캡스톤디자인

결과 보고서

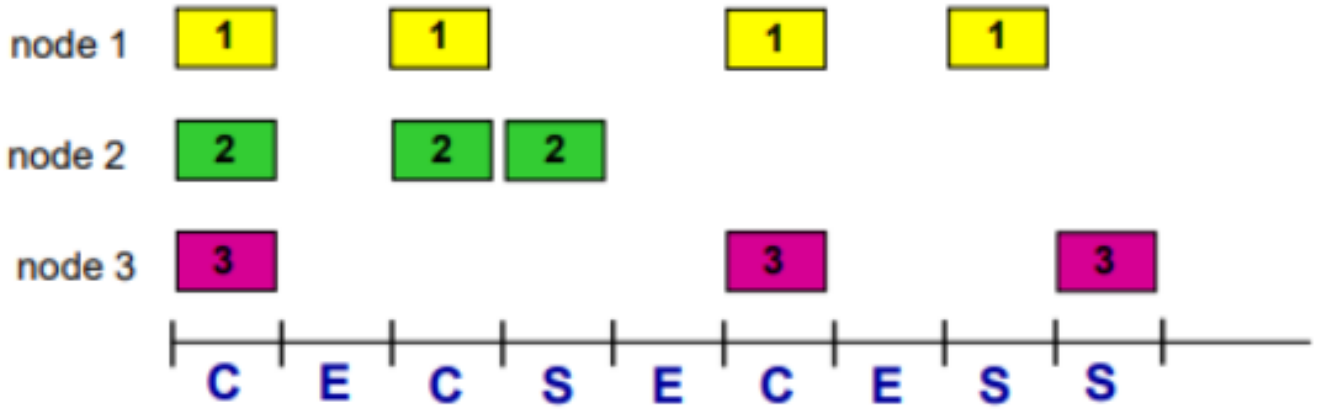
깡통 조
김도원, 윤희준, 이선우, 이지환, 황건하

CONTENTS

• 기존 노드의 프레임 선점 방식(Slotted Aloha)

- 기존 노드의 선점방식은 개수와 용량이 제한된 프레임 안에서
  각 노드들이 선점하여 데이터를 전송하는 방식입니다.

- 이 방식은 노드 간의 충돌이 발생할 경우 일정한 지연시간을 가진 뒤 재전송하는 방식을 가집니다.



• 새로운 문제의 해결방법

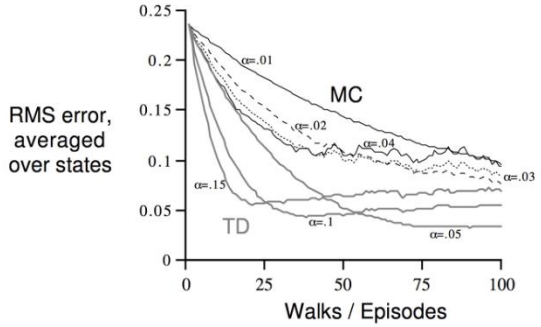
- 이러한 노드 간 충돌 후 재전송 방식은 불가피하다고 여겨집니다.
- 충돌을 최소화하기 위해 비용을 장비에 투자하기에는 너무도 큰 비용이 발생하기 때문에 한계가 있습니다.
- 우리 과제의 목적은 문제 해결을 소프트웨어적인 관점으로 보아 각 노드 간의 선점 시 발생하는 충돌을 자가 학습을 통해 최소화 할 수 있는 강화학습을 도입하고자 합니다.

• 강화학습 알고리즘 적용의 적시성 검토

- 알고리즘을 적용하기에 앞서 노드 통신의 어떤 부분에 알고리즘을 적용하는 것이 좋을지 검토가 필요합니다.
- 어떤 알고리즘을 적용하는 것이 성능을 향상시키는지 검토가 필요합니다.
- 결론적으로 강화학습 알고리즘들의 특성을 파악 후 노드 경쟁 방식에 적용하여 상황별 성능을 비교 분석하기로 했습니다.

• 강화학습 알고리즘 종류와 적용

- MC has high variance, zero bias
  - Good convergence properties
  - (even with function approximation)
  - Not very sensitive to initial value
  - Very simple to understand and use
- TD has low variance, some bias
  - Usually more efficient than MC
  - TD(0) converges to $v_\pi(s)$
  - (but not always with function approximation)
  - More sensitive to initial value



❖ MC는 variance가 높은 대신 bias 가 없고, TD는 variance가 낮은 대신 bias가 있다.

