https://doi.org/10.7840/kics.2023.48.2.196

저궤도 인공위성에 적용되는 심층강화학습 기술 동향

이 현 수◆, 김 중 헌°

Survey on Deep Reinforcement Learning Applied for LEO Satellites

Hyunsoo Lee◆, Joongheon Kim°

요 약

빠른 통신기술의 발전으로, 최근 차세대 이동통신에서 저궤도 위성을 사용하는 연구가 활발히 이루어지고 있다. 저궤도 위성통신에서는 자원 관리, 핸드오버 문제 등 복잡한 문제가 발생하는데, 심층강화학습을 통해 기존의 방법으로 풀기 어려웠던 문제들을 해결할 수 있다. 본 논문에서는 위성통신에서 심층강화학습이 사용된 기술에 대해 살펴본다. 크게 Task의 순서를 배정하는 문제인 Scheduling problem, 한정된 자원을 효율적으로 관리하기 위한 문제인 Resource Allocation, 위성 사용자들의 연결을 원활하게 만들거나 사용자들을 적절하게 분배하기 위한 문제인 User Access Control, 그리고 그 외의 문제로 분류하였다.

키워드 : 저궤도 위성, 심층 강화학습, 에지 컴퓨팅, 최적화, SAGIN
Key Words : LEO Satellite, Deep Reinforcement Learning, Edge Computing, Optimization, SAGIN

ABSTRACT

With the rapid development of communication technology, research using LEO satellites in next-generation mobile communication is being actively conducted. In LEO satellite communication, complex problems such as resource management or handover occur, so deep reinforcement learning method can fix the problems that were difficult to solve with conventional methods. In this paper, we investigated the cases that applied deep reinforcement learning method in satellite communication. It is largely classified into scheduling problem, resource allocation, user access control, and the other problems.

Ⅰ. 서 론

최근 5G 기술을 비롯한 통신 기술이 급격하게 발전함에 따라 통신 원천 기술 개발이 빠르게 진행되고 있다. 위성통신의 경우, 지상에 새로운 기지국을 건설하는 것에 비해 비용이 훨씬 저렴하고, 기존에 있는 지상 기지국과의 통합된 네트워크를 통해 5G를 지원할 수 있으며, 범국가적으로 사용할 수 있다는 점에서 더욱 각광받고 있다. 위성통신 시스템에는 크게 거리를 기준으로, 그림 1에서 보는 것처럼 GEO(Geostationary Earth Orbit), MEO(Medium Earth Orbit), LEO(Low Earth Orbit) 위성이 있다. GEO 위성은 36000km 상공에 위치하며, 이 궤도에 있는 위성들은 지구의 자전 속도와 동일한 공전 속도로 회전하기 때문에, 지상에서 볼 때는 한 지점에 정지해 있는 것으로 보여 정지궤도 위성이라고도 부른

※ 이 성과는 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2022R1A2C2004869).
◆ First Author : Korea University Department of Electrical and Computer Engineering, hyunsoo@korea.ac.kr, 학생회원
° Corresponding Author : Korea University School of Electrical Engineering, joongheon@korea.ac.kr, 종신회원
논문번호 : 202211-283-B-RU, Received November 23, 2022; Revised December 15, 2022; Accepted December 16, 2022

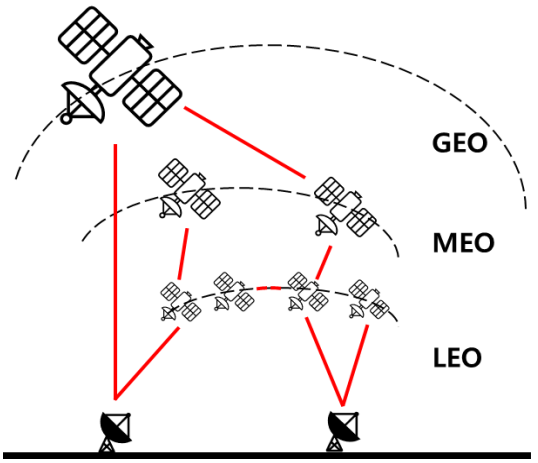

Fig. 1. Conceptual Diagram of LEO, MEO, GEO Satellite

다. 지연 시간이나 Handover의 변동이 적지만, 위성의 발사 및 유지비용이 크다는 단점이 있다. MEO는 2000-20000km 상공에 위치하는 궤도 위성으로, GEO보다 전파 지연이 적지만, 한 지역을 지속적으로 서비스하기 위해 필요한 위성의 개수가 GEO보다 더 많이 필요하다. LEO 위성은 500-2000km 상공에 위치하는데, 지연 시간, 경로 손실, 생산 및 발사 비용이 GEO 및 MEO 위성에 비해 상대적으로 적어서, 산업용으로 널리 사용되고 있다. Tesla사의 CEO 일론 머스크가 이끄는 SpaceX사의 경우, 현재까지 2개의 궤도에 2000여 개의 위성을 이미 발사했고, 최종적으로 40000여 개의 위성을 발사하여 위성 간 연결을 통한 인터넷 서비스를 진행할 것으로 알려져 있다[1].

강화학습은 MDP(Markov Decision Process) 최적화 기법을 이용한 Machine learning 방법의 하나로, 여러 최적화 문제를 푸는데 적합한 기법이다. 기존의 지도 학습과 비지도 학습이 주어진 데이터로 문제를 푼다고 하면, 강화학습은 스스로 데이터를 만들어서 학습을 진행하는 것이다. 어떤 환경(Environment)에 주어진 에이전트(Agent)가 취하는 행동(Action)에 따라, 에이전트의 상태(State)와 보상(Reward)가 달라지는데, 이때의 보상을 최대화시키는 정책(Policy)를 찾는 것을 목표로 하는 것이 강화학습이다[2]. 특히 강화학습에 DNN(Deep Neural Network)를 결합한 심층 강화학습(Deep Reinforcement Learning)은 복잡한 최적화 문제를 해결하는 데 탁월하여, 경로 탐색이나 자원 관리, 무인기 제어 등 통신 분야뿐만 아니라 자율주행, 게임, 금융에 이르기까지 다양한 분야에서 연구가 활발하다[3]. 본 논문에서는 저궤도 인공위성에 심층 강화학습 기술을 적용한 최신 동향을 다양한 측면에서 설명한다.

