보상 최대화를 위한 Deep Q-Network 기반 Adaptive slotted ALOHA 기법

김지나, 박현희*
명지대학교

jnkim@mju.ac.kr, *hhpark@mju.ac.kr

Adaptive Slotted ALOHA Method Based on Deep Q-Network for Rewards Maximization

Jeena Kim, Hyunhee Park*
Myongji Univ.

요 약

본 논문에서는 Adaptive slotted ALOHA 기법에 DQN 을 적용한다. 충돌이 발생한 디바이스는 Q-network 와 target network 간 가중치 값의 차이를 충돌이 발생한 디바이스의 페널티 값으로 사용한다. 따라서, 학습이 종료된 후에는 모든 디바이스가 최적의 행동을 하여 Q 값을 최대화하여, 지속적으로 패킷 전송에 성공한다. 이를 통해, 본 논문에서는 페널티 파라미터에 따른 에피소드 별 누적 보상을 통해 성능을 검증한다. 페널티 파라미터를 적용한 Deep Q-Network 기반 Adaptive slotted ALOHA 기법이 지속적으로 보상이 증가함을 보였다.

Ⅰ. 서 론

Slotted ALOHA 기법[1]은 machine type communication(MTC) 환경에서 다수의 디바이스가 동시에 데이터를 전송할 때 사용한다. 하지만, MTC 환경에서 기존의 slotted ALOHA 기법을 적용하여 다중 접근을 설계하는 경우 충돌 발생 증가를 야기할 수 있다.

이를 해결하기 위해 slotted ALOHA 기법에 Q-learning 을 적용한 Q-learning based slotted ALOHA 기법[2]은 디바이스가 전송 결과에 따라 보상을 얻는다. 디바이스가 패킷 전송에 성공했을 경우 +1 의 보상을 얻으며 같은 타임 슬롯에 2 대이상의 디바이스가 패킷 전송을 시도하여 충돌이 발생했을 경우 -1 의 페널티를 받는다. 그러나, Q-learning based slotted ALOHA 기법은 이진으로 구성된 보상을 얻기 때문에 누적 보상을 최대화하기 힘들다는 문제가 있다.

[2]의 문제를 해결하기 위해 adaptive slotted ALOHA 기법[3]이 제안된다. [3]의 기법은 하나의 타임 슬롯에 단일 디바이스가 패킷 전송을 시도할 경우 전송에 성공하여 +1 의 보상을 얻는다. 그러나, 같은 타임 슬롯에 2 대 이상의 디바이스가 동시에 패킷 전송을 시도했을 경우 충돌이 발생한다. [3]의 기법은 충돌이 발생한 디바이스는 같은 타임 슬롯에서 발생한 총 디바이스 개수를 고려하여 페널티 파라미터가 적용된다. 이는 현재 에피소드 내 타임 슬롯의 혼잡도를 디바이스가 식별하기 위함이다. 디바이스는 전송 결과에 따른 보상 및 페널티를 얻은 후 Q 값을 업데이트 한다. 업데이트 된 Q 값은 Q-table 에 저장되며 이전 상태에서 수행한 행동이 기록된다. 따라서, 학습이 진행되면서 최적의 행동을 수행하여 Q 값을 최대화하는 것을 목표로 한다. 학습이 종료된 시점에서 디바이스는 고유한 타임 슬롯을 점유할 수 있다.

그러나, 상태 공간과 행동 공간이 증가하면서 모든 Q 값을 Q-table 에 저장하기에 적합하지 않다. 이러한 문제를 해결하기 위해 본 논문에서는 Adaptive slotted ALOHA 기법에 Deep Q-network(DQN)[4,5,6]를 적용한다. 제안하는 기법은 target Q 값과 예측된 Q 값의 격차를 페널티 파라미터로 사용하여 학습을 진행한다. 따라서, 학습이 진행될수록 target Q 값과 예측된 Q 값의 격차를 감소시켜 예측된 Q 값이 target Q 값에 근사하는 방향으로 학습된다.

Ⅱ. DQN 기반 Adaptive slotted ALOHA(DQ-ASA) 제안 및 결과

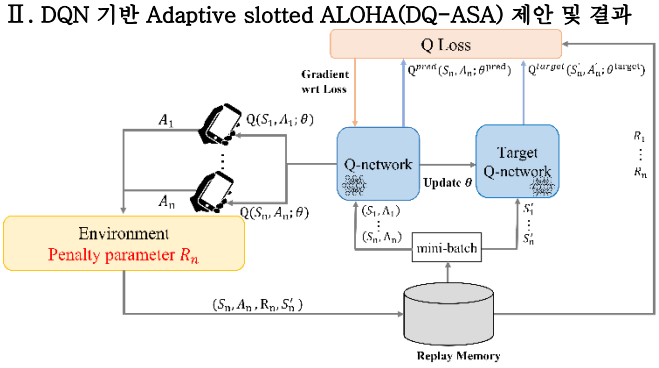

그림 1. DQ-ASA 기법

DQN 기반 Adaptive slotted ALOHA 기법은 Deep Q network 에 Adaptive slotted ALOHA 를 적용한다. 그림 1 은 제안하는 기법인 DQ-ASA 기법의 과정을 도식화한 것이다. Environment 는 디바이스의 행동을 관찰하는 영역으로 디바이스의 행동은 다음 식 (1)과 같이 나타낼 수 있다.

