JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 3, pp. 655-665, Mar. 2025

Check for updates

RL-PoCS: 메타버스 블록체인을 위한 강화학습 기반 동적 검증자 노드 선정 알고리즘

현 기 정*
서강대학교 메타버스전문대학원 박사과정

Reinforcement Learning-Based Proof of Contribution Score: A Reinforcement Learning-Based Dynamic Validator Node Selection Algorithm for Metaverse Blockchain

Ki-Jeong Hyun*

Ph.D. Student, Graduate School of Metaverse, Sogang University, Seoul 04107, Korea

[요 약]

본 논문은 메타버스 블록체인 환경에서의 검증자 노드 선정을 위한 강화학습 기반 동적 알고리즘 Reinforcement Learning based Proof of Contribution Score(RL-PoCS)을 제안한다. Double Deep Q-Network(DDQN)의 강화학습 기법을 기여점수증명(Proof of Contribution Score, PoCS)에 결합한 본 알고리즘은 블록체인 검증자 노드의 기여도를 정량화하고, 이를 기반으로 동적으로 검증자를 선정하여 보상을 분배하는 알고리즘이다. 특히, 메타버스 환경처럼 동적이고 탈 중앙화 된 환경에서의 자원 분배 문제를 해결하기 위해 합리지수를 활용하여 검증자의 기여점수와 실제 보상 간의 공정성을 측정하고, 그 편차를 학습 과정에 반영함으로써, 보다 공정하고 효율적인 검증자 선정을 가능하게 한다. 이를 위해 지속적으로 학습하며 네트워크의 상태 변화에 따라 최적의 검증자를 선정하는 과정을 개선한다. 시뮬레이션 결과, RL-PoCS 알고리즘은 메타버스 블록체인 네트워크에서 공정한 보상 분배를 실현함과 동시에, 확장성과 효율성을 유지하며 탈 중앙화 된 검증 구조를 최적화하는 데 효과적임을 보여준다.

[Abstract]

This paper proposes a reinforcement learning-based dynamic algorithm, namely Reinforcement Learning-based Proof of Contribution Score (RL-PoCS), for validator node selection in metaverse blockchain environments. By integrating the reinforcement learning technique of Double Deep Q-Network (DDQN) with the Proof of Contribution Score (PoCS), this algorithm quantifies the contribution of blockchain validator nodes and dynamically selects validators based on this score to distribute rewards. Specifically, to address the resource allocation problem in dynamic and decentralized environments such as the metaverse, the algorithm employs a rationality index to measure the fairness between the contribution score and the actual rewards of validators. The deviation is incorporated into the learning process, enabling fairer and more efficient validator selection. The algorithm continuously learns and adapts to network state changes, improving the process of selecting optimal validators. Simulation results demonstrate that the RL-PoCS algorithm effectively achieves fair reward distribution, maintains scalability and efficiency, and optimizes decentralized validation structures within metaverse blockchain networks.

색인어 : 메타버스, 강화학습, DDQN, 블록체인, 기여점수증명

Keyword : Metaverse, Reinforcement Learning, Double Deep Q-Network, Blockchain, Proof of Contribution Score

