https://doi.org/10.7840/kics.2024.49.1.88

이기종 무선 네트워크에서 심층 강화학습을 이용한 다중 접근 제어 프로토콜

김 도 원◆, 신 경 섭°

Multiple Access Control Protocol using Deep-Reinforcement Learning in Heterogeneous Wireless Networks

Do-won Kim◆, Kyung-seop Shin°

요 약

최근 이동통신 기기의 종류와 개수가 점차 증가함에 따라 한정된 주파수 대역을 효율적으로 분배할 수 있는 방법이 중요하게 여겨지고 있다. 하나의 네트워크에서 여러 기기가 공존하는 heterogeneous network 환경에서는 기기간의 Multiple Access Control (MAC) 프로토콜이 다르게 적용되기 때문에, 기존의 방식으로는 충돌이 불가피하다. 본 논문에서는 기존의 MAC 프로토콜과 효율적으로 공존할 수 있도록 강화학습을 이용한 MAC 프로토콜을 제안하여 각 기기간의 정보교환이 어려운 경우에도 데이터 전송에서 충돌을 줄일 수 있는 방법을 제안하고, 이종 망 환경에서 성능이 향상됨을 보였다.

키워드 : 이기종 네트워크, 다중 접근 제어, 강화학습, 패킷 충돌
Key Words : Heterogeneous Network, Multiple Access Control, Reinforcement Learning, Packet Collisions

ABSTRACT

With the increasing number and diversity of mobile communication devices, the efficient allocation of limited frequency bands has become a critical concern. In heterogeneous network environments, where multiple devices coexist within a single network, the application of different Multiple Access Control (MAC) protocols to each device leads to inevitable collisions using conventional methods. In this paper, we propose MAC protocol based on reinforcement learning, aiming to achieve efficient coexistence in such heterogeneous networks. By utilizing reinforcement learning, our proposed protocol mitigates collisions in data transmission, even in scenarios with hardness of information exchange between devices and experimental results demonstrate performance improvements in mixed-network environments.

Ⅰ. 서 론

최근 이동통신 기기의 사용량이 증가함에 따라 한정된 주파수 대역을 공유하는 접속 기술이 다양해지기 때문에 기기간 자율적이고 효율적으로 통신 자원을 분배할 수 있는 경쟁방법이 중요해지고 있다. 데이터를 보내기 위해 네트워크 접속 시에는 여러 기기의 전송 신호를 충돌 없이 접속 가능하도록 하기 위해

※ 본 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1064059)
◆ First Author : Sangmyung University, wlsgur479@gmail.com, 학생회원
° Corresponding Author : Sangmyung University, ksshin@smu.ac.kr 정회원
논문번호 : 202307-019-B-SE, Received July 25, 2023; Revised September 13, 2023; Accepted October 7, 2023

multiple access control(MAC) 프로토콜을 사용한다. 비경쟁 기반의 MAC 프로토콜은 주파수 대역을 여러 개의 채널로 나누어 사용자를 분배하는 Frequency Division Multiple Access (FDMA)나 동일 주파수를 시간적으로 분할하여 자신에게 할당된 시간 slot에만 데이터를 전송하게 하는 Time Division Multiple Access (TDMA) 기법, 사용자를 코드 시퀀스로 구분하여 분할한 Code Division Multiple Access (CDMA)가 대표적이다. 또한 경쟁기반의 프로토콜에는 사용자가 무작위로 전송을 시도하여 충돌 감지 및 회피를 위해 충돌 발생 시에 재전송을 시도하는 Random access 기법이 있다. 보낼 데이터 패킷이 있을 때에 비동기 전송을 시도하는 pure-ALOHA기법이나 시간슬롯을 정하여 보내게 되는 slotted ALOHA, 반송파 검출을 통하여 충돌 여부 또는 사용 여부를 확인하는 CSMA/CD, CSMA/CA가 널리 사용되고 있다[1]. 이러한 프로토콜은 동일 프로토콜을 공유하는 다른 기기의 전송을 고려하여 설계하기 때문에 충돌을 의도적으로 회피하거나 재전송을 통한 통신이 가능하다. 하지만 다른 MAC 프로토콜을 이용하는 사용자가 있는 이기종 환경에서는 공존하기가 어려운 부분이 발생할 수 있다.