$$A_n = \begin{cases} 1, \text{when user transmit single packet} \\ 0, \text{when user choose to not transmit a single packet} \end{cases} \quad (1)$$

n은 $n \in N$으로 디바이스의 인덱스를 나타낸다. A_n이 1 일 경우 현재 상태에서 디바이스가 패킷을 보내고자 한다. A_n이 0 일 경우 디바이스가 현재 상태에서 패킷을 보내지 않는다. A_n이 1 인 디바이스는 environment 에서 여러 개로 나뉘어진 타임 슬롯에 패킷 전송을 시도한다. 각 디바이스는 패킷 전송 결과에 따른 보상을 얻으며 전송 결과에 따른 보상은 다음 식 (2)와 같이 나타낸다.

$$R_n = \begin{cases} 1 & , \text{if user } n \text{ transmission sucess} \\ \theta^{target} - \theta^{pred} & , \text{if user } n \text{ collision occurs} \end{cases} \quad (2)$$

단일 디바이스가 타임 슬롯에 전송을 시도했을 경우 전송에 성공한다. 따라서, 디바이스는 +1 의 보상 r_n을 얻는다. 그러나, 2 대 이상의 디바이스가 같은 타임 슬롯에 패킷 전송을 시도할 경우 충돌이 발생한다. 충돌이 발생한 경우 Q-network 의 가중치인 θ^{pred}와 target network 의 가중치인 Q^{target}간의 격차만큼 페널티를 얻는다. θ^{pred}는 Q-network 의 출력인 $Q^{pred}(s_n, a_n)$의 가중치를 나타낸다. 또한, θ^{target}은 target

network 의 출력인 $Q^{target}(s'_n, a_n)$의 가중치를 나타낸다. 이는 다음 상태에서 행동을 수행할 때 두 네트워크의 격차를 감소시켜 target value 에 근사하기 위함이다. 전송 결과에 따른 보상과 함께 (s, a, r, s')로 이뤄진 샘플은 replay buffer 에 저장된다. Replay buffer 에 저장된 샘플은 크기 32 로 mini-batch 가 된다. Q-network 에는 현재 상태와 행동을 나타내는 (s_n, a_n)가 업데이트 되어 현재 에피소드에서 디바이스 n의 Q^{pred}을 계산한다. Target network 에는 (s'_n)가 업데이트 되어 현재 디바이스 n의 Q^{target}을 계산한다. 각 네트워크에서 계산된 $Q^{pred}(S_n, A_n; \theta^{pred})$ 와 $Q^{target}(S_n, A_n; \theta^{target})$ 는 Q-loss 로 업데이트된 후 두 값의 제곱 오차를 계산한다. 이는 손실 함수에 해당하며 다음 식(3)과 같이 구할 수 있다.

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(r_i + \gamma \max_{a'} Q(a', s'_i; Q^{target}) - Q(a_i, s_i; \theta^{pred}))^2 \quad (3)$$

표 1 학습 파라미터 설정

Contents	Value
Episode	100
The number of timeslots	400
The number of devices	100
Batch size	32
Optimizer	Adam
Exploration	ε-greedy
Learning rate	0.01

다음 표 1 은 DQ-ASA 기법의 DQN 학습을 진행하기 위한 파라미터이다. DQ-ASA 기법의 성능 검증을 위해 에피소드 증가에 따른 누적 보상 결과를 도출한다. 누적 보상이 증가할수록 패킷 전송에 성공하는 디바이스가 증가하고 있음을 나타낸다. 본 실험에서는 1 개의 episode 별 400 개의 타임 슬롯이 있다고 가정한다. 전송 채널에 참여하는 디바이스 개수는 100 대이며, 디바이스는 타임 슬롯에 패킷을 전송한다. 디바이스의 경험 샘플은 32 개로 mini-batch 되어 Q-network 와 target network 로 업데이트 된다. Q-loss 손실함수를 계산하여 Adam optimizer 을 적용하여 손실함수를 Q-network 로 업데이트 한다. 탐색은 ε-greedy 기법을 적용하여 0.1 으로 초기화 하여 탐색을 시작한 후 학습이 진행될수록 0.98 의 속도로 지수 감쇄하여 학습 종료 시 0.01 의 탐색을 한다. 마지막으로 학습의 learning rate 는 0.01 로 설정하였다.

