책으로 행동을 결정하는 ST의 전송 성능을 도출하였다. Complete Information을 가정한 경우는 500개의 타임 스텝으로 구성된 에피소드를 500번 반복하면서 학습한 DQN을 사용하여 탐욕 정책으로 행동하면서 전송 성능을 도출하였다.

그림 5와 6은 각각 1차 채널이 점유될 확률과 패킷 발생 확률에 따른 ST의 전송 성능을 도시한다. 결과로부터 알 수 있듯이, 1차 채널이 PU에 의해 점유될 확률이 0.6보다 작거나 패킷 발생 확률이 0.7을 초과하지 않으면 ST는 제안기법을 통해 Complete Information을 가정한 이상적인 경우에 근접한 전송 성능을 얻을 수 있는 것을 확인하였다. 하지만 제안 기법은 동작 모드에 따른 보상을 스스로 관찰하기 위해 타임 슬롯마다 에너지 수집 모드를 필수적으로 수행해야 하기 때문에, 1차 채널이 빈번하게 점유되거나 데이터 패킷이 자주 발생되는 경우 전송 기회가 감소 영향을 크게 받아 전송 성능이 저하되는 것을 확인할 수 있다. 이와 같은 문제를 해결하기 위해 ST가 보유한 에너지가 많으면 Harvest-Then-Transmit 모드로 전환하여, 최소의 에너지만 저장한 후 나머지 에너지를 모두 사용하여 Active 모드를 수행하면서 전송률을 증가시키는 전력 제어 방안을 고려할 수 있다.

## V. 결 론

본 논문에서는 RF 충전 후방산란 CR 네트워크에서 ST가 1차 채널과 상호작용 하면서 받는 보상을 통해 임의 상태에서 최적의 정책을 학습하는 방안을 제안하였다. 1차 채널로부터 받게 될 보상을 스스로 관

찰하기 위해, 본 논문에서는 ST가 에너지 상태의 변화를 감지하여 1차 채널의 점유 상태를 판단할 수 있는 에너지 수집 모드를 활용하는 방안을 제안하였다. ST가 순차적 행동 결정 문제에 접근할 수 있도록 MDP를 통해 제안 기법을 수학적으로 정의하였으며, 최적의 정책을 학습하기 위한 강화학습 알고리즘으로 DQN을 사용하였다. 이 때 심층신경망을 안정적으로 학습시키기 위해 리플레이 메모리를 사용하여 샘플간의 상관관계가 학습에 주는 영향을 완화하였다. 또한 주어진 Learning 구간에서 ST가 효율적으로 최적의 정책을 학습하기 위해 EO 패널티를 할당하는 방안을 제안하였다.

제안 기법을 통해 ST가 달성 가능한 전송 성능을 검증하기 위해, 1차 채널의 완벽한 정보를 알고 있는 이상적인 Complete Information 경우와의 비교를 진행하였다. 제안 기법은 타임 슬롯마다 에너지 수집 모드를 필수적으로 수행하기 때문에, 1차 채널이 높은 확률로 점유되거나 패킷 발생 확률이 상대적으로 높으면 Complete Information을 가정한 경우보다 전송 성능이 저하되는 것을 확인하였다. 하지만 제안 기법은 ST가 에너지 수집 모드를 통해 동작 모드에 따른 보상을 스스로 관찰하므로, 추가적인 시그널링 및 프로세싱으로 인해 발생되는 전력소모를 최소화할 수 있다는 장점을 갖고 있다. 또한 제안 기법은 ST가 스스로 보상을 관찰하고 최적의 정책을 학습하기 때문에 다수 ST가 서로 다른 1차 채널에 접근하는 네트워크 모델로 확장이 용이하다.

## References


[1] S. Lee, R. Zhang, and K. Huang, “Opportunistic wireless energy harvesting in cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4788-4799, Sep. 2013.

[2] D. Hoang, D. Niyato, P. Wang, D. I. Kim, and Z. Han, “Ambient backscatter: A new approach to improve network performance for RF-powered cognitive radio networks,” *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3659-3674, Sep. 2017.

[3] S. Choi and D. I. Kim, “Backscatter communication for wireless-powered communication networks,” *J. KICS*, vol. 40, no. 10, pp. 1900-1911, Oct. 2015.

[4] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, “Ambient backscatter; Wireless communication out of thin air,” in *Proc. ACM SIGCOMM 2013*, Hong Kong, China, Aug. 2013.

[5] *ADG902 RF Switch Datasheet*, available online at https://www.analog.com/media/entechnical-documentation/data-sheets/ADG901_902.pdf

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, 2018.

[7] Y.-C. Liang, K.-C. Chen, G. Y. Li, and P. Mähönen, “Cognitive radio networking and communications: An overview,” *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3386-3407, Sep. 2011.

[8] Y.-C. Liang, Y. Zeng, E. C. Y. Peh, and A. T. Hoang, “Sensing-throughput tradeoff for cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1326-1337, Apr. 2008.

[9] B. Clerckx, R. Zhang, R. Schober, D. W. K. Ng, D. I. Kim, and H. V. Poor, “Fundamentals of wireless information and power transfer: From RF energy harvester models to signal and system designs,” *IEEE J. Sel. Areas in Commun.*, vol. 37, no. 1, pp. 4-33, Jan. 2019.

[10] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, “Wireless networks with RF energy harvesting: A contemporary survey,” *IEEE Commun. Surveys & Tuts.*, vol. 17, no. 2, pp. 757-789, 2nd Quarter 2015.

[11] G. Papotto, F. Carrara, A. Finocchiaro, and G. Palmisano, “A 90-nm CMOS 5-Mbps crystal-less RF-powered transceiver for wireless sensor network nodes,” *IEEE J. Solid-State Cir.*, vol. 49, no. 2, pp. 335-346, Feb. 2014.

[12] A. N. Parks, A. Liu, S. Gollakota, and J. R. Smith, “Turbocharging ambient backscatter communication,” in *Proc. SIGCOMM 2014*, Chicago, USA, Aug. 2014.

[13] Y.-J. Kim, J. S. Bhamra, J. Joseph, and P. P.

Irazoqui, "An ultra-low power RF energy harvesting transceiver for multiple-node sensor application," *IEEE Trans. Cir. & Syst.*, vol. 62, no. 11, pp. 1028-1032, Nov. 2015.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," in *Proc. NIPS 2013*, Lake Tahoe, USA, Dec. 2013.

[15] S. Wu and Y. Shin, "An efficient mode optimization based on reinforcement learning in RF-powered backscatter CRNs," in *Proc. 1st Korea AI Conf.*, Online, Dec. 2020.

**오 선 애** (Shanai Wu)
한국통신학회논문지 vol. 41, no. 9 참조
[ORCID:0000-0001-6572-2838]

**권 민 혜** (Minhae Kwon)
한국통신학회논문지 vol. 41, no. 11, 참조
[ORCID:0000-0002-8807-3719]

**김 진 영** (Jin Young Kim)
한국통신학회논문지 vol. 38C, no. 11 참조
[ORCID:0000-0002-1456-7097]

**신 요 안** (Yoan Shin)
한국통신학회논문지 vol. 41, no. 9 참조
[ORCID:0000-0002-4722-6387]