- 위에서 설명한 Monte Carlo(MC)와 Temporal Difference(TD)의 성능 차이를 비교하기 위해 같은 환경을 만들어 줍니다.
- 두 알고리즘의 차이로는 모든 노드들이 전송을 성공하는 것을 관점으로 보는 것(MC)과 노드들이 전송을 시도하는 각 step(TD)을 관점으로 보는 것에 대한 차이입니다.
- 여기서 우리는 두 알고리즘이 노드, 슬롯의 개수, 프레임의 전송 용량에 따라 성능이 좋은 알고리즘이 다르다는 것을 알 수 있었습니다.

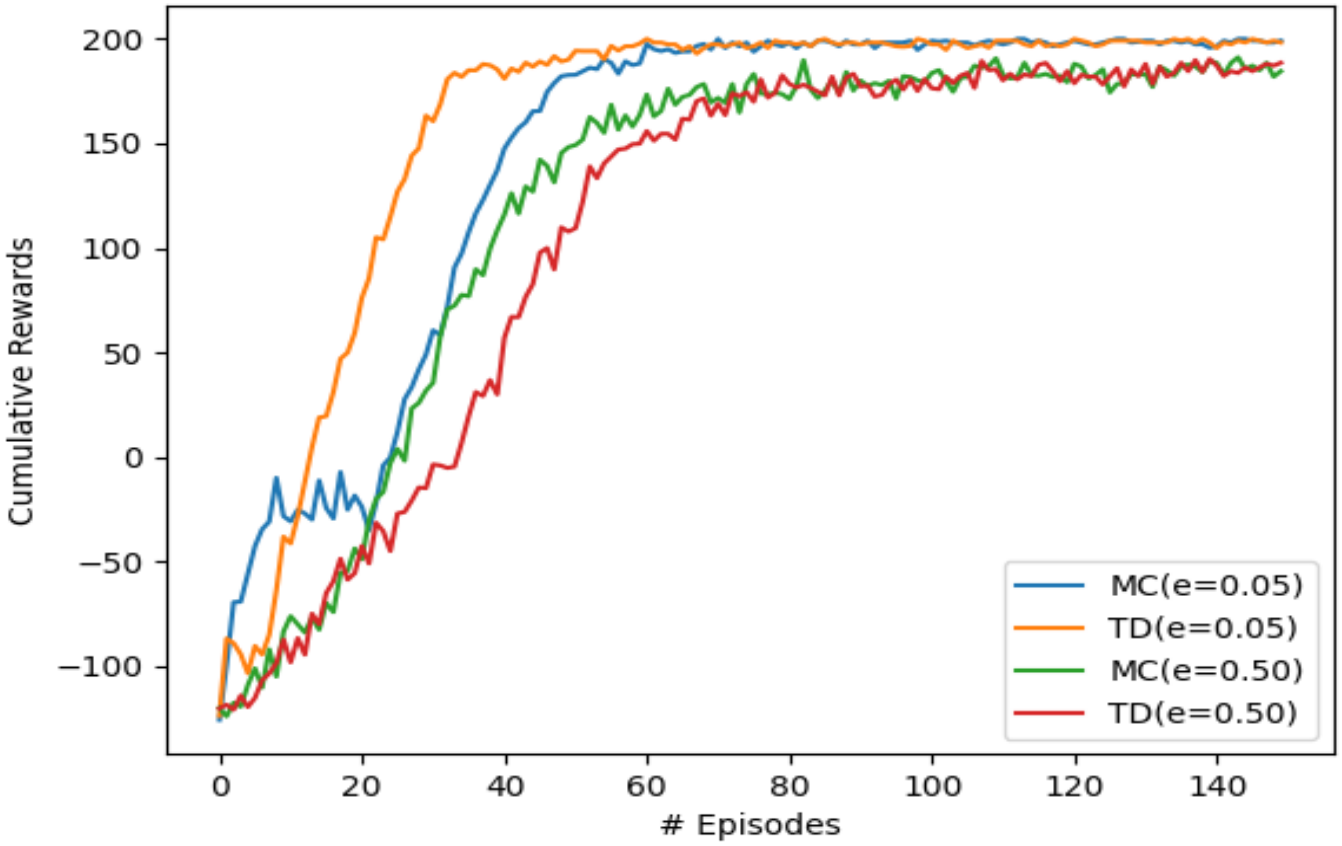
• 강화학습 알고리즘 연구과정과 결과

- 알고리즘의 성능 향상을 Factor별로 비교하기 위해 두 가지 경우로 나누어 실험을 진행하였습니다.
- 첫 번째 실험에서는 epsilon-greedy 알고리즘 적용 부분에 사용하는 epsilon을 다르게 하였습니다.
- 두 번째 실험에서는 MC와 TD알고리즘에 Episode가 종료되고 받는 Reward인 R2의 크기를 다르게 하였습니다.