본 연구에서는 서로 다른 MAC 프로토콜을 채택한 무선 네트워크 환경에서 공통된 스펙트럼을 분배할 수 있는 방법을 모색한다. 공통된 무선 스펙트럼을 공유하는 이기종 네트워크 환경에서 하나의 사용자가 이용가능한 스펙트럼을 독점하는 경우는 이상적인 상황이 아니기 때문에, 사용자는 다른 네트워크 기기의 스펙트럼 점유를 고려하여 통신할 수 있도록 한다[2].

본 논문에서는 MAC 프로토콜 설계를 위해 심층 강화학습을 이용한다. 심층 강화학습은 기존의 강화학습에 신경망을 추가하여 복잡하고 고차원의 문제를 해결하고자 한다. 강화학습은 agent와 환경 사이에서 일어나는 문제에 대해 행동을 선택하고 그에 대한 보상을 최대화하기 위한 학습을 진행하는 프레임워크이다. 여기에 딥러닝 신경망을 도입하여 복잡한 문제에 대해 기존 방법에 비해 뛰어난 성능을 보이는 것이 증명되었다[3, 4]. 우리는 주어진 상황을 판단하여 통신 참여자가 스스로 판단을 내릴 수 있도록 하는 강화학습의 대표적인 알고리즘으로 알려진 deep Q-network (DQN)을 이용하여 기존의 방식과 공존할 수 있는 새로운 MAC protocol을 제안한다.

Ⅱ. 본 론

하나의 네트워크에서 여러 가지의 기기가 데이터 패킷을 전송하는데 많은 충돌과 간섭이 일어난다. 우리의 목표는 각 기기가 효율적으로 전송하기 위하여 상호간의 충돌을 최소화하고, 정해진 MAC 프로토콜을 넘어서 자율적으로 기기가 판단하여 다른 기기와 공존하도록 한다[5].

그림 1에서는 TDMA MAC 프로토콜과 q-ALOHA 프로토콜이 공존하는 네트워크 환경의 예를 보여주고 있다. TDMA는 보낼 데이터 패킷이 존재한다면 정해진 time frame에 따라서 전송을 수행하고 q-ALOHA는 정해진 일련의 확률에 따라서 데이터 패킷을 전송한다. 이 과정에서 다른 네트워크 기기에 대한 고려사항이 존재하지 않기 때문에 충돌이 일어날 확률이 매우 높다. 이러한 충돌을 줄이고 효율적으로 통신을 하기 위하여 우리는 강화학습을 이용하여 MAC 프로토콜을 설계한다.

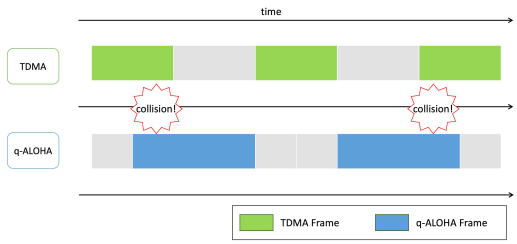

그림 1. TDMA와 q-ALOHA의 패킷 충돌
Fig. 1. Packet Collision between TDMA and q-ALOHA

2.1 Q-learning

강화학습은 agent가 환경을 관찰하고 행동을 선택하여 나온 여러 가지 경험을 통해 보상을 최대화할 수 있는 최적의 행동을 결정한다. 이러한 과정에서 시나리오의 한 step을 episode라 하며, 환경에서 관찰한 상태를 state로 한다. 사용자와 환경은 서로 관찰정보를 주고받으며 다음의 행동을 결정하고, 그에 따른 보상을 얻게 된다. Episode가 진행된 후에 사용자는 이전의 행동과 상태, 보상을 이용하여 학습을 진행하고 시나리오를 반복하게 된다. 이 학습에서의 궁극적인 목표는 얻은 보상의 누적합을 통해 결정할 행동이 다음의 보상을 최대화할 수 있는지를 판단하여 다음의 행동을 결정하는 것이다. 일반적으로 agent는 행동을 선택하는데 필요한 정책 π를 결정하게 되며, 이는 학습을 통한 경험에서 최적의 정책 π^*를 찾게 된다.

