JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 11: 1500~1506, Nov. 2020

단말간 직접 통신 네트워크를 위한 심층 강화학습 기반 분산적 스케쥴링 알고리즘

정무웅[1] · 김륜우[2] · 반태원[3*]

A Distributed Scheduling Algorithm based on Deep Reinforcement Learning for Device-to-Device communication networks

Moo-Woong Jeong[1] · Lyun Woo Kim[2] · Tae-Won Ban[3*]

[1]Ph.D. Postdoctoral Research Fellow, Dept. of Information and Communication Engineering, Gyeongsang National University, TongYeong, 23064 Korea
[2]Undergraduate Student, Dept. of Information and Communication Engineering, Gyeongsang National University, TongYeong, 23064 Korea
[3*]Associate Professor, Dept. of Information and Communication Engineering, Gyeongsang National University, TongYeong, 23064 Korea

요 약

본 논문에서는 오버레이 단말 간 직접 (Device-to-Device : D2D) 통신 네트워크를 위한 강화학습 기반 스케쥴링 문제를 연구한다. 강화학습 모델 중 하나인 Q-learning을 이용한 D2D 통신 기술들이 연구되었지만, Q-learning은 상태와 행동의 개수가 증가함에 따라서 높은 복잡도를 유발한다. 이러한 문제를 해결하기 위하여 Deep Q Network (DQN) 기반 D2D 통신 기술들이 연구되었다. 본 논문에서는 무선 통신 시스템 특성을 고려한 DQN 모델을 디자인하고, 피드백 및 시그널링 오버헤드를 줄일 수 있는 DQN 기반 분산적 스케쥴링 방식을 제안한다. 제안 방식은 중앙집중식으로 변수들을 학습시키고, 최종 학습된 파라미터를 모든 단말들에게 전달한다. 모든 단말들은 최종 학습된 파라미터를 이용하여 각자의 행동을 개별적으로 결정한다. 제안 방식의 성능을 컴퓨터 시뮬레이션을 통하여 분석하고, 최적방식, 기회주의적 선택 방식, 전체 전송 방식과 비교한다.

ABSTRACT

In this paper, we study a scheduling problem based on reinforcement learning for overlay device-to-device (D2D) communication networks. Even though various technologies for D2D communication networks using Q-learning, which is one of reinforcement learning models, have been studied, Q-learning causes a tremendous complexity as the number of states and actions increases. In order to solve this problem, D2D communication technologies based on Deep Q Network (DQN) have been studied. In this paper, we thus design a DQN model by considering the characteristics of wireless communication systems, and propose a distributed scheduling scheme based on the DQN model that can reduce feedback and signaling overhead. The proposed model trains all parameters in a centralized manner, and transfers the final trained parameters to all mobiles. All mobiles individually determine their actions by using the transferred parameters. We analyze the performance of the proposed scheme by computer simulation and compare it with optimal scheme, opportunistic selection scheme and full transmission scheme.

키워드 : 단말간 직접 통신, 기계학습, 강화학습, DQN, 스케쥴링 알고리즘
Keywords : Device-to-Device communication, Machine learning, Reinforcement learning, DQN, Scheduling algorithm

Received 24 September 2020, **Revised** 25 September 2020, **Accepted** 5 October 2020

*** Corresponding Author** Tae-Won Ban(E-mail:twban35@gnu.ac.kr, Tel:+82-55-772-9177)
Associate Professor, Department of Information and Communication Engineering, Gyeongsang National University, Tongyeong, 23064 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.11.1500 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

폭발적으로 증가하는 모바일 데이터 트래픽을 처리하기 위하여 무선 이동통신 네트워크의 전송 용량을 증가시킬 필요가 있다. 무선 이동통신 네트워크 전송 용량의 개선을 위한 기술 중 하나로 단말 간 직접 (Device-to-Device Communication : D2D) 통신 기술에 관한 많은 연구가 진행되어왔다. D2D 통신은 크게 셀룰러 통신과 같은 주파수 대역을 사용하는 언더레이 D2D 통신 (Underlay D2D Communication)과 셀룰러 통신과 다른 주파수 대역을 사용하는 오버레이 D2D 통신 (Overlay D2D Communication)으로 나뉜다. 언더레이 D2D 통신은 높은 주파수 효율성을 얻을 수 있지만, 셀룰러 네트워크와의 상호간섭으로 인하여 실용성이 낮다. 반면에, 셀룰러 네트워크와 서로 다른 주파수 대역을 사용하는 오버레이 D2D 통신 네트워크에 대한 링크 스케쥴링 기법들이 연구되었다 [1-3].

