심층 강화 학습을 활용한 무선 네트워크 환경에서의 동적 주파수 자원 할당에 관한 연구

최시현, 최시영, 김정준, 박세웅
서울대학교 전기정보공학부 뉴미디어통신연구소

shchoi@netlab.snu.ac.kr, sychoi@netlab.snu.ac.kr, jjkim@netlab.snu.ac.kr, sbahk@snu.ac.kr

A Study on the Deep Reinforcement Learning for Dynamic Frequency Resource Allocation in Wireless Network Environment

Sihyun Choi, Siyoung Choi Jungjun Kim and Saewoong Bahk
Department of Electrical and Computer Engineering, INMC, Seoul National University

요 약

본 연구의 목적은 무선 네트워크 환경에서의 동적 주파수 자원 할당에 심층 강화 학습이 적용될 수 있음을 확인하기 위함이다. 이를 위해 서로 다른 mobility 와 Quality of Service 을 가지는 사용자들이 하나의 station 에 연결되어 있을 때, Deep Q-Learning 을 적용하여 시스템 내에서 모든 사용자들의 utility 총합을 최대화하도록 Deep Q-Network 가 주파수 자원을 동적으로 관리하는 것을 학습할 수 있는지 확인한다. Python 을 기반으로 한 시뮬레이션을 통해 이를 검증하였다.

Ⅰ. 서 론

2016 년 AlphaGo 의 등장 이후, 심층 강화 학습을 기존의 주요 공학 문제들에 적용하고자 하는 연구들이 활발하게 이루어지고 있다. 그 이유는 그간 많은 발전을 이루어 왔던 수학적인 optimization 또는 heuristic 한 방법들이 이제는 명확한 한계를 보이고 있고, 이러한 상황에서 심층 강화 학습이 그 돌파구가 될 수 있을 것이라 기대하기 때문이다.

통신에서의 무선 자원 할당을 위한 utility maximization 역시 이러한 맥락에서 심층 강화 학습을 적용할만한 가치가 있는 문제이다. 일반적으로 Quality of Service(QoS) 관점에서 사용자가 체감하는 throughput 에 대한 만족도를 나타내는 utility function 은 non-convex 한 모양을 가지고 있기 때문에 심층 강화 학습을 적용한다면 legacy 방법들보다 성능, 시간 복잡도 그리고 확장성에서 더 좋은 결과를 기대할 수 있다.

본 논문에서는 서로 다른 mobility 와 QoS 을 가지는 사용자들이 하나의 station(STA)에 연결되어 있는 간단한 시나리오를 고려한다. 이러한 시나리오에서 심층 강화 학습의 기본 알고리즘인 Deep Q-Network(DQN)의 학습을 통해 STA 이 시스템 내 모든 사용자들의 utility 총합을 최대화하도록 한정된 주파수 자원을 동적으로 관리할 수 있음을 보였다 [1].

Ⅱ. 본론

1) 심층 강화 학습과 DQN

강화 학습은 agent 가 주어진 environment 속에서 현재 자신의 state(S)을 인식하고 이에 대한 reward(R)을 최대화하는 방향으로 action(A)을 취하는 policy(π)을 찾아나가는 알고리즘이다.

Q-Learning 은 강화 학습에서도 가장 기본이 되는 알고리즘이다. Q-Learning 에서는 Q-Value 라는 값을 정의하는데 이는 어떤 S 에서 어떤 A 을 취했을 때 그 행동이 가지는 가치를 나타내며 다음과 같이 정의된다.

$$Q_\pi(S,A) = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots | S_t, A_t]$$

여기서 t는 time step 그리고 γ는 discount factor 이다.

Q-Value 는 다음의 수식에 따라 매 step 마다 갱신된다.

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_a Q(s_{t+1}, a))$$

여기서 α는 learning rate 이다. 어떤 S 에서 Q-Value 을 최대화하는 A 을 취함으로써 π을 찾아나간다.

Q-Learning 은 Markov decision process 에서 transition probability model 을 알지 못하더라도 environment 가 알려주는 next S 와 next A 만 가지고도 학습을 진행 할 수 있는 model free 알고리즘이다. Unknown environment 에서 episodic data 만을 가지고도 학습이 가능하기 때문에 주변 environment 을 구체적으로 파악하기 어려운 STA 가 주파수 자원을 관리하기에 매우 적합한 알고리즘이다.

DNN 을 이용하여 Q-Value 함수를 근사한 것이 DQN 이다. 이를 통해 복잡한 environment 에서도 Q-Network 가 학습을 할 수 있을 것이라 기대해 볼 수 있다.

2) 시스템 모델

S 는 STA 로부터 사용자들의 거리 및 속력 그리고 QoS(desired capacity)로 구성하였고, A 은 해당 시점에서 각 사용자들에게 할당되는 bandwidth 이다. 이때 총 bandwidth 10 MHz 이며 resolution 은 1 MHz 로 설정하였다. R 은 모든 사용자들의 utility 총합이며 utility 함수는 다음과 같이 정의되고 이는 사용자가 체감하는 capacity 에 대한 만족도를 나타낸다 [2].

$$U(c) = \begin{cases} \beta_1 e^{p_1(c-c^d)} & \text{if } c < c^d \\ 1-(1-\beta_2)e^{-p_2(c-c^d)} & \text{if } c \geq c^d \end{cases}$$

여기서 c는 channel capacity, c_d는 desired channel capacity, β와 p는 곡선의 기울기를 조정하는 parameter 이다.