Q-learning은 강화학습에서 주로 이용되는 알고리

즘으로 정책 π에 따라 환경 상태에서 agent가 결정한 행동에 따른 누적 보상에 해당하는 행동-가치함수 $Q^{\pi}(s,a)$를 학습하게 된다. 아래의 식 (1)과 같은 행동-가치 함수에서 s는 환경으로부터 관찰한 state 정보를 나타내며, a는 선택한 action을 나타낸다. 이 s와 a를 통해 최적의 value-function Q(s,a)를 학습하여 보상을 최대화한다.

$$Q_{\pi}(s,a) = E_{\pi}[r_{t+1} + \gamma r_{t+2} + \cdots | s,a]. \quad (1)$$

따라서 보상을 최대화 할 수 있는 최적의 행동을 선택할 수 있고 최적의 가치함수를 찾을 수 있다면 Bellman optimality equation에 기반하여 최적의 정책 π^*를 구할 수 있다.

$$\pi^*(s) = \underset{a}{argmax}\, Q(s,a). \quad (2)$$

2.2 Deep Q-network

Q-learning에서 상태-행동 쌍의 수가 많을 때 학습을 진행하게 되면 연산량이 많아지게 된다. 이에 따라 강화학습을 적용할 수 있는 환경이 제한되기 때문에 우리는 Q-network를 이용하여 기존 Q-learning에 신경망을 추가한 DQN을 사용하기로 한다.

DQN은 기본적으로 환경에 대한 정보를 input으로 받아 학습하고 선택할 수 있는 행동의 Q-value를 output으로 하여 보상을 최대화할 수 있는 행동을 선택하도록 네트워크의 연결계수의 값을 학습한다. 이 때, main-network와 target-network의 2가지 신경망을 사용하여 학습으로부터 계산되는 main-network의 예측가중치를 target-network로 복사하여 오차를 역전파 방식으로 업데이트하는 학습 방법이다. 오차는 식 (3)과 같은 방식으로 계산한다.

$$L(\theta) = \left(r + \gamma \underset{a'}{max} Q(s', a'; \theta^{-}) - Q(s,a;\theta)\right)^2, \quad (3)$$

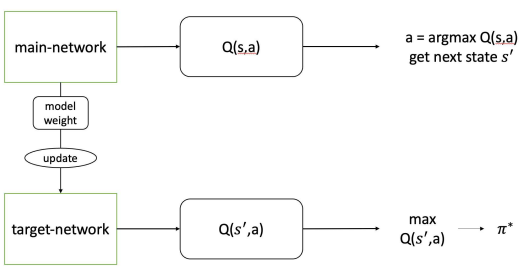

그림 2. DQN 학습 과정
Fig. 2. Training process of DQN

2.3 강화학습 관련 시스템 모델

2.3.1 Agent

Heterogeneous wireless network 환경에서 데이터 패킷을 전송할 수 있는 각 기기를 agent라고 한다. 각 agent는 기지국으로 데이터 패킷을 보내고, 그에 대한 응답으로 ACK를 받아 전송의 성공유무를 판단한다.

네트워크 내에는 다수의 agent가 존재할 수 있으며 각 agent는 개별적으로 행동하기 때문에 전송시도에서 충돌이 일어날 가능성이 있다. 우리가 제안하는 deep reinforcement learning(DRL) MAC 프로토콜은 개별적으로 동작하는 환경에서 자율적으로 판단하여 agent간 충돌을 줄이고 전송 시도를 효율적으로 할 수 있도록 돕는다.