http://dx.doi.org/10.9728/dcs.2025.26.3.655

Received 16 February 2025; **Revised** 10 March 2025
Accepted 19 March 2025

***Corresponding Author; Ki-Jeong Hyun**

Tel: +82-70-4235-1495
E-mail: vennyhyun@gmail.com

Ⅰ. 서 론

메타버스는 가상현실, 증강현실, 블록체인, 인공지능 등 첨단 기술이 융합된 디지털 공간으로, 현실 세계와 유사한 복잡한 경제 구조와 사회적 상호작용을 포함하는 세상이다[1]. 메타버스라는 용어는 1992년 닐 스티븐슨의 사이버펑크 소설 Snow Crash에서 처음 등장했다[2]. 이 소설에서는 가상 공간과 현실이 혼합된 디지털 세계를 그렸으며, 현재 우리가 사용하는 메타버스 개념의 시초가 되었다. 메타버스는 물리적 세계와 유사한 형태로 사용자가 3D 디지털 공간에서 아바타를 통해 서로 상호작용할 수 있는 환경을 제공하며, 실제 현실을 복제한 형태로 발전해 왔다. 홀렌센, 코틀러, 오프레시닉 또한 메타버스를 “우리가 물리적 세계에서 일하는 방식을 디지털로 모방한 것”으로 정의하며, 사용자들이 물리적 세상과 비슷한 방식으로 가상 공간에서 사회적, 경제적 활동을 할 수 있다고 설명한다[3]. 이러한 메타버스는 가상 경제 구조와 사용자 활동이 매우 복잡하며, 블록체인 기술은 그 안에서 투명성과 보안을 보장하는 필수적인 역할을 수행한다[1],[4],[5]. 특히, 메타버스 내에서 사용자는 아바타로서 경제 활동에 참여하고, 그 기여도에 따라 보상을 받을 수 있는 메커니즘이 필수적 요소이다. 또한, 메타버스에서 블록체인은 사용자들이 디지털 자산에 대한 진정한 소유권을 갖고, 투명하고 공정한 시스템 내에서 상호작용하며, 다양한 가상 세계를 자유롭게 넘나들 수 있도록 하는 핵심적인 기반 기술이다. 중앙화된 플랫폼의 한계를 넘어 사용자 중심의 개방적이고 안전한 메타버스 생태계를 구축하는 데 필수적인 역할을 한다.

블록체인 네트워크에서 검증자 노드는 네트워크의 안전성과 신뢰성을 유지하는 핵심 요소이다. 기존의 Proof of Work(PoW)는 높은 연산 비용과 에너지 소비 문제로 인해 메타버스의 실시간 트랜잭션을 처리하는 데 비효율적이며, 네트워크 확장성이 낮아 대규모 사용자 환경을 원활하게 지원할 수 없다. Proof of Stake(PoS)는 자본을 기반으로 네트워크 참여 권한을 부여하는 구조이므로, 기여도가 아닌 보유 자산에 따라 네트워크 영향력이 결정되어 메타버스 내에서 공정한 기여 평가를 보장하기 어렵다[6]. 특히, 이러한 환경에서는 사용자들의 기여도를 기반으로 한 공정한 보상 분배가 중요하여, 이를 반영할 수 있는 노드 선정 알고리즘이 필요하다[7]. 이러한 맥락에서 RL-PoCS는 메타버스 환경에서 빠르게 변화하는 조건에 적응할 수 있도록 설계되었고, 기여도 기반의 보상 체계를 통해 검증자들에게 인센티브를 제공하는 방식으로 진행한다. 이를 통해, 메타버스 블록체인의 탈 중앙화를 촉진시킬 수 있다

강화학습(Reinforcement Learning, RL)은 환경과의 상호작용을 통해 에이전트가 최적의 행동 전략을 학습하는 방법으로, 블록체인의 검증자 노드 선정에 효과적으로 적용될 수 있다[8]. 특히, DDQN은 심층 신경망을 사용하여 복잡한 상태 공간에서 최적의 정책을 학습할 수 있는 강화학습 알고리즘으로, 네트워크의 상태 변화에 따라 실시간으로 적응할 수 있는 동적 검증자 노드 선정을 가능하게 한다[9],[10].

따라서, 본 논문에서는 기여도를 평가하여 검증자 노드를 선정하는 PoCS를 합의 알고리즘으로 채택하고, 이를 강화학습 기반으로 확장시킨 RL-PoCS 알고리즘을 제안한다.

이 논문의 기여점은 다음과 같다:

- **RL-PoCS 알고리즘 제안:** 기존의 블록체인 검증자 선정 방법에 비해 강화학습과 새로운 PoCS 알고리즘을 도입함으로써, 블록체인 검증자 노드를 동적으로 선정하여 보다 공정한 보상 방식으로 개선한다.
- **강화학습을 통한 동적 환경 적응:** 메타버스의 동적이고 탈중앙화 된 환경에서 발생하는 빠른 상태 변화에 대응하기 위해 DDQN 기반 강화학습을 적용하여, 검증자 선정 과정에서 네트워크 기여 활동에 맞춰 학습하고 적응하는 알고리즘을 개발한다.
- **시뮬레이션을 통한 실효성 검증:** 시뮬레이션 실험을 통해 제안된 RL-PoCS 알고리즘이 검증자 선정의 공정성과 효율성에서 우수한 성능을 보였으며, 메타버스 블록체인 네트워크에서 실용적으로 적용될 수 있음을 보여준다.