한편, 네트워크 환경 변화에 능동적으로 적응하는 강화학습을 이용한 무선 통신 네트워크 기술들이 많은 주목을 받고 있다. 이러한 강화학습 기술 중 Q-learning을 D2D에 활용하기 위한 연구들이 진행되었다 [4-6]. Q-learning은 상태와 행동의 가치를 판단하는 지표인 Q 값의 개수가 상태와 행동의 개수에 비례하여 증가하는 문제점이 있다. 이러한 문제를 해결하기 위하여 Deep Q Learning (DQN) 기반 D2D 통신 기술들이 연구되었다 [7-9]. 논문 [7]에서는 모바일 기기에 콘텐츠를 미리 분배하여 D2D 통신으로 다른 모바일 유저에게 콘텐츠를 제공하는 캐시 가능 D2D 통신 네트워크가 연구되었다. 구체적으로, D2D 통신 네트워크의 트래픽과 딜레이를 최소화하기 위한 콘텐츠 분배 문제를 풀기 위하여 DQN 기반 캐싱 전략 기법을 제안하였다. 논문 [8]은 소형셀 또는 모바일에 콘텐츠를 미리 저장하여 사용자들에게 콘텐츠를 직접 전달하는 모바일 엣지 캐싱 네트워크를 고려하였으며, 해당 네트워크에서 전력 소비를 최소화하기 위한 Q-learning 기반 및 DQN 기반 콘텐츠 분배 알고리즘을 제안하였다. 논문 [9]에서는 캐시 가능 D2D 통신 네트워크에서 다수의 기지국 간 정보 교환으로 발생하는 오버헤드를 줄이기 위하여 분산적 DQN 기반 콘텐츠 분배 기법을 제안하였다.

다양한 연구에도 불구하고, DQN을 활용한 오버레이 D2D 네트워크에서의 스케쥴링 문제에 대한 기존 연구는 존재하지 않는다. 그리고, 본 연구에서 고려하는 오버레이 D2D 네트워크는 셀룰러 통신과 다른 주파수를 사용하고 기지국의 개입 없이 단말 간 직접 통신하기 때문에, 오버레이 D2D 네트워크를 위하여 분산적 알고리즘이 필수적으로 요구된다. 따라서, 본 논문에서는 오버레이 D2D 통신 네트워크의 특성을 고려하기 위한 DQN 모델을 디자인하고, 기지국과 모바일 간 정보 교환으로 발생한 오버헤드를 줄일 수 있는 DQN 기반 분산적 스케쥴링 알고리즘을 제안한다. 본 논문의 나머지 구성은 다음과 같다. Ⅱ장에서는 D2D 통신 네트워크의 시스템 모델과 D2D 통신 네트워크를 위한 DQN 모델을 설명한다. Ⅲ장에서 DQN 모델 기반 분산적 스케쥴링 알고리즘을 제안한다. Ⅳ장에서 컴퓨터 시뮬레이션을 이용해서 제안 방식의 성능을 다양한 환경에서 분석하고, Ⅴ장에서 결론을 맺는다.