Ⅱ. 전통적인 LEO 위성의 최적화 방법

지구 관측에서의 Scheduling 문제는 NP-hard에 속하여, 이론적으로 단일 heuristic 알고리즘으로 정확한 해를 구하기 어렵다. 실제로 기존에 제안된 알고리즘의 경우, 소형화하거나 제약 조건을 상세히 둔 지표들을 통해서 문제를 해결하는 경향을 보였다[4,5]. 대부분의 심층 연구는 Greedy algorithm[4], Genetic algorithm[6] 등의 Heuristic algorithm 또는 Meta-Heuristic algorithm을 중점에 두었다. Heuristic rule은 좋은 솔루션을 얻는 방법이지만, 일반적인 경우에서 얻기 어려워 많은 변인을 통제하여 실험을 진행하기 때문에 응용 범위가 제한된다. 따라서 실제 상황에서 원하는 결과를 얻기 위해서는 현재의 Heuristic rule보다 좋은 방법이 필요하다.

최근 5G를 비롯한 통신 기술의 발달로, IoT 장치는 수많은 애플리케이션 및 서비스에서 중요한 역할을 하게 되었다[7]. 특히, IoT 장치를 배포하여 환경을 감시하여 산업의 자동화, 지능형 교통 관리 등 새로운 사업 분야가 탄생하기도 했다. 지능형 교통 관리는 매우 동적인 입력을 빠르게 처리해야 하므로 지연 지향 사물인터넷 작업 스케줄링(Delay-Oriented IoT Task Scheduling) 방법을 사용하게 되는데, IoT 장치의 한정된 연산량 때문에 카메라 내의 이미지 또는 영상 처리 같은 작업을 장치에서 수행하게 되면 심각한 수준의 서비스 지연이 발생하고, 장치의 수명에도 해를 끼칠 수 있다. 따라서 이를 해결하기 위해 에지 컴퓨팅(Edge computing)을 도입하게 되었다[8]. 에지 컴퓨팅을 통해 IoT장치는 연산 작업을 인근의 지상 기지국(Base station)으로 오프로드할 수 있어 작업 실행 대기 시간과 IoT 장치의 전력 소비를 동시에 줄일 수 있다.

위성에서는 무선 자원이 충분하지 않기 때문에, 전력이나 주파수 등 자원을 효율적으로 관리하는 것은 중요하다. 그에 따라, 한정된 무선 자원을 요청된 용량에 맞게 유연하게 할당하는 방법이 연구되고 있다. 이를 해결하기 위한 Resource allocation 문제에 있어서, 전송 효율과 광대역 커버리지는 가장 중요한 요소에 속한다. 효율적인 전송과 넓은 수신 지역 확보를 위해 Multi-Beam Satellite(MBS)가 도입되었다. 전통적인 동적 대역폭 할당 방법은, Simulated Annealing(SA) 알고리즘을 사용하여 빔에 걸쳐서 대역폭을 구성한다[9,10]. 특히 동적 대역폭 할당 방법은 빔 전반에 걸쳐 요청된 주파수 용량을 만족시키는 데

적합함을 알 수 있다.

최근 수년간의 무선통신망 사용자의 증가로, 기존의 지상 네트워크만으로는 급증하는 장치에 대한 서비스의 보장이 어려워졌다. 이를 해결하기 위해 지상 기지국(이동 기지국)을 늘리는 것은 공간 차지와 비용 문제 때문에 적합하지 않다. 따라서 그림 2와 같이 High Altitude Platform Station(HAPS)나 Unmanned Aerial Vehicle(UAV)을 이용한 중계 통신을 통해 액세스 제어에 도움을 주는 방법이 활발히 연구되고 있다. 한편 기존의 방법으로는 User Equipment(UE)가 Received Signal Strength(RSS)가 가장 강한 기지국으로 접속하는 경향이 있었다[11]. 하지만, HAPS와 UAV 등 빠르게 움직이는 이동 기지국(Mobile Base Station)이 있는 경우, 기지국과 사용자 단말기 사이의 RSS가 급격하게 변할 수 있어 기지국 간의 Handover가 많이 발생하고, 이는 잦은 전송 중단으로 인한 QoS의 감소를 야기한다. 그리고 많은 수의 사용자 단말기가 높은 RSS를 가진 한 기지국에 연결될 경우, 특정 기지국에서 무선 액세스 혼잡이 발생하여 처리량이 낮아지는 문제가 발생할 수 있다. 따라서 많은 기지국이 있는 경우 다중 사용자 접속을 조정하기 위한 해결책이 필요하다.

HAPS나 UAV처럼 빠르게 움직이는 Base station을 사용하는 경우에는 사용자 액세스를 제어할 때 이러한 NT-BS(Non-Terrestrial Base Station)의 궤적을 고려해야 한다. NT-BS는 일반적으로는 원이나 타원 등 특정한 궤도로 움직이지만, 사용자들과의 원활한 연결을 위해 최적화된 궤적으로 움직이기도 한다[12-14]. NT-BS간의 통신은 셀룰러 시스템의 형태를 띤

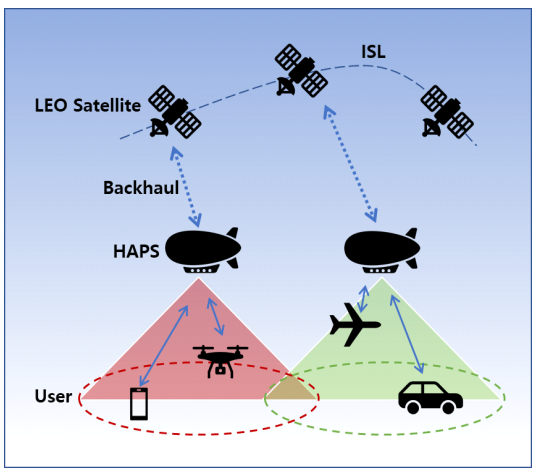

Fig. 2. Relay communication between satellites, HAPS, and users

NTN(Non-Terrestrial Network)를 통해서 연결되는데, 이런 통신 환경에서 NT-BS간의 액세스 용량을 효율적으로 분배하는 문제도 활발히 연구되고 있다.

원활한 자원 활용을 위해, 서로 다른 위성 시스템을 구축한 후 통합하는 것도 하나의 방법이 될 수 있다. 예를 들면, 중계 위성 시스템은 빔을 Scheduling하여 데이터를 반환하고, 통신 위성 시스템은 빔 할당, 부반송파 할당 및 전력 할당을 포함한 다차원 자원 할당(Multidimension Resource Allocation)을 사용하여 사용자에게 서비스를 제공한다. SDN(Software Defined Network)이나 NFV(Network Function Virtualization) 등을 통해, 위성 간의 링크(ISL, Inter-Satellite Link) 없이 상호 통신을 할 수 있는 방법이 제안되기도 했다[15-17].