03 결론 &기대효과

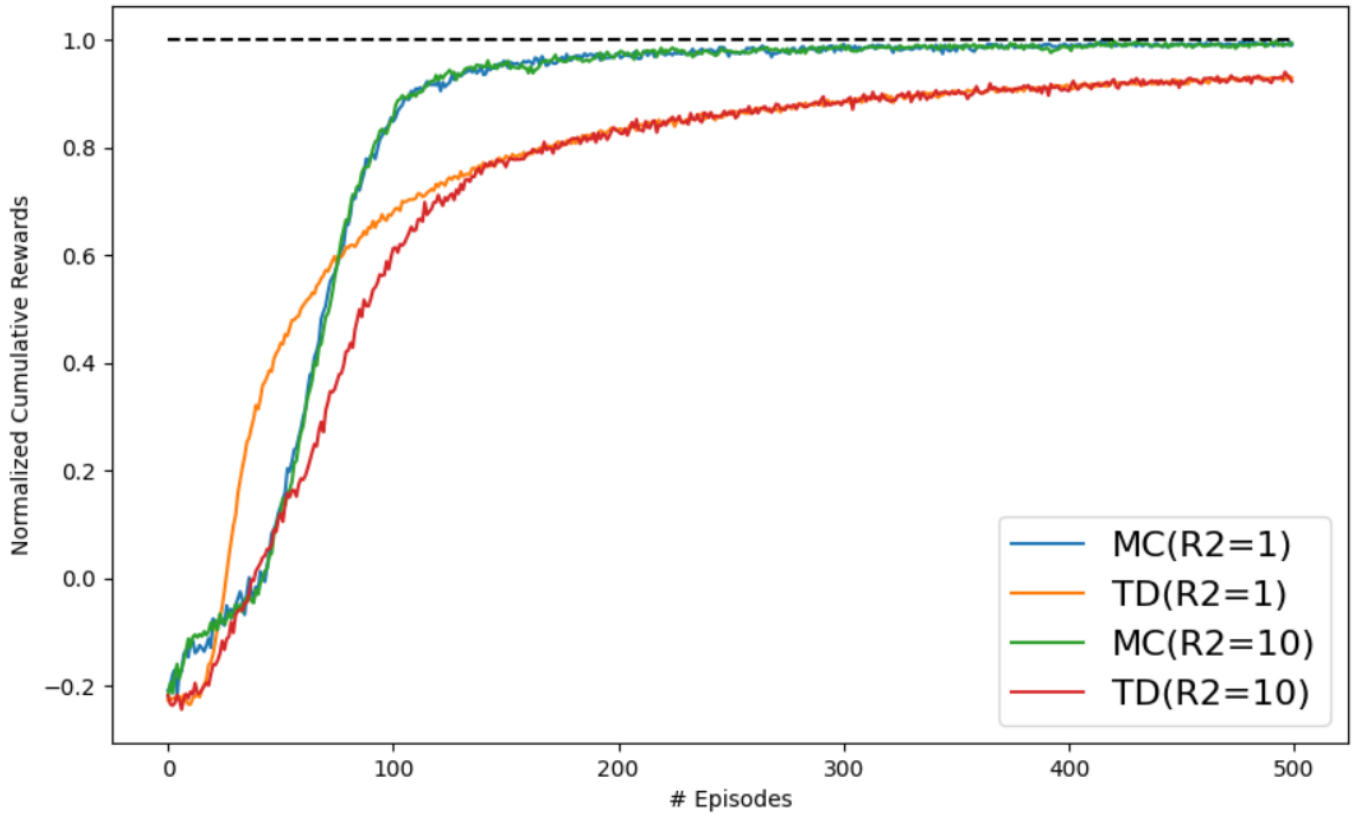
- Figure 1

- 실험 1의 결과입니다. 그래프의 Legend에 보이는 e(ε)의 값이 작을수록 알고리즘의 다음 Action 선택이 현재Q-Table의 최대 값에 대해 Greedy해지는 성질이 있습니다.
- 두 알고리즘에 주어지는 Reward의 크기는 같습니다.
- 그래프의 결과에서 보이듯 e의 값이 작을 수록 MC가 TD에 비해 더 빨리 수렴하는 것을 알 수 있습니다.
- 하지만 e의 값이 커질 경우 두 알고리즘 모두 성능이 떨어지는데, 알고리즘간의 상대적인 성능은 e의 값이 작을 경우와는 반대인 양상을 볼 수 있습니다.



• Figure 2

- 실험2의 결과입니다. 각 Node가 선택을 하였을 때 주는 Reward의 값을 R1, Episode가 끝났을 때 주는 Reward 값을 R2로 나누어 R1=R2일 때와 R1<R2일 때로 나누어 실험을 진행했습니다.
- 그래프는 500개의 random seed를 수행한 후, 받은 전체 reward를 평균값으로 표준화 하여 나타내었습니다.
- MC의 경우에는 R2의 크기가 달라져도 큰 영향을 받지 않고 비슷한 그래프를 그리는 경향을 보였습니다.
- TD의 경우에는 R1<R2인 경우가 수렴속도, 변동 폭의 안정성 부분 모두 떨어지는 경향을 보였습니다.



• 강화학습 알고리즘을 활용한 노드 선점방식 개선

- 본 연구에서는 간단한 알고리즘(MC와 TD)을 이용하여 결과를 도출 해보았습니다. 이 연구를 통하여 알고리즘 별로 놓인 Factor와 환경에 따라 다른 성능 향상이 보인다는 것을 확인하였습니다.
- e값이 작을수록 greedy한 선택을 보다 많이 하게 되므로, 목표치에 빠르게 수렴하는 경향성을 확인하였습니다.
- Reward(R1, R2)의 차이에 따라서는 MC와 TD에 제한적인 영향을 미치며, 경우에 따라서는 성능에 변동성을 줄 수 있습니다.
- 불가피한 노드 간의 충돌을 알고리즘을 통하여 최대한 줄여 나갈 수 있도록 하고자 합니다.
- 물리적인 작업을 통한 통신 시스템 증축보단 소프트웨어적으로 보다 간결한 작업 수행이 될 것이라 미루어 짐작하고 있습니다.