2.3.2 Action

Agent가 선택가능한 행동으로는 데이터 패킷을 전송하는 것과 전송하지 않는 대기상태 두 가지로 나누어 볼 수 있다. 식 (4)에 따라 DRL agent와 other agent는 각각 다른 행동을 선택한다. DRL agent는 신경망을 통한 학습으로 자율적으로 행동을 선택하고 other agent는 미리 정의된 자신의 MAC 프로토콜에 따라 각자 다른 방식으로 행동을 선택하게 된다.

$$a_{DRL},\ a_{other} = \begin{cases} 0 & not\ to\ transmit \\ 1 & to\ transmit \end{cases}, \quad (4)$$

각 agent의 행동을 한 쌍으로 하여 episode가 진행됨에 따른 행동 a_t를 식 (5)와 같이 정의한다.

$$a_t = [a_{DRL},\ a_{other}]. \quad (5)$$

2.3.3 Reward

환경에서 action 정보에 따라 각 agent는 다음의 scalar 값으로 보상을 받을 수 있다. 데이터 패킷을 성공적으로 전송하고 ACK를 수신하였을 때 식 (6)과 같이, 1의 보상을 받을 수 있고, 전송을 시도하였지만 정상적인 ACK 수신이 불가한 경우 충돌로 간주하여 -1의 보상을 받게 된다.

각 agent가 얻은 보상을 묶어 아래의 식 (7)과 같이 reward vector r_{t+1}로 정의하며 N은 agent의 수, n은 agent의 index를 의미한다. 이는 충돌 여부에 의한 보상 값을 포함하여 학습에 사용될 state 정보에 사용한다.

$$r_{t+1}^{n} = \begin{cases} -1 & collided \\ 0 & transmit\ fail \\ 1 & transmit\ success \end{cases}, \quad (6)$$

$$r_{t+1} = [r_{t+1}^{1},\ r_{t+1}^{2},\ r_{t+1}^{3},\ \ldots,\ r_{t+1}^{N}], \quad (7)$$

Heterogeneous 환경을 고려하여 각 r_{t+1}^{n}을 얻는 agent는 다른 MAC 프로토콜을 사용할 수 있고, 이에 따라 각각 다른 보상을 받을 수 있다.

2.3.4 State

State는 observation state와 environment state로 나뉜다. Observation state는 모든 agent가 선택하는 행동의 정보를 모두 고려하여 나타낸다. 표 1은 두 개의 agent가 존재하는 상황에서 observation state, s_o를 정의한다[6].

environment state는 DRL agent의 action a_{DRL}, observation state s_o, reward r_{t+1}의 정보를 모두 포함하는 s_t로, 식 (8)과 같이 나타낸다.

$$s_t = [a_{DRL},\ s_o,\ r_{t+1}] \quad (8)$$

현재 agent가 선택한 action에 대하여 전체적인 observation state s_o와 reward vector r_{t+1}로 state s_t를 구성하여 DQN의 batch memory에 저장된다. 이때, 표 2에서의 state 저장길이에 따라 과거 정보를 사용하여 본 논문 시뮬레이션 환경에서 state는 4×8 (state 저장길이 × s_t의 길이)의 matrix로 이루어져 학습에 사용되게 된다.

Episode의 timestep마다 experience memory에 s_t가 저장되고, mini-batch를 설정하여 experience batch sample $(s_t, a_t, r_{t+1}, s_{t+1})$을 이용하여 학습을 진행한다.

표 1. Network에서 가능한 observation state
Table 1. Possible observation state in network

N	$a_{DRL} = 0, \quad a_{other} = 0$
U	$a_{DRL} = 0, \quad a_{other} = 1$
S	$a_{DRL} = 1, \quad a_{other} = 0$
F	$a_{DRL} = 1, \quad a_{other} = 1$

표 2. 시뮬레이션에 사용된 파라미터
Table 2. Parameters for simulation.

Parameters	Value
node 수	2
state 저장길이	4
memory 버퍼 크기	1000
network update 빈도	200
미니배치 크기	64
총 episode 수	5e-4
discount factor	0.9
learning rate	0.01
optimizer	RMSPprop
backoff window size	4