Ⅲ. 강화학습 개요

강화학습은 기계 학습의 한 종류로, 특정 환경에서 행동을 취했을 때, 이 행동이 적합한 행동일지, 잘못된 행동일지를 나중에 판단하고 그에 대해 보상을 제공함으로써 반복을 통해 스스로 학습하도록 하는 방법이다[18]. 강화학습에는 에이전트와 환경이라는 두 가지 구성 요소가 존재한다. 에이전트는 주어진 환경에서 자신의 행동을 결정하고, 환경은 그에 맞는 보상을 제공한다. 이러한 보상은 즉각적인 보상(Immediate reward) 또는 누적 보상(Cumulative reward)의 형태로 주어지는데, 일반적으로 여러 행동들을 취한 뒤에 보상이 주어지는 형태로 결정된다.

심층강화학습은 강화학습에 신경망을 결합한 방식이다. 2013년 영국 DeepMind사에서 Deep Q-Network (DQN)를 개발하며, 게임 분야에서 사람을 훨씬 능가하는 인공지능을 개발한 것이 그 시초이다[19]. 최초의 심층강화학습 방법인 DQN은 에이전트가 행동가치 함수에 근거하여 행동을 선택하는 가치 기반 알고리즘으로, 손실 함수를 기반으로 가치 네트워크를 학습시켜 최적의 정책을 찾아내는 방식이다. 그동안 컴퓨터가 정복할 수 없다고 여겨졌던 영역이었던 바둑도 2016년 DeepMind사의 알파고(AlphaGo)의 승리로 끝난 것을 비롯하여, 이후 다양한 심층강화학습 알고리즘의 개발이 진행되면서 그동안 해결할 수 없었던 현실 세계의 광범위한 문제들이 해결되고 있다.

싱글에이전트 강화학습은 환경 내에 최적화시킬 에이전트가 단일 에이전트인 경우를 말하며, 여러 에이전트가 협력하여 동시에 최적화를 진행하는 경우는 멀티에이전트 강화학습이라 부른다. 강화학습이 실제로 적용되는 분야인 자동차, 로보틱스 등의 분야를 고

려할 때 다수의 에이전트를 고려하는 것은 필수적이기 때문에, 멀티에이전트 강화학습에 대한 연구는 필수적이다. 이 때 기존의 방식처럼 중앙 집중형(Centralized) 방식으로 수행한다면, 공동 행동공간(Joint action space)의 크기가 매우 커지기 때문에 학습이 어렵고, 분산형(Decentralized) 방식으로 접근한다면 연산 능력 감소 및 협력 데이터 감소로 인해 원활하게 학습이 이루어지기 어렵다. 따라서, 알고리즘의 훈련 과정에서는 모든 상태 정보를 이용하고, 실행 시에는 각 에이전트의 상태 정보만을 활용하는 중앙 집중형 훈련, 분산형 실행 (CTDE) 방식이 주목받고 있다.

본 논문에서는 저궤도 위성 분야에 심층강화학습을 적용한 연구로 한정했지만, 강화학습 방식은 높은 성능을 무기로 다양한 연구에 활용되고 있다. 모빌리티 분야에서는, 여러 대의 자율주행 드론이 협력하면서 일정 지역을 감시하는 체계를 확립하면서, 에너지 효율을 최적화하기 위해 강화학습 방법이 사용되었다[20]. 또한 의료 딥러닝 분야에서는, 환자에게 마취를 진행할 때 얼마만큼의 프로포폴을 주입해야 하는지에 대한 문제를 강화학습으로 풀기도 했다[21]. 이처럼 심층강화학습은 바둑처럼 행동 공간이 매우 큰 문제나, 드론의 이동 혹은 약물 주사와 같이 에이전트가 연속적인 행동 공간을 갖는 문제에 사용하기 적합한 방법이다.

Ⅳ. 강화학습을 이용한 LEO 위성의 최적화 방법

기존의 위성통신에서의 최적화 방법에 비해, 심층강화학습을 이용할 경우 많은 이점을 얻을 수 있다. 특히, LEO보다 낮은 궤도에서 외진 지역과 통신량이 많은 지역을 모두 커버할 수 있는 HAPS의 활용이 두드러지고 있다[22]. 드론이나 HAPS처럼 위성과 다른 종류의 플랫폼 사이의 통신을 하게 되는 이기종 모델의 경우, 동적 모델링을 통해 Scalability를 낮출 수 있어 성능 면에서 우수하다. 또한, 심층강화학습에서 사용하는 MDP(Markov Decision Process)가 가져다주는 이점이 크다. MDP는 시간이 지남에 따라 상태가 변하는 행동을 취하고, 그에 따른 보상을 얻게 되는 프로세스이다. 따라서 기존에의 방법이 새로운 이벤트 발생에 대처가 어려운 반면, 심층강화학습을 이용하면 이러한 MDP의 특성을 통해 새로운 이벤트 발생에 유연하게 대처할 수 있다. MDP는 주어진 환경에서 최대의 보상을 얻을 수 있는 최적의 정책을 학습하는 것을 목표로 하는데, 기존의 NP-Hard 최적화 문제의 추론 시간을 줄여 Real-Time을 보장할 수 있다.

4.1 Scheduling

Scheduling은 위성에서 처리할 task의 순서를 배정하는 문제로, 이를 해결하기 위해 멀티에이전트 심층강화학습을 기반으로 온라인 분산 위성 협력 관측 스케줄링 알고리즘이 제안되었다[23]. 실시간 서비스는 위성의 중요한 기능 중 하나로, 분산형 스케줄링을 사용할 때 중앙집중형 스케줄링보다 더 확장성이 좋다. 그러나 기존의 분산 위성 스케줄링은 작업을 조정하기 위해 위성 간에 매우 많은 양의 통신이 필요하기 때문에 실시간 스케줄링을 지원하기 어려웠다. 이를 해결하기 위해 다중 에이전트 심층 강화학습(MADRL)을 이용한 해결 방안이 제시되었다. 이를 이용하면 위성들이 각각의 결정 정책을 공유하지만, 결정을 내리는 데이터나 내부 상태에 관한 데이터를 공유할 필요가 없기 때문에, 통신량이 줄어들어 분산 스케줄링을 원활하게 진행할 수 있게 한다. 기존의 방법인 CNP(Contract Net Protocol), 그리고 모든 미래의 작업과 저장소의 소비를 확인할 수 있는 이상적인 경우인 OMN(hypothetical OMNiscient scheduler)과 비교했을 때, MADRL을 이용한 방법의 총 이득은 CNP에 비해 약 5% 향상되었고, 이상적인 성능에 93.5%에 도달하였다. 에이전트가 우선 순위가 높은 작업부터 진행하도록 하기 때문에 기존의 방법에 비해 성능이 향상되었다. CNP는 중복을 피하기 위해 계산을 위한 중심 노드를 항상 유지해야 하기 때문에 MADRL보다 많은 작업을 수행하지만, CNP는 향후 도착할 높은 우선 순위의 작업을 미리 예측할 수 없기 때문에 심층강화학습 알고리즘의 경우 더 나은 성능을 보여주는 것을 확인할 수 있다.