- MC와 TD 이외의 다양한 알고리즘을 연구하고 직접적인 경험을 통하여 통신 노드 간의 선점방식에 강화학습을 적용할 때에 더 복잡하지만 상황에 맞는 알고리즘을 적용한다면 더 좋은 결과를 얻을 수 있다는 것을 알 수 있었습니다.

03 결론 &기대효과

- 논문 첨삭

밀집된 통신기기의 무선접속 알고리즘에
적용할 수 있는 강화학습 알고리즘에 대한 비교 연구

이선우, 김도원, 윤희준, 이지환, 황건하, 신경섭
상명대학교

fhfgdyd986@naver.com, wlrtqur479@gmail.com, jungle128@gmail.com, jj8620@gmail.com, aromasy@naver.com, kwhin@smu.ac.kr

A Study on the Reinforcement Algorithm based Medium Access Algorithm in a Densely Deployed Wireless Networks

Lee Seon Woo, Kim Do Won, Yoon Hee Jun, Lee Ji Hwan,
Hwang Geon Ha, and Shin Kyung Seop
Sangmyung University

요 약

본 논문은 기존 강화학습을 활용한 통신 프로토콜 연구에서 주로 다루어진 Q-learning 기반 알고리즘보다 더 기본적인 강화학습 알고리즘인 Temporal-Difference(TD)와 Monte-Carlo(MC)를 통신 알고리즘에 적용하여 강화학습 방식에 따른 네트워크 접속성능의 비교분석을 진행하였다. 실험을 위해 알고리즘은 대용량의 데이터를 작은 단위로 나누어 저장하는 다중 사용자 환경을 가정하였다. 본 논문에서는 TD와 MC 알고리즘에 동일한 reward를 주기만 ε-greedy 방식을 사용하는 경우에 ε 값을 다르게 설정하여 성능을 비교하는 실험과 ε 값을 고정하고 reward를 다르게 하는 실험을 수행하였다. 실험 결과 TD와 MC를 활용한 노드 간 통신에 ε값이 작을수록 모두 성능이 향상되는 경향성을 확인하고, 최종 reward의 차이에 따라 TD의 성능이 변화가 더 크다는 것을 확인하였다.