Ⅱ. 시스템 모델

본 논문에서 고려하는 오버레이 D2D 통신 네트워크에서는 전송 단말과 수신 단말이 각각 N_L개씩 존재하며 한 쌍씩 링크로 구성되어 있다. h_{ij}^t는 t번째 시간 슬롯에서 측정된 전송 단말 j와 수신 단말 i 사이의 채널 계수를 나타낸다. 모든 채널 계수는 평균이 0이고 분산이 1인 복소 정규 분포를 따르고, 독립적이고 동일하게 분포되어 있다고 가정한다. 시분할 이중화 기법을 고려하므로, 채널 간 대칭성으로 인하여 상향 링크와 하향 링크의 채널 계수가 동일하다고 가정한다. 따라서, 각 전송 단말은 수신 단말들로부터 받은 요청 메시지를 통하여 자신과 해당 수신 단말 간 채널 계수를 완벽하게 측정할 수 있다. 준정적 시간 페이딩 채널 (quasi-static fading channel)을 고려하여 채널 계수들은 데이터를 전송하는 시간 슬롯 내에서 일정한 값을 가지고, 시간 슬롯마다 독립적으로 변한다. 그리고 p_i^t는 t번째 시간 슬롯에서 전송 단말 i의 전송 전력을 나타내며, 모든 전송 단말의 전송 전력은 P_{tx}로 일정하다고 가정한다. t번째 시간 슬롯에서 수신 단말 i의 수신 SINR (signal-to-interference-plus-noise ratio)은 다음과 같이 계산된다.

$$\gamma_i^t = \frac{\rho I_i^t \| h_{ii}^t \|^2}{\sum_{k=1,k\neq i}^{N_L} \rho_k^t I_k^t \| h_{ik}^t \|^2 + 1}, \ 1 \le i \le N_L \quad (1)$$

$\rho = P_{\text{tx}}/N_0$는 전송 전력대비 잡음 비 (signal-to-noise: SNR)를 나타내고, 여기서 N_0는 백색 가우시안 잡음의 평균 전력을 나타낸다. 그리고 I_i^t는 t번째 시간 슬롯에서 전송 단말 i의 전송 여부를 나타낸다. 예를 들어, 전송 단말 i가 t번째 시간 슬롯에서 데이터를 전송하면 $I_i^t = 1$, 그렇지 않으면 $I_i^t = 0$이다.

$\boldsymbol{H}^t = \left\{\boldsymbol{h}_1^t, \cdots, \boldsymbol{h}_i^t, \cdots, \boldsymbol{h}_{N_L}^t\right\}^{\mathrm{T}}$를 시간 슬롯 t에서의 모든 채널 계수들의 집합으로 정의하고, 여기서 $\boldsymbol{h}_i^t = \left\{h_{1i}^t, \cdots, h_{ji}^t, \cdots, h_{N_L i}^t\right\}$는 전송 단말 i와 모든 수신 단말 사이의 채널 계수를 나타내는 열 벡터이다. 그리고, $\boldsymbol{I}^t = \left\{I_1^t, \cdots, I_{N_L}^t\right\}$는 슬롯 t에서의 모든 전송 단말들의 전송 여부를 나타내는 이진수들의 집합으로 정의한다. 시간 슬롯 t에서 채널 집합 $\boldsymbol{H}^t$와 $\boldsymbol{I}^t$에 따른 전송률의 합은 $C(\boldsymbol{H}^t, \boldsymbol{I}^t) = \sum_{i=1}^{N_L} \log_2\left(1+\gamma_i^t\right)$로 계산된다.

본 논문에서는 D2D 네트워크의 전송률의 합을 최대화하도록 하는 링크 스케쥴링 문제를 고려하고, 해당 문제는 다음과 같이 정의된다.

$$\max_{\boldsymbol{\rho}^t} \ C(\boldsymbol{H}^t, \boldsymbol{I}^t) \quad s.t. \ \boldsymbol{I}_i^t \in \{0,1\}, \forall i \quad (2)$$