최적의 지구 관측 계획을 개발하기 위해, AEOSSP(Agile Earth Observation Satellite Scheduling Problem)을 정의하여, Planning 및 Scheduling 혼합 문제를 강화학습 알고리즘을 통해 최대의 총 보상을 갖게 되도록 설계하였다[24]. 기존에는 이 문제를 풀기 위해 Heuristic algorithm을 이용했지만, Heuristic algorithm의 경우에는 사용하는 데이터의 규모가 확장되면 복잡도가 크게 상승하여 좋은 솔루션을 찾기가 어렵다. 본 모델에서는 2계층의 Attention mechanism을 가진 Encoder-decoder 구조에 RNN(Recurrent Neural Network)을 활용한다. 논문에 사용된 Attention mechanism은 목표하는 작업과 입력되는 작업 간의 연관성을 분석하여 확률분포에 반영하는 역할을 수행한다. 가장 연관성이 높은 작업

은 높은 Attention을 받아 다음 작업으로 채택될 수 있다. Training scale을 각각 20, 50, 100으로 다르게 한 결과, 동일한 알고리즘 실행 시간에서도, 작업량이 가장 많은 경우에 특히 성능이 두드러지게 증가하였다.

기존의 에지 컴퓨팅에서 지상 기지국으로의 오프로드에만 의존하는 것은 IoT 에지 컴퓨팅의 성능을 확실하게 보장하기 어렵고, 특히 이동 기지국의 수가 적은 지역에 있거나 BS 근방에서 사용할 수 없는 경우, 일반적인 IoT 장치는 전력이 크지 않아 작업 오프로드를 위한 장거리 전송을 지원할 수 없다. 릴레이 통신을 지원하기 위해, Space-Air-Ground 통합 네트워크의 IoT Task Scheduling 문제를 해결하는 데 심층 강화학습이 적용되었다[8]. UAV가 IoT 장치에서 연산 작업을 수집하면, UAV는 작업을 직접 처리할지, 또는 위성이나 HAPS 등의 이동 기지국으로 작업을 분배할지를 결정하여 위성과 UAV로 지상 네트워크를 강화하게 되었다. UAV의 에너지 용량이 한정되어 있기 때문에, 작업을 직접 처리하거나 분배하는 문제에 있어서 계산 지연을 최소화하기 위해 작업 스케줄링 정책을 설계해야 한다. 또한 지상 기지국과 UAV, 위성의 작업 처리 능력이 각기 다르기 때문에, 기능에 따라 적절한 작업을 분배해 주어야 하고, 각 Agent가 현재 도착한 작업을 처리할 에너지 소비량과 앞으로 도착할 예정인 작업을 처리할 에너지의 소비 계획을 모두 고려해야 한다. 이를 위해 DOTS (Delay-Orientated IoT Scheduling) 구조를 도입하였는데, 이 구조는 그림 3에서 볼 수 있다. UAV는 IoT 장치에서 작업을 수집하고, 컴퓨팅 대기열(Computing queue) 내에서 작업을 로컬로 처리한다. 처리 혹은 오프로드되지 않은 작업들은 UAV의 컴퓨팅 대기열에 머무른다. UAV가 컴퓨팅 대기열에서 BS나 위성으로 작업을 오프로드할 수도 있다. 이때 전달되지 않은 작업들은 전달 대기열(Forwarding Queue)에 머무른다. IoT 장치에서 새로 도착한 작업은 UAV의 컴퓨팅 대기열에 저장되고, 컴퓨팅 대기열이 가득 차면 새로 도착한 작업이 삭제된다. UAV는 계속 사전 지정된 경로를 따라 비행하면서 다음 작업을 찾는다. 수집된 작업을 원활하게 분배하기 위해, Online scheduling 문제를 CMDP (Constraint Markov Decision Process) 문제로 공식화하고, 심층강화학습 알고리즘을 설계한다. CMDP를 사용함으로써 UAV가 통신하고 계산하는 데 사용하는 에너지 용량을 반영하여 작업 지연에 소요되는 평균 시간을 최소화할 수 있다. 심층강화학습 알고리즘은 각 state에 대해 UAV가 소비하는 에너지 용량을 초과하는지를 측정하는 변수 'risk'의 값을 평가하고, 이에 따라 최적의 정책을 학습하면서 지연과 'risk'를 최소화하는 최적의 변수를 찾게 된다.