Ⅲ. 실 험

제안된 MAC protocol의 검증을 위하여 실험은 python으로 진행하였고, 신경망 학습을 위하여 Tensorflow의 Keras 라이브러리를 이용하였다. 신경망 네트워크는 4개의 layer로 구성되며 1개의 GRU layer와 이후 3개의 Dense layer로 구성된다. 사용한 activation function은 relu이며 optimizer로는 RMSProp을 사용하였다. 모델의 구조는 그림 3과 같이 확인할 수 있다. 우리의 알고리즘과 비교로 사용한 알고리즘은 q-ALOHA, TDMA, Fixed Window(FW)이다. q-ALOHA는 평상시에는 0.5의 q-value의 확률로 전송을 시도하지만 일정 시점에서 0 또는 0.2의 q-value를 갖도록 설계하였고, TDMA는 미리 정해진 자신의 time slot에만 전송을 시도하도록 설계하였다. FW는 backoff 시간을 window size와 counter를 기반으로 무작위로 생성한 후, 충돌이 발생할 때 마다 counter를 증가시킨다. Episode가 진행될 때 마다 1씩 감소시키고, backoff 시간이 0이 될 때 전송을 하는 방법으로 설계하였다. 시뮬레이션에 사용된 파라미터는 표 2와 같다. 이기종 네트워크 환경에서 agent는 서로 MAC 프로토콜을 알지 못하는 상태이기 때문에 자신의 상태와 ACK를 통한 observation state를 종합