Ⅰ. 서론

네트워크 환경이 모바일 기기와 IoT 기기의 증가로 통신량이 급증함에 따라 효율적인 데이터 저장 프로토콜의 필요성이 커져가고 있다. 이에 부응하여 추가적으로 통신시설을 증축하지 않고도 네트워크의 효율성을 높일 수 있는 강화학습을 활용한 통신 프로토콜 알고리즘이 많이 제안되어왔다. 기존 연구에서는 라우팅 프로토콜에 강화학습의 일종인 Q-learning을 활용하여 노드 경로 선택을 최적화하는 방식으로 네트워크의 효율성을 높이는 연구가 주를 이뤘다. [1][2].

본 논문에서는 통신 알고리즘에 강화학습을 적용하여 서로 다른 노드가 동시에 데이터를 저장하여 충돌하는 상황을 방지하고자 한다. 기존에 연구가 많이 이루어진 노드는 서로의 상황을 모르고 일방적으로 데이터를 저장하므로 통신을 위한 알고리즘으로 model-free prediction이 적합하다. model-free 알고리즘은 환경에 대한 모델을 세우기 어려운 경우에 적용할 수 있고, 여러 도메인에 동일한 방법을 적용할 수 있다는 장점이 있다. prediction은 선택한 policy를 따랐을 때의 value function을 찾는 문제이며 최적의 해를 찾기보다 value 로의 수렴 여부가 중요하다. 본 논문에서는 기존 연구에서 주로 다루어진 Q-learning 기반 알고리즘 보다 더 기초적인 강화학습 알고리즘 중 Monte-Carlo(MC)와 Temporal-Difference(TD) 두 가지 방식을 통신노드의 미디엄 액세스 프로토콜에 적용하여 성능을 비교함으로써 네트워크의 효율성을 더 높일 수 있는 방안을 비교분석하였다.

Ⅱ. 본론

MC는 episode를 진행하며 return을 저장해두고 episode가 종료된 이후에 그 결과로 학습한다. return은 discounted reward 이며 다음과 같다.

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T \quad (1)$$

위의 return에 따라 policy π에 대한 value function은 다음과 같다.

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] \quad (2)$$

Episode가 진행됨에 따라 다음과 같이 value function이 업데이트된다.

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t)) \quad (3)$$

MC는 실행한 후에 통계를 내는 단순한 휴리스틱 episode가 끝나야만 적용이 가능한데, TD는 MC와 다르게 완전하지 않은 episode에서도 학습이 가능하다. 이 경우 value function은 다음과 같이 업데이트할 수 있다.

$$V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)) \quad (4)$$

MC에서의 value function에서 G_t는 discounted reward의 합이었지만, TD에서는 episode가 완료되어야 하는 G_t 대신에 다음 step의 예측으로 현재 예측을 업데이트하는 $R_{t+1} + \gamma V(S_{t+1})$를 사용한 것을 볼 수 있다.

본 논문에서 진행된 실험은 다음과 같은 환경을 가정하여 설계되었다. 10Mbyte의 데이터를 1Mbyte 씩 나누어 저장하여 10개의 slot으로 이루어진 stage에 데이터를 다 저장할 경우 한 episode가 끝난 것으로 가정하고 1Mbyte를 저장하는 것을 한 step으로 가정하였다. 또한 MC 알고리즘은 한 episode가 끝났을 때마다, TD 알고리즘은 한 step을 수행할 때마다 policy를 업데이트한다는 차이가 있다.