본 실험의 기준을 얻고자 페널티 파라미터를 적용시키지 않은 DQ-ASA 기법과 비교한다. 페널티 파라미터를 적용하지 않은 경우는 디바이스의 전송 결과에 따라서 성공한 경우 +1 의의 보상을 얻고 충돌이 발생한 경우 -1 의 페널티를 얻는다. 에피소드 증가에 따른 DQ-ASA 기법의 누적 보상 결과는 다음 그림 2 와 같다.

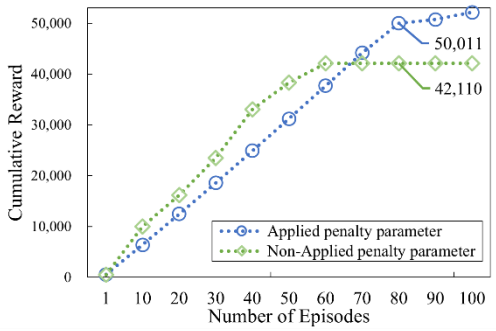

그림 2. 페널티 파라미터 적용 여부에 따른 누적 보상

그림 2 의 Applied penalty parameter 는 제안하는 DQ-ASA 기법을 나타내며, Non-Applied penalty parameter 는 페널티 파라미터를 적용하지 않은 DQ-ASA 기법을 나타낸다. 학습 초반에는 페널티 파라미터를 적용하지 않은 DQ-ASA 기법이 누적 보상이 높은 것으로 나타난다. 그러나, 50 에피소드가 지난 후 누적 보상 값이 점차 수렴하는 것으로 나타난다. 이는 페널티 Non-Applied penalty parameter 는 학습이 진행될수록 누적 보상이 크게 증가하지 않음으로 판단된다.

반면에, Applied penalty parameter DQ-ASA 기법은 학습이 진행될수록 지속적으로 누적 보상이 증가하고 있음을 나타낸다. 따라서, 학습이 진행될수록 제안하는 DQ-ASA 기법에서 디바이스들이 패킷 전송에 성공하고 있음을 나타낸다. 이는 학습이 진행될수록 적용된 페널티 파라미터인 예측된 Q 값의 가중치 θ^{pred}와 target Q 값의 가중치 θ^{target}의 격차가 점차 감소하여 예측된 Q 값이 target Q 값에 근사하고 있음으로 판단된다.

Ⅲ. 결론

본 논문에서는 DQN 기반 Adaptive slotted ALOHA(DQ-ASA) 기법을 제안했다. DQ-ASA 기법은 MTC 와 같이 다수의 디바이스가 연속적으로 패킷 전송하는 환경에서 누적 보상 성능을 검증했다. DQ-ASA 기법은 에피소드가 증가할수록 지속적으로 누적 보상 값이 증가함을 나타냈다. 따라서, 다수의 디바이스들의 패킷 전송 성공이 증가하고 있음으로 판단한다. 향후 연구에서는 DQ-ASA 기법에서 target Q 값과 예측된 Q 값의 차이의 다양한 정책들을 적용하여 DQ-ASA 기법의 적합한 정책을 찾는 것을 목표로 한다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2021-0-00990, 설명가능한 인공지능 기반 무선랜 네트워크 시스템 고도화 핵심 기술 연구, No. 2021-0-00368, 6G 서비스를 위한 인공지능/머신러닝 기반 자율형 MAC 개발)

참 고 문 헌

[1] L. C. Kalalas, F. Vazquez-Gallego, L. Alonso, and J. Alonso-Zarate, “Goodbye, aloha!,” IEEE access, vol. 4, pp. 2029-2044, Apr. 2016.

[2] Y. Chu, P. D. Mitchell, D. Grace, “ALOHA and Q-Learning based medium access control for Wireless Sensor Networks,” 2012 International Symposium on Wireless Communication Systems (ISWCS), p. 511-515, Aug. 2012

[3] J. Kim and H. Park, “단말의 활성화를 고려한 Q-learning 기반의 Adaptive slotted ALOHA 기법,” Proc. Symp. KICS, p. 1012-1013. Jun. 2022.

[4] V. Mnih, et al, “Human-level control through deep reinforcement learning,” nature, vol. 518(7540), p. 529-533, Feb. 2015.

[5] M. A. Jaddon, et al. “Deep Reinforcement Learning for Random Access in Machine-Type Communication,” 2022 IEEE Wireless Communications and Networking Conference (WCNC), May. 2022.

[6] O. Naparstek, K. Cohen. “Deep multi-user reinforcement learning for distributed dynamic spectrum access,” IEEE transactions on wireless communications vol. 18(1), p. 310-323, Jan. 2019.