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 4, 8)]            0

 gru (GRU)                   (None, 64)                14208

 dense (Dense)               (None, 64)                4160

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 4)                 260

=================================================================
Total params: 22,788
Trainable params: 22,788
Non-trainable params: 0
```

그림 3. 학습에 활용된 DQN 모델
Fig. 3. DQN model used for DRL

하여 action을 선택하게 된다.

실험을 통해서 이기종 노드가 서로 충돌을 최대한 피하도록 행동하는 것을 확인하고자 하였으며, q-ALOHA는 일정 주기마다 idle 시간이 존재한다고 가정하였다.

그림 4는 DRL MAC 프로토콜과 Q-aloha 를 비교하였을 때 reward의 누적보상합의 평균값을 나타낸 그래프이다. q-ALOHA가 평균적으로 0.5의 확률로 전송을 하고 reward를 받을 때, DQN은 대부분 0에 근접한 reward를 받는 것을 보여준다. q-ALOHA가 0의 reward를 받는 시점은 전송을 하지 않는 idle 상태로, DQN은 학습을 통하여 최대의 보상에 가깝게 받는 것을 확인하였다.

누적 보상합을 이용하여 network update 빈도수로 이동평균을 구한 후 0-1 사이 값으로 정규화하여 데이터 throughput을 확인하였다. q-ALOHA 노드가 통신을 하는 중에는 DQN 노드가 충돌을 우려하여 전송 시도를 하지 않는 것을 확인할 수 있다. q-ALOHA 노드가 idle 상태로 진입하고자 하면 DQN 노드는 전송을 준비하여 높은 throughput을 보이는 것을 그림 5와 같이 확인할 수 있다.

그림 6은 일정 시간마다 slot에 전송을 시도하는 TDMA와 비교하여 throughput을 나타낸 그래프이다. TDMA는 일정 정해진 시간에 slot에 데이터를 전송하게 되는데, DQN은 그 시점을 학습하고 충돌이 일어나지 않는 시간대에 전송하려는 경향을 확인할 수 있다.

그림 7은 TDMA와 q-ALOHA의 throughput을 나타낸 그래프이다. TDMA는 일정 시간마다 자신이 보낼 데이터가 있다면 전송을 시도하고자 하며, q-ALOHA는 일정 확률에 따라서 데이터 패킷을 전송하게 된다. 이 둘은 서로의 MAC 프로토콜을 알지 못하지만 고려하지도 않기 때문에 서로 충돌횟수가 잦아져 throughput이 충분하게 나오지 않는 것을 확인할

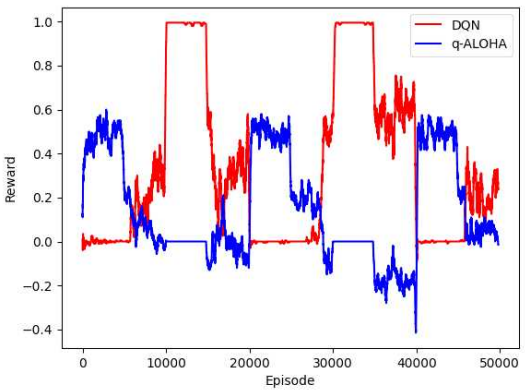

그림 4. DQN과 q-ALOHA의 누적 보상
Fig. 4. Cumulative rewards of DQN and q-ALOHA

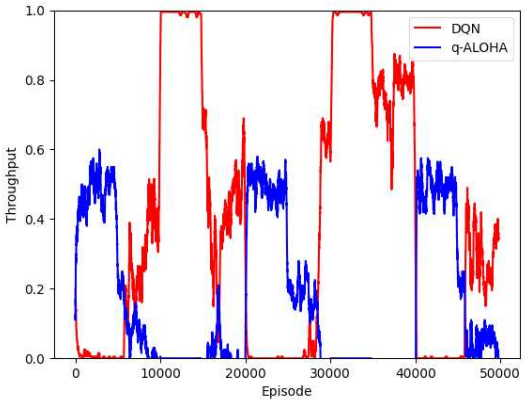

그림 5. DQN과 q-ALOHA의 throughput
Fig. 5. Throughput of DQN and q-ALOHA

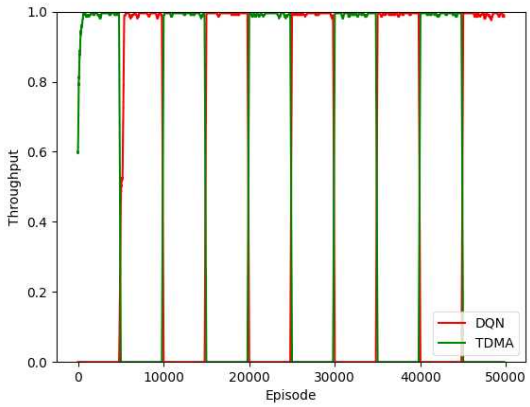

그림 6. DQN과 TDMA의 throughput
Fig. 6. Throughput of DQN and TDMA

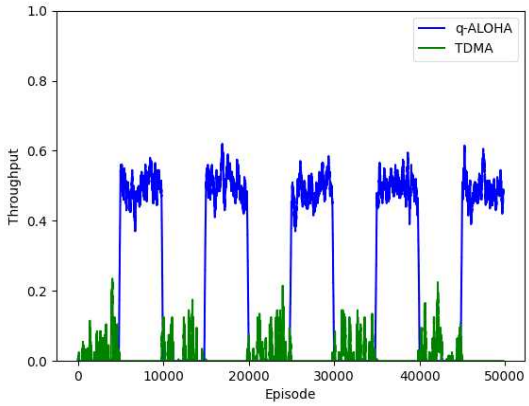

그림 7. q-ALOHA와 TDMA의 throughput
Fig. 7. Throughput of q-ALOHA and TDMA

수 있다.

그림 8은 DQN과 FW의 throughput을 나타낸 그래프이다. FW는 전송을 시도하는데 충돌이 발생한다면 counter 값을 1씩 증가시키며, episode가 진행됨에 따라 backoff 값을 1씩 감소시켜 backoff 값이 0이 된다면 전송을 하게 되는 방법이다. 이에 FW가 충돌이 발생한 시간에 따라 counter를 증가시키며 대기하다가 threshold값에 도달해 전송을 시도하게 되면, DQN은 이를 학습하여 그 시점에서는 전송확률을 낮추게 되는 모습으로 throughput 측면에서 공존하는 모습을 보인다.

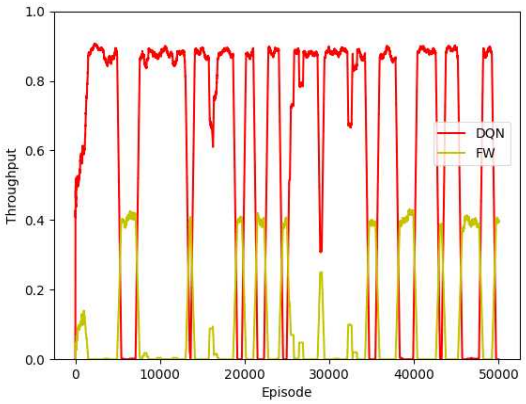

그림 8. DQN과 FW의 throughput
Fig. 8. Throughput of DQN and FW

Ⅳ. 결 론