Ⅱ. 배경지식

메타버스는 디지털 공간에서 사용자가 다양한 경제 활동을 수행하는 가상 생태계로, 이를 뒷받침하는 블록체인 기술은 메타버스의 경제 구조를 더욱 강화한다[4],[10]. 메타버스 내에서 사용자들은 사회적 상호작용, 경제 활동, 협력 활동 등을 통해 기여도를 쌓고 가상 화폐나 토큰으로 보상받는 경제적 선순환 구조를 만들 수 있다. 이러한 기여점수를 관리하고 투명하게 기록하는 데 필요한 요소가 블록체인 검증자 노드이다[11],[12].

메타버스 블록체인의 검증자 노드는 메타버스에서 발생하는 수많은 트랜잭션의 유효성을 검증하고, 블록 제안 및 확인 과정을 통해 네트워크의 분산 합의를 이끌어내는 기능을 담당한다. 이를 신뢰할 수 있는 형태로 블록체인 네트워크에 기록함으로써 메타버스 경제의 신뢰성을 높인다. 이를 통해 네트워크의 보안성을 강화하고, 블록체인의 핵심 원칙인 탈중앙화를 유지한다[11],[12]. 특히, PoS와 DPoS와 같은 합의 메커니즘에서 검증자는 네트워크의 안정성과 신뢰성을 보장하는 필수적인 구성 요소로 작용한다. 만약 검증자 노드가 없다면, 블록체인은 중앙화 된 시스템이 되어 탈 중앙화라는 본래의 가치를 잃을 수 있다. 검증자 노드를 선정함에 있어, PoCS는 블록체인 네트워크에서 노드의 기여도를 공정하게 평가하고 합의 과정을 개선하기 위한 합의 알고리즘이다. PoCS는 단순히 노드의 지분만을 평가하는 기존의 PoS 방식과 달리, 블록 생성 빈도, 네트워크 활동, 장기 참여도 등의 다

양한 기여 요소를 고려한다. 이를 통해 PoCS는 노드의 실제 기여도를 반영한 공정한 보상과 검증 기회를 제공하며, 중앙화 문제를 완화하고 네트워크의 지속 가능성을 높인다. 특히 PoCS는 Practical Byzantine Fault Tolerance(PBFT)와 결합하여 허가형 블록체인에서의 안정성과 신뢰성을 강화할 수 있으며, 메타버스의 복잡한 환경에서도 효율적으로 적용될 수 있다[13].

강화학습은 에이전트가 환경과 상호작용하면서 결정을 내리는 학습 방식으로, 보상을 통해 피드백을 받고 장기적으로 누적 보상을 극대화하기 위해 행동을 최적화하는 계산적 접근법이다. 강화학습의 주요 요소에는 에이전트, 상태, 행동, 정책, 보상 및 가치 함수 등이 있으며, 이들은 강화학습의 핵심 개념을 구성한다[14]. 그림 1은 에이전트-환경 상호작용 과정을 보여준다.

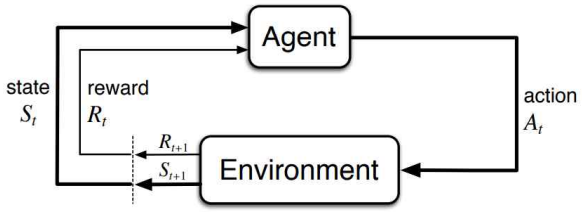

그림 1. 에이전트-환경 간 상호작용[15]
Fig. 1. Agent-environment interaction[15]