작업의 중요도가 다른 경우, 더 중요한 작업을 우선

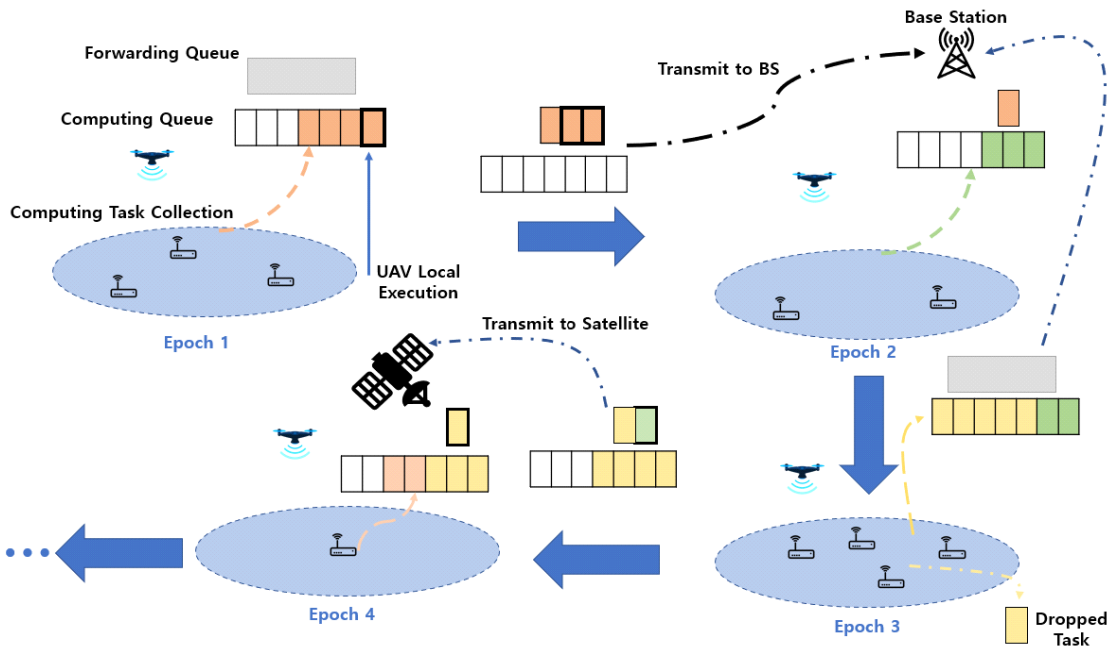

Fig. 3. Delay-Oriented IoT Task Scheduling Scheme

적으로 처리하는 것이 작업의 효율을 높일 수 있다. 지구 관측 위성의 경우에, 긴급한 작업에 대한 응답 속도를 향상시키기 위해 강화학습 알고리즘을 사용하였다[25]. 위성의 컴퓨팅 자원이 제한되어 기존의 알고리즘으로는 실시간 Scheduling을 실현하기 어렵기 때문에, Scheduling의 안정성을 향상시키고, 계산 복잡도를 줄이기 위해, 두 개의 층으로 구성된 계층 강화학습이 채택되었다. 각 계층에 대한 MDP를 기반으로 Scheduling 모델을 만들고, 강화학습 알고리즘의 하나인 Q-Learning을 통해 모델을 최적화하였다. 위성은 정해진 전략에 따라 긴급한 작업이 무작위로 도착하면 현재 환경에 대한 상태 판단을 내린다. 그 후 Q-Matrix와 결합된 탐색 정책을 수립하고, 위성은 이 정책에 따라 정해진 상태 s에서 행동 a를 선택한다. 선택한 동작을 실행 후, 위성은 주어진 식 (1)에 따라 q값을 업데이트하는 데 사용되는 $R^{a_{ss}}$라는 보상을 받게 된다.

$$Q_{l+1}(s,a) \leftarrow (1-\alpha) Q_l(s,a) + \alpha[r + \gamma \max Q_l(s',a')] \quad (1)$$

위성은 작업 스케줄링을 수행하고, 환경은 위성의 행동에 대한 보상으로 피드백을 제공한다. 여러 번의 피드백을 거친 후, 위성은 가장 큰 보상을 얻을 수 있는 동작을 선택하게 된다.

4.2 Resource Allocation

Resource allocation은 위성에서 활용할 수 있는 주파수, 채널 등 한정된 자원을 효율적으로 관리하기 위한 문제이다. Resource allocation의 한 예로, 위성에서 통신을 위해 쏘는 빔 간의 간섭 완화와 효율적인 자원 분배를 통해 전송 효율을 높이고 트래픽의 복잡도를 낮추는 방법이 제안되었다[26]. 대규모 위성통신의 경우, 복잡한 채널 조건과 트래픽 부하로 인해 최적화 과정의 복잡도가 증가하여 MBS의 계산 복잡도가 커진다. 빔 간의 간섭을 줄이기 위해 위성 빔의 크기를 크게 하면, 단일 에이전트, 즉 위성의 심층 강화학습을 위한 행동 공간의 크기가 커지기 때문에, 시간 복잡도가 증가한다. 따라서, 단일 에이전트를 멀티 에이전트로 확장하여 게임이론의 협력 게임(Cooperative game), 협상 게임(Bargaining game)을 이용한 Cooperative Multi-agent 강화학습을 설계한다. 이 게임에서는 MBS의 각 빔이 DQN의 에이전트의 역할을 하며, 이 빔들은 게임 과정에서의 플레이어가 된다. 협력 게임 방식에서, 플레이어들은 빔 간의 정보 공유를 통해 모든 플레이어의 이익을 극대화하는 방향으로 결정을 내리게 되어, 이를 통해 최적의 대역폭 할당 방법을 찾게 된다. 이를 통해 전송 효율을 높이고, 복잡도를 낮추어 성능을 향상시킬 수 있다.

위성 사물인터넷(SIoT, Satellite Internet of Things)은 다수의 LEO 위성으로 구성된 우주 공간 네트워크로, SIoT에서 에너지 측면을 기준으로 효율적인 채널 할당을 위해 심층강화학습 기반의 'DeepCA' 알고리즘이 연구되었다[27]. 이 알고리즘은 LEO 위성의 동적 기능 모델링을 용이하게 하는 Sliding block scheme을 도입하고, SIoT의 동적 채널 할당 문제를 MDP로 표현한다. 그런 다음 최적의 채널 할당을 위한 심층강화학습 알고리즘이 적용된다. 위성은 각 시간 슬롯에서 모든 노드의 대역폭 자원 요청과 현재의 채널 할당 상태를 관찰하고, 에이전트는 이 결과를 바탕으로 신경망을 통해 전력 할당 작업을 수행한다. 이때 신경망에 대한 입력으로 상태공간을 정규화한다. 전력을 할당한 후에 에이전트는 그에 따른 보상을 받고, 노드의 현재 채널 상태 및 대역폭 수요를 재확인하며 다음 상태로 이동하는 과정을 거쳐 각 노드의 요청에 따른 가장 높은 보상을 받을 수 있는 채널을 할당한다.

다중 빔 전송 안테나를 갖춘 위성에서의 동적 자원 할당 문제를 풀기 위해 심층강화학습 방법이 적용되었다[28]. 동적 자원 할당은 네트워크 성능을 향상시키는 핵심적인 기술로, 예상되는 장기적 자원 활용을 최대화하는 정책을 찾는 것이다. 기존의 동적 자원 할당 문제는 OBP(On-Board Payload)와 다수의 빔으로 높은 데이터 속도와 대용량 서비스를 충족시킬 수 있었지만, 빔간 간섭 때문에 서비스 품질(QoS)이 저하되는 문제점이 있었다. 그림 4에서 묘사한 것처럼 위성을 에이전트로, User terminal을 환경으로 모델링하는 심층강화학습 프레임워크(Deep Reinforcement Learning Framework)를 제안하여, 강화학습 환경을 기반으로 상태 정보를 받아 이를 이미지 텐서로 변환하는 State reformulation을 진행한 후, 그 결과를 Deep Q-Network에 전달하면 심층강화학습을 진행한 뒤 위성이 행동을 선택하도록 한다.