다양한 기기가 서로 공존할 수 있는 heterogeneous 무선 네트워크 환경에서 우리는 강화학습을 활용하여 기기간 충돌로 인한 throughput 감소를 개선하기 위해 효율적으로 데이터를 전송할 수 있는 MAC 프로토콜을 제안했다. 기존의 방식과의 차별점으로 강화학습을 이용하여 이기종의 MAC프로토콜을 알지 못하여도 ACK에 따라서 전송결과와 네트워크 상태를 확인하고 학습하여 효율적인 전송을 할 수 있도록 하였다. 강화학습으로는 널리 알려진 Q-learning에 Q-network 신경망을 추가한 DQN을 사용했다.

시뮬레이션 결과를 통해 우리가 제안한 DRL 프로토콜은 학습한 이후 다른 기기가 전송을 시도할 때에는 자신이 보내지 않는 경향을 보이고, 다른 기기가 idle 상태인 것으로 판단하면 전송을 통해 throughput을 높이는 모습을 보여주었다. 이를 통해 각 기기가 공평하게 slot을 나누어 사용하며 서로 공존하는 모습으로 강화학습이 MAC 프로토콜의 충돌방지에 도움이 된다는 사실을 확인할 수 있었다.

하지만 본 연구의 한계점으로는 heterogeneous 환경에서의 다른 기기의 MAC 프로토콜이 일정 frame에 고정적으로 전송하거나, 확률적으로 전송하지만 idle 시간이 존재하는 경우를 가정하였기 때문에 다른 기기가 우리의 DRL 프로토콜과 같이 자율적으로 동작하는 경우에는 서로가 greedy하게 전송을 할 우려가 있다. 향후 한 네트워크 내에 다양한 DRL 프로토콜을 지닌 기기가 있는 환경에서 네트워크 활용성을 최대화 할 수 있는 후속 연구를 진행할 계획이다.

References

[1] Behrouz A. Forouzan, *Computer Network*, McGraw Hill Korea, 2012

[2] Y. Xu, et al., “A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges,” *IEEE Commun. Surv. & Tuts.*, vol. 23, no. 2, pp. 668-695, Feb. 2021. (https://doi.org/10.1109/COMST.2021.3059896)

[3] K. Arulkumaran, et al., “Deep reinforcement learning: A brief survey,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26-38, Nov. 2017. (https://doi.org/10.1109/TMC.2021.3057826)

[4] J. Ryu, J. Kwon, and J. Joung, “Timeslot scheduling with reinforcement learning using a double deep q-network,” *J. KICS*, vol. 47, no. 7, pp. 944-952, Jul. 2021. (https://doi.org/10.7840/kics.2022.47.7.944)

[5] Y. Yu, T. Wang, and S. C. Liew, “Deep-reinforcement learning multiple access for heterogeneous wireless networks,” *IEEE J. Sel. Areas in Commun.*, vol. 37, no. 6, pp. 1277-1290, Jun. 2019. (https://doi.org/10.1109/JSAC.2019.2904329)

[6] Y. Yu, S. C. Liew, and T. Wang, “Multi-agent deep reinforcement learning multiple access for heterogeneous wireless networks with imperfect channels,” *IEEE Trans. Mobile Comput.*, vol. 21, no.10, pp. 3718-3730, Oct. 2022. (https://doi.org/10.1109/TMC.2021.3057826)

김 도 원 (Do-won Kim)

2022년 2월 : 상명대학교 컴퓨터과학과 졸업
2022년 9월~현재 : 상명대학교 컴퓨터과학과 석사과정
<관심분야> 통신공학, IoT 네트워크, 인공지능
[ORCID:0009-0003-9155-7890]

신 경 섭 (Kyung-seop Shin)

2009년 1월 : KAIST 전기 및 전자공학과 졸업
2011년 2월 : KAIST 전기 및 전자공학과 석사
2015년 2월 : KAIST 전기 및 전자공학과 박사
2017년 9월~2020년 3월 : 세명대학교 컴퓨터학부 조교수
2020년 3월~현재 : 상명대학교 컴퓨터과학과 조교수
<관심분야> 이동통신, IoT 네트워크, 인공지능
[ORCID:0000-0002-3867-1921]