식 (2)와 같은 최적 문제는 매우 큰 SNR 환경에서는 가장 좋은 채널 하나를 선택하는 기회주의적 선택 기법이 최적 방법이고, 매우 작은 SNR 환경에서는 모든 링크들이 통신하는 전체 전송 방식이 최적 방법이다 [10]. 따라서 우리는 환경에 따라서 적응적으로 전송 링크를 선택하는 DQN 기반 스케쥴링 알고리즘을 연구한다. 일반적으로 강화학습 문제는 MDP (Markov Decision Process)라는 확률 모델로 수식화한다. 매 단계마다 에이전트는 상태 s에서 확률적 정책 (Stochastic policy) π에 근거하여 행동 a를 취하고, 새로운 상태 s'에서 보상 r을 얻는다. S는 상태들의 유한 집합으로 정의되며, 매 시간 슬롯마다 $s(\in S)$가 확률적으로 변경된다. A는 행동들의 유한 집합으로 정의되며, 에이전트들은 상태 s에서 π에 근거하여 행동 집합에 속한 행동 중 하나를 선택한다. P는 전이 확률의 집합으로 정의되며, $P_a(s,s')$는 행동 상태 $s(\in S)$에서 행동 $a(\in A)$를 취하였을 때, 상태 $s'(s' \neq s \ \& \ s' \in S)$로 전이될 확률을 나타낸다. R은 얻을 수 있는 보상들의 유한 집합을 나타내며, $r_a(s,s')$는 상태 s에서 행동 a로 인하여 상태 s'로 전이되었을 때 얻는 보상으로 정의된다. 각 에이전트는 하나의 에피소드 동안 최종 보상이 가장 크도록 행동을 선택한다.

강화학습 모델을 무선 통신 환경에 적용하기 위해서 다음의 2가지 무선 통신 네트워크의 특징을 고려한다. 첫 번째, 모든 단말들의 그 어떠한 행동과 무관하게 단계마다 상태가 확률적으로 변경된다. 두 번째, 무선 통신의 현재 보상은 현재의 행동과 현재의 상태에 의하여 즉각적으로 결정된다. 위의 2가지 특징은 수학적으로 각각 모든 a, s, s'에 대해서 $P_a(s,s') = P(s,s')$와 $r_a(s,s') = r_a(s)$로 표현된다. 일반적으로 하나의 에피소드는 여러 단계로 구성되어 있고, 강화학습은 하나의 에피소드에 대하여 최종 보상이 최대가 되도록 학습한다. 반면에, 해당 무선 통신 네트워크의 특징을 고려할 때, 최종 단계의 행동 가치를 최대화하는 강화학습 문제를 개별 단계의 보상을 최대화하는 문제로 변환할 수 있다 [11]. 그리고 보상의 총합을 나타내는 Q 값은 벨만 방정식 (Bellman equation)에 의하여 $Q(s,a) = r_a(s) + \gamma \max_{a^t} Q(s',a')$로 주어진다. γ는 미래 보상의 중요도를 나타내는 할인계수 (discount factor)이다. 논문 [11]에서 $\gamma > 0$일 때, Q 값이 과대평가되고 γ가 DQN의 성능을 저하시킨다는 것을 증명하였다. 따라서 본 논문에서 하이퍼 파라미터인 할인계수를 $\gamma = 0$으로 설정한다.

Ⅲ. D2D 네트워크를 위한 DQN 기반 스케쥴링 기법

오버레이 D2D 네트워크는 기지국이 D2D 통신에 간섭하는 상황을 고려하지 않기 때문에, 분산적 알고리즘이 필수적이다. 따라서, 본 논문에서는 오버레이 D2D 네트워크에서의 DQN 기반 스케쥴링을 위한 분산형 알고리즘을 제안한다.

t번째 시간 슬롯에서 전송 단말 i는 수신 단말의 요청

메시지로부터 $\boldsymbol{h}_i^t$를 측정할 수 있다. 각 단계의 채널 상황에 따라서 전송 단말이 취해야 할 행동이 달라지기 때문에, 채널 정보는 행동 선택에서 매우 중요한 요소이다. 하지만, 제안하는 분산적 알고리즘의 경우, 모든 전송 단말들이 같은 심층 강화학습 네트워크 파라미터를 사용하기 때문에, $\boldsymbol{h}_i^t$를 그대로 사용하면 신호 채널 정보와 간섭 발생 채널 정보 간 구분이 어려워진다. 따라서 2가지의 채널 정보를 구분하기 위하여, 각 단말의 신호 채널을 가장 앞으로 위치시켜서 채널 정보를 $\tilde{\boldsymbol{h}}_i^t = \{h_{ii}, h_{1i}, \cdots, h_{(i-1)i}, h_{(i+1)i}, h_{(i+1)i} \cdots, h_{Ni}\}$와 같이 변환한다. 그리고 안정적 학습을 위하여, 다음과 같이 정규화된 채널 정보가 학습 네트워크의 입력으로 사용된다.