HSN(Heterogeneous Satellite Communication Network)은 GEO와 LEO 등 다른 고도에서 운용되는 위성 시스템 간의 협력을 수행하는 네트워크로, HSN에서의 자원 활용을 관리하는 방법에 관한 연구가 수행되었다[29]. 메모리 용량이나 에너지 소비, 궤도 역학 등 예측하지 못한 요인들이 QoS 또는 자원 관리 문제에 영향을 줄 수 있는데, 다양한 목표를 동시에 최적화하기 위해 합성 목표함수를 만들어 여러 목표들 사

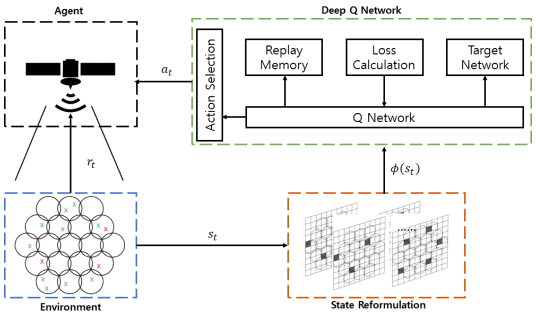

Fig. 4. Deep Reinforcement Learning Framework in beamforming method

이의 균형을 맞추기 위해 Policy network와 Q network를 활용한다. MDP를 기반으로, DRL은 각 단계에서 Local Optimal Policy를 선택하고, 그 후에 전체 정책에 따라 DRL은 이익 극대화에 도달할 때까지 지속적으로 결정을 최적화하게 된다. 이 Multi-objective learning 알고리즘에서는 적어도 2개 이상의 파레토 최적 해를 얻을 수 있고, Multi-objective learning 알고리즘을 여러 위성이 지능적으로 협력할 수 있도록 Multi-agent로 확장할 수 있다. Multi-agent 환경에서는 에이전트의 다양한 변수들 때문에 State-action space가 커져 MARL(Multi-agent Reinforcement Learning)의 복잡도가 크게 증가하고, 각 에이전트가 다른 에이전트들과 서로 경쟁 또는 협력하면서 에이전트들의 학습이 고르게 이루어지지 않을 수 있다. 이때 에이전트들의 학습 안정도를 높이고 계산 용량을 낮추기 위해, 진행되는 훈련은 중심에서 진행하고, 실행 단계는 분산된 방식을 채택하면 더 나은 성능을 가질 수 있다.

4.3 User Access Control

User Access Control은 지상이나 HAPS등 위성 사용자들의 연결을 원활하게 만들거나, Capacity에 따라 적절하게 연결을 분배하는 문제에 심층강화학습을 적용한 사례이다. NT-BS(Non-Terrestrial Base Station)은 드론이나 비행선, 위성 등에 탑재된 이동 기지국을 일컫는 말로, 지상의 고정 기지국과 다르게 빠른 속도로 이동하기 때문에, 통신 환경에서 Handover가 자주 발생하게 된다[30]. 잦은 Handover로 인해 QoS가 낮아지는 것을 해결하기 위해서, NT-BS의 Backhaul에서 중앙집중형 에이전트가 DQN의 매개변수 훈련을 진행하고, 매개변수는 NT-BS에서 UE로 전달되어, UE에서 학습이 수행된다. 이때 심층강화학습 에이전트는 강화학습 환경과의 지속적인 상호 작용을 통해 변형

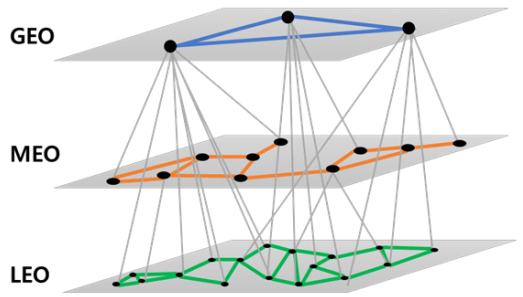

Fig. 5. 3-Layer Integrated Satellite Network Graph

패턴을 학습하여 최적의 액세스 전략을 얻을 수 있다. 어떤 사용자와 연결하는 것이 이득인지를 계산한 보상을 축적하여, 각각의 상황에서의 최대 이득인 즉각적 보상을 최대화하는 것이 아닌, 전체 상황에서의 누적 보상을 최대화하는 방향으로 설계하기 때문에, 빈번한 Handover를 피하고 Long-term throughput을 최대화할 수 있다.

그림 5에서 나타낸 것과 같은 GEO, MEO, LEO의 3계층 통합 위성 네트워크에서의 Capacity management 문제는 복잡한 구조로 인해 네트워크 연산 복잡도가 매우 커지게 된다[31]. 따라서 네트워크 모델을 구성하고, 위성 간의 용량을 계산하기 위해 상대적으로 낮은 복잡도의 모델을 제시하였다. 시간 확장 그래프를 이용하여, 복잡도가 낮은 용량 계산 알고리즘과 시간 구조 기반의 Augmenting path searching을 적용하여, 기존의 방법에 비해 검색 공간을 크게 줄여 계산 복잡도를 낮추었다. 그런 다음, Q-Learning을 기반으로 최적의 누적 Capacity 관리 알고리즘을 제안하였다. Q-Learning의 경우 중앙 집중형 학습을 진행하여 사용자가 많은 시나리오에서 상태 공간의 크기가 커질 수 있는데, 이때 Deep Q-Learning을 적용하면 성능 향상에 도움이 될 수 있다.

4.4 Others

심층강화학습은 위성 네트워크에서의 결함 진단에서 사용되기도 한다. SAGS(Space-Air-Ground- Sea) 네트워크에서, 개별 네트워크에 장애가 있는 경우 전체 네트워크의 QoS가 감소할 수 있다[32]. 이에, Beyond-5G 기술을 적용하여 다중 접속 엣지 컴퓨팅 및 클라우드 서비스를 활용해 결함이 있는 드론을 진단하는 알고리즘을 활용할 수 있다. 이 알고리즘은 Cubature Kalman filter를 기반으로 한 Radial bias function neural network를 활용하여, 드론이 제공하는 데이터를 기반으로 고장 감지를 수행한다. 이 때

데이터 수집 경로는 다중 접속 엣지 컴퓨팅 및 클라우드 서버에서, 심층 강화학습 알고리즘인 DDPG(Deep Deterministic Policy Gradient)를 이용하여 에너지 효율을 증가시킨다. 이를 통해 기존의 오류 감지 알고리즘에 비해 정밀도 증가와 에너지 소비 감소 효과를 얻을 수 있다.