채널 모델은 거리에 따른 path loss 만을 고려하였고, capacity 는 Shannon's formula 을 사용하였다 [3].

3) Scenario 및 Simulation Setup

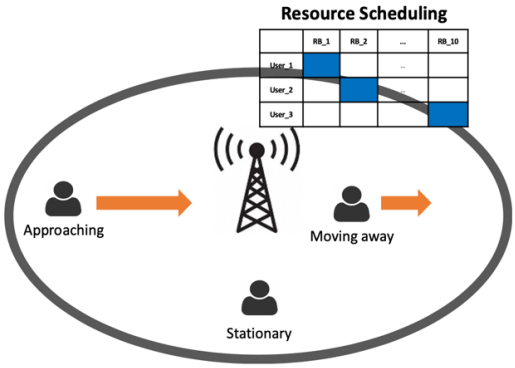

그림 1. 주파수 자원 관리 모델

하나의 STA 가 총 사용자 3 명의 무선 주파수 자원을 관리하고 있으며 각각 100 km/h 의 속력으로 다가오는 또는 멀어지는 그리고 가만히 있는 mobility 상태를 가정하였다. 주파수 자원은 매 1 ms 마다 할당된다. 또한 3 명의 사용자는 그림 2 처럼 각각 다른 utility function 을 가지고 있다.

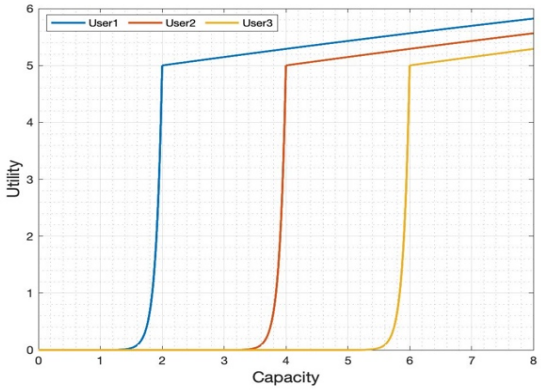

그림 2. Utility function

기타 DQN 알고리즘과 관련된 hyper parameter 은 해당 알고리즘 논문을 참고하였다 [1].

시뮬레이션을 통해 STA 이 사용자 3 명의 mobility 와 QoS 을 고려하여 한 episode 에서의 utility 총 합이 최대가 되도록 매 time step 마다 적절하게 한정된 주파수 자원을 할당할 수 있는지 확인하고자 한다.

4) Simulation Result

그림 3 에서 실험 결과 그래프의 x 축은 학습 진행에 따라 반복된 episode 의 횟수이고 y 축은 step 당 사용자 3 명의 utility 합의 평균이다. 총 episode 의 수는 100 이며 한 episode 내에서 총 step 수는 1000 으로 설정하였다.

첫 episode 에서는 학습된 것이 없으므로 완전히 무작위로 무선 주파수 자원을 할당하는 것이라고 볼 수 있다. 이 때 step 당 utility 합의 평균은 7.5~7.6 사이의 값을 갖는다.이후 episode 가 반복될 수록 점점 학습이 되어 80 episode 근방에서 utility 합의 평균이 수렴하는 것을 볼 수 있다. 이때 Step 당 utility 합의 평균은 11.1~11.3 사이의 값을 보여주고 있다.

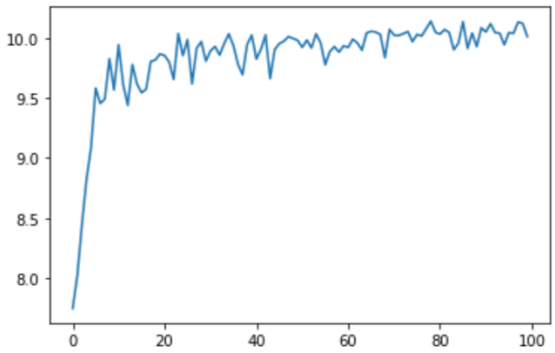

그림 3. Learning curve

결론적으로 심층 강화 학습을 적용함으로써 utility 총합에서 대략 40~50%의 성능의 향상을 얻을 수 있음을 확인하였다.

Ⅲ. 결론

본 논문에서는 간단한 시나리오에서의 시뮬레이션을 통해 심층 강화 학습을 활용하여 무선 네트워크 환경에서 STA 이 서로 다른 mobility 와 QoS 을 가지는 사용자들의 utility 총합을 최대화하도록 한정된 주파수 자원을 동적으로 할당할 수 있음을 보였다.

기존 연구에서는 거리에 따른 path loss 만을 고려하는 간단한 채널 모델을 가정하였다. 추후 연구에서는 shadowing 또는 multi path 그리고 Doppler effect 와 같은 채널 요소들을 보완함으로서 좀 더 현실적이고 복잡한 시나리오에서도 심층 강화 학습이 적용 가능한지 살펴보고자 한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2017R1E1A1A01074358)

참 고 문 헌

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

[2] Tan, Liansheng, et al. "Utility maximization resource allocation in wireless networks: Methods and algorithms." IEEE Transactions on systems, man, and cybernetics: systems 45.7 (2015): 1018-1034.

[3] T. Cover and J. Thomas, "Elements of Information Theory," Wiley Interscience, 2nd Ed.