$$s_i^t := \boldsymbol{\Gamma}_i^t = \frac{1}{\gamma_{ii}^t}\{\tilde{\boldsymbol{h}}_i^t\} \tag{3}$$

행동 집합은 $A := \{0, 1\}$로 정의되며, 전송 단말 i는 행동 집합 내의 행동 중 하나를 선택하여 전송 여부를 결정한다 ($a_i^t \in A$). 그리고 시스템의 성능을 최대화하기 위해서는 단계마다 전송률의 합을 최대화하는 문제를 풀어야 한다. 따라서 전송 단말 i의 행동에 대한 보상은 $r_i^t := C(\boldsymbol{H}^t, \boldsymbol{\Gamma}^t)$로 결정되고, 모든 전송 단말의 보상들은 전송률의 합으로 모두 동일하다.

제안된 심층 강화학습 기반 D2D 스케쥴링 방식은 데이터 수집 및 학습 단계와 분산적 실행 단계로 나누어진다. 전송 단말들은 분산적으로 학습이 가능하지만, 빠른 학습을 위하여 중앙 컨트롤러가 전송 단말의 채널들을 모아서 한 번에 학습을 진행한다. 특정 시간 슬롯 동안 데이터 수집 단계와 학습 단계가 반복되고, 특정 시간 슬롯 이후, 중앙 컨트롤러는 최후에 업데이트된 DQN 파라미터들을 모든 전송 단말들에게 전파하고 분산적 실행 단계가 시작된다.

그림 1은 t번째 시간 슬롯에서 데이터 수집 및 DQN 학습 단계의 절차 구조를 나타낸다. 중앙 컨트롤러는 모든 전송 단말들로부터 모든 채널 정보를 수집한다. 신호 채널 정보와 간섭 채널 정보를 구분하기 위하여, $\boldsymbol{s}^t = \boldsymbol{\Gamma}^t = \{\Gamma_1^t, \cdots, \Gamma_{N_L}^t\}$를 생성한다. 중앙 컨트롤러는 입력 $\boldsymbol{s}^t$를 바탕으로 전송 단말 당 행동 개수만큼의 행동 가치 함수 $Q(\boldsymbol{s}^t, \boldsymbol{a}; \theta_q)$들을 DQN의 출력으로 얻는다. 이

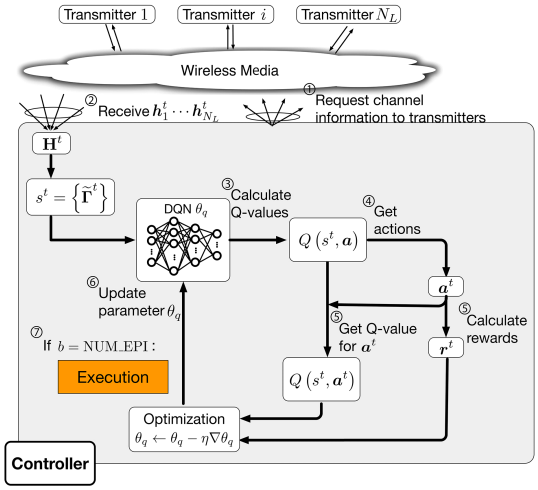


Fig. 1 The flow chart of data collection and learning process

때, DQN을 이용하여 예측한 행동 값 $\hat{\boldsymbol{a}}^t$는 전송 단말 당 가장 큰 $Q(\boldsymbol{s}^t, \boldsymbol{a}; \theta_q)$를 가지는 행동들로 결정된다. DQN 학습 단계 동안, 국소 최적해에 수렴하는 것을 방지하기 위하여 ϵ-greedy 탐색이 사용된다. 선택된 행동은 ϵ_t보다 작으면 $\hat{\boldsymbol{a}}^t$로 선택되고, ϵ_t보다 크면 모든 행동들 중에서 무작위로 행동을 하나 선택한다. ϵ_t는 아래와 같이 결정된다.