앞서 언급한 것과 같이, 저궤도 위성 네트워크와 HAPS 등의 NTN을 연결하는 것은 현재 차세대 네트워크 연구에서 활발하게 이루어지고 있다. 통신 효율성을 극대화하려면 궤도를 도는 위성과 UAV의 궤적과의 연결을 최적화해야 하는데, 네트워크 토폴로지가 시간의 흐름에 따라 빠르게 변하고, 행동 공간의 크기가 매우 크기 때문에 최적화 문제를 풀기 어렵다. 이 어려움을 극복하기 위해, 행동 차원 감소 기술을 결합한 MADRL 방법으로 문제를 해결하였다 [33]. CTDE 방식을 통해 에이전트 간 정보가 공유된 상태에서 학습을 진행하고, 실행할 때는 분산형 방식을 적용하여 처리량과 에너지 효율 면에서 우수한 성능을 보였다.

V. 결론 및 향후 발전방향

본 논문에서는 고도 2000km 이하에서 서비스하는 저궤도 인공위성에서 심층강화학습을 적용한 동향 및 기술 이슈에 관해 기술하였다. 이 중 Scheduling, Resource allocation, User access control, Fault diagnosis 문제에서 각각 심층강화학습을 적용하게 되면 기존의 방법보다 우수한 성과를 낼 수 있다는 것을 확인하였다. 이 외에도 저궤도 위성통신에 있어 다른 문제에도 심층강화학습 기술을 활용하면 효과를 볼 수 있을 것으로 기대한다.

심층강화학습이 인공위성에 적용될 때 가장 크게 고려되어야 할 점은 모델을 훈련시킬 때의 계산 용량이다. 위성에서는 전력 및 계산 용량이 제한적이기 때문에, 훈련은 지상 기지국처럼 계산 용량이 더 큰 Terminal에서 진행하고, 이를 이용해 위성의 자원 할당 및 분배, 고장 진단 등 다양한 분야로 사용하는 연구가 활발히 이루어지고 있다. 특히 다수의 위성이 한 궤도, 또는 여러 궤도에 분포하는 저궤도 위성의 경우 멀티에이전트 심층강화학습을 적극적으로 활용할 수 있을 것으로 보인다. 통신량을 줄이기 위해 Distributed learning 등을 함께 활용하면, 위성통신을 비롯하여 많은 분야에 심층강화학습을 적용하는 시도가 늘어날 것으로 전망한다.

References

[1] 조준우, 김태윤, 김재현, “6G 초공간 : 위성통신의 현재와 미래,” *The Mag. IEIE*, vol. 47, no. 5, pp. 34-45, May 2020.

[2] S. J. Shin, C. L. Cho, H. S. Jeon, S. H. Yoon, and T. Y. Kim, “A survey on deep reinforcement learning libraries,” *Electron. and Telecommun. Trends*, vol. 34, no. 6, pp. 87-99, Dec. 2019. (https://doi.org/10.22648/ETRI.2019.J.340608)

[3] M. Shin, J. Kim, and M. Levorato, “Auction-based charging scheduling with deep learning framework for multi-drone networks,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4235-4248, May 2019. (https://doi.org/10.1109/TVT.2019.2903144.)

[4] M. Lemaître, G. Verfaillie, F. Jouhaud, J.-M. Lachiver, and N. Bataille, “Selecting and scheduling observations of agile satellites,” *Aerospace Sci. and Technol.*, vol. 6, no. 5, pp. 367-381, Sep. 2002. (https://doi.org/https://doi.org/10.1016/S1270-9638(02)01173-2)

[5] D. Y. Liao and Y. T. Yang, “Imaging order scheduling of an earth observation satellite,” *IEEE Trans. Syst., Man and Cybernetics, Part C (Appl. and Rev.)*, vol. 37, no. 5, pp. 794-802, Aug. 2007. (https://doi.org/10.1109/TSMCC.2007.900668)

[6] Y. Q. Li, M. Q. Xu, and R. Wang, “Scheduling observations of agile satellites with combined genetic algorithm,” in *Proc. Int. Conf. Natural Computation*, vol. 3, pp. 29-33, Haikou, Hainan, China, Nov. 2007. (https://doi.org/10.1109/ICNC.2007.652)

[7] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, “A survey of machine and deep learning methods for internet of things (IoT) security,” *IEEE Commun. Surv. & Tuts.*, vol. 22, no. 3, pp. 1646-1685, Thirdquarter 2020.

[8] C. Zhou, W. Wu, H. He, P. Yang, F. Lyu, N. Cheng, and X. Shen, “Deep reinforcement learning for delay-oriented IoT task scheduling

in SAGIN," *IEEE Trans. Wireless Commun.*, Oct. 2020. (https://doi.org/10.1109/TWC.2020.3029143)

[9] S. Salcedo-Sanz, R. Santiago-Mozos, and C. Bousono-Calzon, "A hybrid Hopfield network-simulated annealing approach for frequency assignment in satellite communications systems," *IEEE Trans. Syst., Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 2, pp. 1108-1116, Apr. 2004. (https://doi.org/10.1109/tsmcb.2003.821458)

[10] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science,* vol. 220, no. 4598, pp. 671-680, May 1983. (https://doi.org/10.1126/science.220.4598.671)

[11] X. Hu, Y. Zhang, X. Liao, Z. Liu, W. Wang, and F. M. Ghannouchi, "Dynamic beam hopping method based on multi-objective deep reinforcement learning for next generation satellite broadband systems," *IEEE Trans. Broadcasting,* vol. 66, no. 3, pp. 630-646, Sep. 2020. (https://doi.org/10.1109/TBC.2019.2960940)

[12] Y. Wu, G. Hu, F. Jin, and J. Zu, "A satellite handover strategy based on the potential game in LEO satellite networks," *IEEE Access*, vol. 7, no. 133, pp. 641-652, Sep. 2019. (https://doi.org/10.1109/ACCESS.2019.2941217)

[13] R. Musumpuka, T. M. Walingo, and J. M. Smith, "Performance analysis of correlated handover service in LEO mobile satellite systems," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2213-2216, Nov. 2016. (https://doi.org/10.1109/LCOMM.2016.2604311)