에이전트는 환경으로부터 보상 또는 처벌의 형태로 피드백을 받으며, 이러한 피드백을 사용해 자신의 행동을 조정하고 성능을 개선해 나간다[9],[16]. DQN은 Q-Learning을 딥러닝과 결합하여 고차원 상태 공간에서도 최적의 행동을 학습할 수 있도록 만든 알고리즘이다[17],[18]. 하지만, Q 값의 추정에서 과대평가 편향(Overestimation Bias) 문제가 있어, 이를 해결하기 위해 Double Q-learning의 개념을 DQN에 적용한 것이 Double DQN 알고리즘이다[18],[19]. 이 DDQN과 PoCS를 조합한 것이 RL-PoCS 알고리즘으로, 블록체인 검증자 선정 방식은 기존의 PoS와 차별되고, PoS 방식에서는 검증자가 보유한 자산의 양에 따라 검증자로 선정되고 보상[6]을 받는 반면, RL-PoCS는 메타버스 내 다양한 활동에 따른 기여도를 바탕으로 검증자를 선정한다. 이를 통해 단순히 지분 보유 양 보다는 메타버스 이용자의 활동과 네트워크에 대한 기여도를 포함시킨 알고리즘 적용으로 보다 공정한 보상이 가능하다. 메타버스의 복잡한 환경에서 활동에 따라 적응적으로 검증자를 선정함으로써, 기여도에 따른 인센티브 구조가 더욱 강화되고, 메타버스 내 경제 활동이 원활하게 이루어질 수 있도록 촉진한다.

RL-PoCS 알고리즘은 강화 학습을 활용하여 블록체인 환경에서 검증자 노드의 선정을 최적화하고, 보상 분배의 공정성과 효율성을 향상시키기 위해 설계되었다. 이 알고리즘의 동적이고 적응적인 학습 방식은 지속적인 상호작용, 빠른 변화, 분산된 거버넌스가 특징인 메타버스 블록체인 생태계에 적합하다. 메타버스의 동적이고 분산된 생태계는 사용자, 서비스 제공자, 콘텐츠 크리에이터 등 다양한 이해관계자가 지속적으로 상호작용하는 가상 세계로, 데이터 저장, 공유, 대량의 거래와 데이터 교환이 발생한다[10]. 이러한 상호작용을 처리하기 위해서는 활동을 검증하고 네트워크의 무결성을 유지할 수 있는 공정하고 확장 가능한 합의 메커니즘이 필요하다. 메타버스에서는 수요, 사용자 참여, 디지털 자산 소유권 등이 빠르게 변화하는 동적인 환경이 특징적이다. 이러한 환경에 대비하기 위해 RL-PoCS 알고리즘은 강화 학습을 기반으로 설계되어 변화에 적응력이 뛰어나다. RL-PoCS 알고리즘은 Reasonable Percentage(RP) 편차를 사용하여 공정성을 확보하며, 기여도를 기반으로 검증자를 공정하게 선정함으로써 권력의 중앙집중화를 방지하고, 과소 대표된 노드에게도 검증 역할이 돌아가게 되어, 공평한 보상을 획득한다. 이러한 RL-PoCS 알고리즘은 탈중앙화, 기여도 중심의 인센티브, 공정성의 원칙을 강화함으로써, 메타버스에서 사용자가 기여를 통해 토큰이나 자산을 획득하는 블록체인 기반 경제는 검증자가 자신의 기여도에 비례하여 보상을 받는 RL-PoCS로부터 혜택을 받을 수 있다.

요약하자면, RL-PoCS는 메타버스의 동적이고 탈 중앙화된 환경에 적합하게 설계된 공정하고 적응적이며 확장 가능한 검증자 선정 및 보상 분배 방식을 제공한다. 이 알고리즘은 기여도에 기반하여 자원을 공정하게 할당하고, 탈중앙화를 촉진하며, 진화하는 생태계에서 지속적으로 학습할 수 있다.

Ⅲ. 관련 연구

블록체인 합의 알고리즘에 대한 기존 연구는 주로 PoW, PoS, 그리고 Delegated Proof of Stake(DPoS)와 같은 고전적인 방법에 초점을 맞추어 왔다. PoW는 계산 복잡도가 높아 에너지 소모가 크고, PoS는 지분에 따른 보상 불균형이 발생할 수 있다. 이러한 문제를 해결하기 위한 연구로, PoCS가 제안되었으며, 이는 노드의 기여도를 기반으로 공정한 검증자 선정을 목표로 한다. 최근에는 강화학습을 이용하여 블록체인 시스템에서의 효율성을 개선하려는 시도들이 등장하고 있으며, 강화학습 알고리즘을 활용한 연구들이 진행되고 있다. 그러나 기존 연구들은 메타버스처럼 동적 환경에서의 검증자 노드 선정에 관한 주제가 부족한 상황이다.