$$\epsilon_t = \epsilon_1 + \frac{t-1}{N_E - 1}(\epsilon_{N_E} - \epsilon_1) \tag{4}$$

ϵ_1과 ϵ_{N_E}는 각각 미리 설정된 확률의 최댓값과 최솟값을 나타낸다. DQN은 아래와 같이 행동에 대한 실제 전송률과 예측된 Q 값의 차이를 최소화하도록 학습한다.

$$\theta_q^* = \underset{\theta_q}{\arg\min} \frac{1}{2}(Q(s, a; \theta_q) - r_a(s))^2 \tag{5}$$

식 (5)을 구하기 위하여 기울기 하강 (gradient descent) 기법을 이용하고, θ_q에 대한 경사도 (gradient)를 구하는 식과 θ_q를 업데이트하는 방식은 다음과 같다.

$$\nabla \theta_q = \left(Q(s, a; \theta_q) - r_s^a\right) \nabla_{\theta_q} Q(s, a; \theta_q) \tag{6}$$

$$\theta_q \leftarrow \theta_q - \eta \nabla \theta_q \tag{7}$$

η는 학습률을 나타낸다. N_E번째 시간 슬롯 이후, 최

종적으로 중앙 컨트롤러에서 학습된 네트워크 파라미터 값은 모든 모바일로 전파되고, 분산적 실행 단계가 실행된다. 분산적 실행 단계에서 각 전송 단말은 모든 수신 단말들의 요청 메시지를 통하여 $\boldsymbol{h}_i^t$를 측정하고 s_i^t를 얻는다. s_i^t를 학습 네트워크에 입력하여 a_i^t를 결정한다. 특정 시간에 모든 전송 단말들은 각각 결정된 행동을 동시에 실행한다.

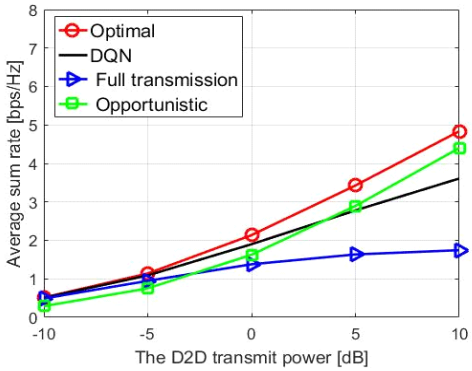
Ⅳ. 시뮬레이션 결과

본 장에서는 제안된 스케쥴링 알고리즘을 컴퓨터 시뮬레이션을 통하여 분석한다. 그리고 제안된 스케쥴링 알고리즘을 최적 방식, 채널 이득이 가장 큰 D2D 링크의 전송 단말만 전송하는 기회주의적 선택 방식, 모든 D2D 링크들이 동시에 통신하는 전체 전송 방식과 비교한다.

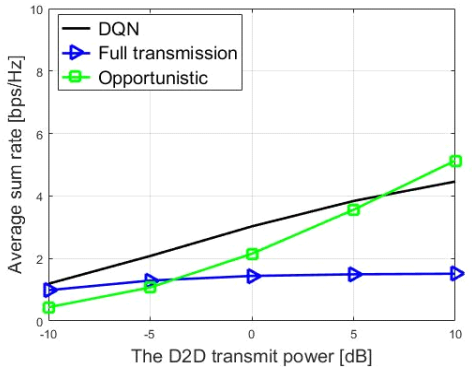
본 컴퓨터 시뮬레이션에서, DQN의 학습 모듈을 파이썬 텐서플로우 기반으로 개발하였다. 학습 네트워크는 입력층, 은닉층, 출력층 하나씩으로 구성되어 있고, 층 사이의 활성화 함수로 렐루 (Relu) 함수가 사용되었다. 학습률은 $\eta = 0.0001$로 주어졌으며, ϵ-greedy 탐색에 사용되는 확률 최솟값과 최댓값은 각각 $\epsilon_1 = 0.1$ $\epsilon_{N_E} = 0.00001$로 주어졌다.