실험은 총 2가지 방식으로 진행되었으며 이는 다음과 같다. 실험 1은 epsilon (ε) -greedy 알고리즘을 활용하여 ε 값이 작을수록 예측 value 값이 가장 큰 행동을 선택할 가능성이 크고 ε 값이 클수록 무작위적으로 행동을 선택한다는 점을 고려하여 ε 값에 따른 randomness가 각 알고리즘의 성능에 미치는 영향을 분석하기 위한 실험이다. 이를 확인하고자 TD/MC 알고리즘에 0.5, 0.05의 ε값을 주어 얼마나 성공적으로 데이터를 저장하였는지를 비교하였다. Fig1 그래프는 실험 1의 결과를 보여준다.

실험 2는 TD/MC 알고리즘에 reward가 미치는 영향을 비교 분석하기 위해 설계되었다. reward를 R1(한 step을 수행하고 받는 reward), R2(모든 step이 끝난 episode 종료 후 받는 reward)로 구분하여 R1=R2인 경우, R1<R2인 경우로 reward 값을 고정하여 실험하였다. Fig2 그래프는 실험 2의 결과를 보여준다.

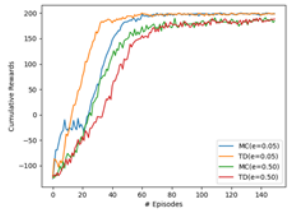


Fig 1. ε 값에 따른 알고리즘의 성능변화

위의 그래프는 MC와 TD에 각각 고정된 reward 값에 서로 다른 ε값을 주어 나온 결과이다. ε의 값이 클수록 다음 step에 random한 선택을 할 확률이 높아진다. 위의 그래프에서 볼 수 있듯, ε의 값이 작을수록 TD가 MC에 비해 더 빨리 수렴하게 되어 유리한 결과를 보여주지만, ε의 값이 커질 경우 그 반대의 결과가 나오게 된다.

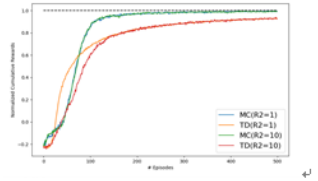


Fig 2. Reward에 따른 알고리즘의 성능변화

위의 그래프는 reward 값에 따른 MC와 TD의 결과이다. 기본적으로 R1을 1로 부여하고, R1=R2일 때와 R1<R2일 때를 비교하였다. 그래프는 각 unit이 episode 당 받은 전체 reward를 서로 다른 random seed에 대한 평균값으로 표준화하여 나타낸 결과이다. 500개의 random seed를 적용한 환경으로, ε-greedy 방식으로 action을 선택했다. TD는 step마다 random한 확률을 가져 exploration적인 선택을 하게 했고, MC는 첫 선택을 해당 step에 고정적으로 가져가게 하여 TD에 비해 더 greedy한 action 선택을 하였다. 위의 그래프에서 볼 수 있듯이, MC의 경우 reward의 차이에 큰 영향을 받지 않아 비슷한 그래프를 그리는 경향을 보였고, TD는 R2가 적절에 따라 수렴속도와 변동폭의 안정성이 다소 떨어지는 경향을 보였다.

Ⅲ. 결론

본 논문에서는 MC와 TD를 사용한 노드 간 통신에 ε값이 작을수록 빠르게 목표치에 수렴하는 경향성을 확인하였다. ε값이 작을 경우, MC와 TD 모두 임의적인 선택이 아닌 greedy한 선택을 더 많이 하게 되므로 ε값이 큰 경우보다 더 빨리 수렴하게 된다. 반면, reward의 크기 변화가 MC에 미치는 영향은 제한적이며 TD의 경우 Bias가 생겨 성능에 변동성이 적지는 부작용이 나타날 수 있다. 결과적으로 두 알고리즘 모두에 결정적인 영향을 미치는 요소는 ε값으로, TD가 더 greedy한 선택을 할수록 MC보다 유리해지는 경향을 보인다는 것을 확인하였다.

ACKNOWLEDGMENT

본 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1064069).

참 고 문 헌

[1]김희정, 정진우, "주기적 Q-table 업데이트를 활용한 무선 라우팅 프로토콜," 한국통신학회논문지 45.12 (2020): pp. 2099-2106.
[2]김희원, 조호신, "수중 센서 네트워크에서 강화학습 기반의 라우팅 프로토콜," 한국통신학회논문지 45.10 (2020): pp. 1716-1719.

THANK YOU