[14] E. Papapetrou and F. Pavlidou, "Analytic study of doppler-based handover management in LEO satellite systems," *IEEE Trans. Aerospace and Electronic Syst.,* vol. 41, no. 3, pp. 830-839, Jul. 2005. (https://doi.org/10.1109/TAES.2005.1541433)

[15] L. Bertaux, S. Medjiah, P. Berthou, S. Abdellatif, A. Hakiri, P. Gelard, F. Planchou, and M. Bruyere, "Software defined networking and virtualization for broadband satellite networks," *IEEE Commun. Mag.*, vol. 53, pp. 54-60, Mar. 2015. (https://doi.org/10.1109/MCOM.2015.7060482)

[16] J. Ding, et al., "Service provider competition and cooperation in cloud-based software defined wireless networks," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 134-140, Nov. 2015. (https://doi.org/10.1109/MCOM.2015.7321982)

[17] M. Sheng, et al., "Toward a flexible and reconfigurable broadband satellite network: Resource management architecture and strategies," *IEEE Wireless Commun. Mag.*, vol. 24, no. 4, pp. 127-133, Aug. 2017. (https://doi.org/10.1109/MWC.2017.1600173)

[18] J. Kim, "심층 강화 학습 기술 동향," *Broadcasting and Media Mag.*, vol. 27, no. 2, pp. 26-34, Apr. 2022.

[19] V. Mnih, et al., "Playing atari with deep reinforcement learning," *arXiv preprint, arXiv: 1312.5602*, CoRR, 2013. (https://doi.org/10.48550/arXiv.1312.5602)

[20] W. J. Yun, et al., "Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control," *IEEE Trans. Ind. Informat.,* vol. 18, no. 10, pp. 7086-7096, Oct. 2022. (https://doi.org/10.1109/TII.2022.3143175)

[21] W. J. Yun, et al., "Hierarchical deep reinforcement learning-based propofol infusion assistant framework in anesthesia," *IEEE Trans. Neural Netw. and Learn. Syst.(Early Access)*, pp. 1-12, Jul. 2022. (https://doi.org/10.1109/TNNLS.2022.3190379)

[22] Z. Jia, M. Sheng, J. Li, D. Zhou, and Z. Han, "Joint HAP access and LEO satellite backhaul in 6G: Matching game-based approaches," *IEEE J. Sel. Areas in Commun.,* vol. 39, no. 4, pp. 1147-1159, Apr. 2021. (https://doi.org/10.1109/JSAC.2020.3018824)

[23] L. Dalin, W. Haijiao, Y. Zhen, G. Yanfeng, and S. Shi, "An online distributed satellite cooperative observation scheduling algorithm based on multiagent deep reinforcement learning," *IEEE Geosci. and Remote Sensing*

Lett., vol. 18, no. 11, pp. 1901-1905, Nov. 2021. (https://doi.org/10.1109/LGRS.2020.3009823)

[24] M. Chen, Y. Chen, Y. Chen, and W. Qi, "Deep reinforcement learning for agile satellite scheduling problem," in *Proc. IEEE SSCI*, pp. 126-132, Xiamen, China, Dec. 2019. (https://doi.org/10.1109/SSCI44817.2019.9002957)

[25] L. Ren, X. Ning, and J. Li, "Hierarchical reinforcement-learning for real-time scheduling of agile satellites," *IEEE Access*, vol. 8, pp. 220523-220532, Nov. 2020. (https://doi.org/10.1109/ACCESS.2020.3040748)

[26] X. Hu, S. Liu, R. Chen, W. Wang, and C. Wang, "A deep reinforcement learning-based framework for dynamic resource allocation in multibeam satellite systems," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1612-1615, Aug. 2018. (https://doi.org/10.1109/LCOMM.2018.2844243)

[27] B. Zhao, J. Liu, Z. Wei, and I. You, "A deep reinforcement learning based approach for energy-efficient channel allocation in satellite internet of things," *IEEE Access*, vol. 8, pp. 62197-62206, Mar. 2020. (https://doi.org/10.1109/ACCESS.2020.2983437)

[28] S. Liu, X. Hu, and W. Wang, "Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems," *IEEE Access*, vol. 6, pp. 15733-15742, Feb. 2018. (https://doi.org/10.1109/ACCESS.2018.2809581)

[29] B. Deng, C. Jiang, H. Yao, S. Guo, and S. Zhao, "The next generation heterogeneous satellite communication networks: Integration of resource management and deep reinforcement learning," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 105-111, Apr. 2020. (https://doi.org/10.1109/MWC.001.1900178)

[30] Y. Cao, S. -Y. Lien, and Y. -C. Liang, "Deep reinforcement learning for multi-user access control in non-terrestrial networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1605-1619, Mar. 2021. (https://doi.org/10.1109/TCOMM.2020.3041347)

[31] C. Jiang and X. Zhu, "Reinforcement learning based capacity management in multi-layer satellite networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4685-4699, Jul. 2020. (https://doi.org/10.1109/TWC.2020.2986114)

[32] X. Wang, H. Lin, H. Zhang, D. Miao, Q. Miao, and W. Liu, "Intelligent drone-assisted fault diagnosis for B5G-Enabled space-air-ground-space networks," *IEEE Trans. Netw. Sci. and Eng.*, vol. 8, no. 4, pp. 2849-2860, Dec. 2021. (https://doi.org/10.1109/TNSE.2020.3043624)

[33] J. -H. Lee, J. Park, M. Bennis, and Y. -C. Ko, "Integrating LEO satellites and Multi-UAV reinforcement learning for hybrid FSO/RF non-terrestrial networks," *IEEE Trans. Veh. Technol.(Early Access)*, 2022. (https://doi.org/10.1109/TVT.2022.3220696)

이 현 수 (Hyunsoo Lee)

2021년 2월 : 숭실대학교 전자정보공학부 졸업

2021년 3월~현재 : 고려대학교 전기전자공학과 석박통합과정

<관심분야> 전자공학, 통신공학, 모빌리티, 강화학습

[ORCID:0000-0003-1113-9019]

김 중 헌 (Joongheon Kim)

2004년 2월 : 고려대학교 컴퓨터학과 졸업

2006년 3월 : 고려대학교 컴퓨터학과 석사

2014년 8월 : University of Southern California Computer Science 박사 졸업

2016년 3월 : 중앙대학교 소프트웨어대학 조교수

2019년 9월~현재 : 고려대학교 전기전자공학부 부교수

<관심분야> Stochastic Optimization, Mobility, Reinforcement Learning, Quantum Deep Learning

[ORCID:0000-0003-2126-768X]