그림 2는 $N_L = 5, 20$인 경우에, 전송 SNR에 따른 평균 전송률의 합을 나타낸다. $N_L = 20$일 때, 최적 방식의 평균 전송률은 매우 높은 계산 복잡도 때문에 컴퓨터 시뮬레이션으로 얻을 수 없다. 높은 SNR 구간에서, 제안 방식의 평균 전송률은 최적 방식과 기회주의적 방식보다 낮다. 하지만, 본 논문에서 사용되는 최적 방식과 기회주의적 선택 방식은 채널 상태 정보 수집과 최적 전송 조합을 결정하기 위한 별도의 중앙집중식 제어 장치를 요구하기 때문에, 실제 환경에서 구현이 어렵다. 높은 SNR 구간에서 D2D 링크의 수가 증가할 때, 제안 방식과 기회주의적 방식 간 성능 차이는 줄어든다. 예를 들면, $N_L = 5, \rho = 10$ [dB]인 경우에, 기회주의적 방식 대비 제안 방식의 비율이 80%인 반면에, $N_L = 20$, $\rho = 10$ [dB]인 경우에, 그 비율은 87%로 증가한다.

그림 3은 $\rho = 10$ [dB] 일 때, 제안 방식의 D2D 링크 개수에 따른 평균 전송률의 합을 나타낸다. 링크 개수가 적은 상황에서, 기회주의적 선택 방식이 제안 방식보다 더 좋은 성능을 보인다. 하지만, 링크 개수가 증가함에 따라서 성능 차이는 감소한다. 일정 링크 개수 이상에서는 제안 방식이 기회주의적 선택 방식보다 더 좋은 성능을 보인다.



(a) $N_L = 5$



(b) $N_L = 20$

Fig. 2 Average sum-rates for varying the D2D transmit SNR when $N_L = 5, 20$

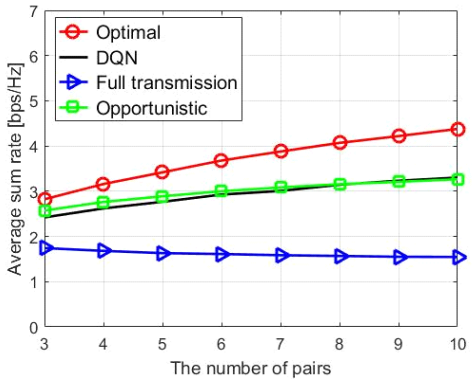


Fig. 3 Average sum-rates for varying the number of pairs when $\rho = 10$

Ⅴ. 결 론

본 논문에서는 오버레이 D2D 통신 네트워크에서 강화학습을 이용한 스케쥴링 문제를 연구하였다. 오버레이 D2D 통신 네트워크에 적합한 DQN 모델을 디자인하고, 피드백과 시그널링 오버헤드를 줄이기 위한 DQN 기반 분산적 스케쥴링 방식을 제안하였다. 학습 단계에서, 중앙 컨트롤러가 모든 전송 단말로부터 모든 채널 정보들을 수신하고, DQN 파라미터들을 학습시킨다. 특정 시간 슬롯 동안 학습이 진행된 이후에, 중앙 컨트롤러가 최종적으로 업데이트된 파라미터들을 모든 전송 단말들에게 전달한다. 분산적 실행 단계에서, 모든 단말들은 최종 업데이트된 파라미터를 이용하여 각자의 행동을 개별적으로 판단한다. 제안된 방식을 컴퓨터 시뮬레이션을 통하여 분석하였고, 최적 방식, 기회주의적 선택 방식, 전체 전송 방식과 비교하였다. 낮은 전송 전력에서는, 제안된 방식은 기회주의적 선택 방식과 전체 전송 방식보다 높은 평균 전송률을 나타내고, 제안 방식과 최적 방식 간 성능 차이가 크지 않다. 높은 전력 수준에서, 중앙집중식인 기회주의적 선택 방식의 평균 전송률이 제안 방식보다 높지만, 전체 D2D 링크의 수가 증가할수록 두 방식 간 성능 차이는 줄어든다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(Ministry of Education) (No. 2020R1I1A3061195, Development Of Wired and Wireless Integrated Multimedia-Streaming System Using Exclusive OR-based Coding).

REFERENCES

[1] M. Sheng, H. Sun, X. Wang, Y. Zhang, T. Q. S. Quek, J. Liu, and J. Li, "Ondemand scheduling: achieving QoS dierentiation for D2D communications," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 162-170, Jul. 2015.

[2] J. Lyu, Y. H. Chew, and W.-C. Wong, "A Stackelberg Game Model for Overlay D2D Transmission With Heterogeneous Rate Requirements," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8461-8475, Oct. 2016.

[3] J. Xu and C. Guo, "Scheduling Stochastic Real-Time D2D Communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6022-6036, Jun. 2019.

[4] Siba Narayan Swain, Rahul Thakur, and C. Siva Ram Murthy, "Design and stochastic geometric analysis of an efficient Q-Learning based physical resource block allocation scheme to maximize the spectral efficiency of Device-to-Device overlaid cellular networks," *Computer Networks*, vol. 119, pp. 71-85, Mar. 2017.

[5] Z. Fan, X. Gu, S. Nie, and M. Chen, "D2D power control based on supervised and unsupervised learning," *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, pp. 558-563, 2017.

[6] X. Fang, T. Zhang, Y. Liu, and Z. Zeng, "Multi-Agent Cooperative Alternating Q-Learning Caching in D2D-Enabled Cellular Networks," *2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, pp. 1-6, 2019.

[7] J. Yin, L. Li, Y. Xu, W. Liang, H. Zhang, and Z. Han, "Joint Content Popularity Prediction and Content Delivery Policy for Cache- Enabled D2D Networks: A Deep Reinforcement Learning Approach," *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Anaheim, CA, USA, pp. 609-613, 2018.

[8] J. Tang, H. Tang, X. Zhang, K. Cumanan, G. Chen, K.-K. Wong, J. A. Chambers, "Energy Minimization in D2D-Assisted Cache-Enabled Internet of Things: A Deep Reinforcement Learning Approach," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5412-5423, Aug. 2020.

[9] R. Li, Y. Zhao, C. Wang, X. Wang, V. C. M. Leung, X. Li, T. Taleb, "Edge Caching Replacement Optimization for D2D Wireless Networks via Weighted Distributed DQN," *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, Seoul, Korea (South), pp. 1-6, 2020.

[10] T. Ban and B. C. Jung, "On the Link Scheduling for Cellular-Aided Device-to-Device Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 9404-9409, Nov. 2016. doi: 10.1109/TVT.2016.2519461.

[11] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power Allocation in Multi-User Cellular Networks: Deep Reinforcement Learning Approaches," *IEEE Transactions on Wireless Communications*, 2020. doi: 10.1109/TWC.2020.3001736.

정무웅(Moo-Woong Jeong)

2014년 2월 경상대학교 정보통신공학과 학사
2014년 3월~2020년 8월 경상대학교 정보통신공학과 박사
2020년 9월~현재 경상대학교 박사후 연구과정
※관심분야 : 차세대이동통신, MIMO, D2D, Reinforcement Learning

김륜우(Lyun Woo Kim)

2014년 02월~현재 경상대학교 정보통신공학과 학사
※관심분야 : D2D, Machine Learning

반태원(Tae-Won Ban)

1998년 2월 경북대학교 전자공학과 학사
2000년 2월 경북대학교 전자공학과 석사
2010년 2월 KAIST 전기및전자공학과 박사
2000년 2월~2012년 8월 KT 네트워크부문
2012년 9월~현재 경상대학교 정보통신공학과 부교수
※관심분야 : 이동통신, 자원관리, 간섭관리, 협력 및 중계통신, 인지통신, 주파수 공유, 차세대 이동통신 시스템