PoCS는 사용자의 다양한 기여도를 반영한 검증자 선정 메커니즘으로, 메타버스 내에서의 활동, 사회적 기여, 경제적 기여 등을 종합하여 검증자를 선정한다. 기존 PoS의 자산 기반 검증자 선정 방식에서 더 나아가, 복잡한 기여 요인을 반영하는 PoCS는 메타버스 이용자의 기여도 측정과 보상에 적합한 방식이다. RL-PoCS는 이 알고리즘을 강화 학습 방식으로 재구현한 알고리즘이다[13]. 하지만, PoCS 방식은 노드의 기여도에 따라 보상을 분배하는 공정한 시스템을 목표로 하지

만, 고정된 규칙과 비동적인 보상 분배 메커니즘은 네트워크의 변화에 실시간으로 제대로 적응하지 못하는 한계를 가진다. PoCS 합의 알고리즘에서는 노드의 기여도를 평가하기 위해 기여 점수를 사용한다. 노드 i 의 기여 점수는 다음과 같이 계산된다. 수식 (1)은 기여점수를 구하는 수식이다.

$$CS_i = \sum_{k=1}^{n} W_k S_k \quad (1)$$

수식 (1)에서, CS_i는 기여 점수의 합, S_k는 기여 요소의 활동 점수, W_k 는 가중치, n은 전체 기여 요소의 수이다. 무엇보다, 현재 라운드에서 검증자로 선택된 노드는 수식 (2)처럼 기여 점수에서 차감된다.

$$CS_i = \max(0, CS_i - \gamma \times CS_i) \quad (2)$$

수식 (2)에서, $\gamma = \frac{|S|}{|N|}$ 이다. 반면에, 선택되지 않은 노드는 기여 점수가 증가한다.

$$CS_i = CS_i + CSD_i \quad (3)$$

수식 (3)에서 CSD_i 는 기여점수의 델타 값이다. 이러한 방식을 통해, PoCS는 검증자 선출 및 보상 분배에 여러 기여 지표를 반영하여 PoS보다 공정한 대안을 제공한다[13].

MRL-PoS는 다중 에이전트 강화 학습을 활용하여 네트워크의 보안과 신뢰성을 강화하고, 악의적인 행위를 효과적으로 대처할 수 있는 메커니즘이다. PoS 방식을 강화학습과 결합하여 검증자 노드 선출 과정을 최적화하고, 악의적인 노드 탐지와 평판 기반 투표를 통해 보안성과 신뢰성을 강화하려고 했고, 각 에이전트가 네트워크 참여 시 동적으로 생성된 평판 테이블을 바탕으로 리더를 선정하며, 이를 통해 악의적인 행위를 억제하고 검증자의 신뢰성을 유지하려는 목표를 달성하고자 하였다. MRL-PoS에서 검증자 노드를 선정하는 방식은 각 에이전트가 네트워크 참여 시 동적으로 생성된 평판 테이블(Reputation Table)을 기반으로 투표하여 리더를 선정하는 방식이다. 평판 테이블은 각 노드의 신뢰성과 성능을 평가할 수 있는 다섯 가지 지표로 구성되어 있다. 첫 번째 지표는 정확성(Accuracy)으로, 이전 블록 검증의 정확성을 평가하여 노드의 신뢰성을 측정한다. 두번째는 트랜잭션 보유 경향(Transaction Holding Tendency)으로, 트랜잭션을 불필요하게 보유하는 경향이 높은 노드는 악의적인 의도를 가질 가능성이 있어 이를 평가한다. 세번째는 처리 지연(Processing Delay)으로, 트랜잭션을 처리하는 데 걸리는 시간으로 노드의 효율성을 평가한다. 네번째는 과도한 처리 능력(Super Processing Power)으로, 과도한 컴퓨팅 파워를 가진 노드가 네트워크를 지배하려는 의도가 있는지 평가한다. 마지막으로 불법 트랜잭션 탐지 능력(Illegitimacy Detection Ability)은 불법 트랜잭션을 탐지하고 차단하는 능력을 평가하여 네트워크 보안에 기여한다. 이 평판 기반 투표 메커니즘을 통해 리더를 선정함으로써 악의적인 노드의 검증자 선출을 방지하고, 네트워크의 공정성과 신뢰성을 강화할 수 있다[21]. 하지만, 메타버스의 복잡한 환경에서 동적 적응성이 필요하지만, 평판 테이블의 지표는 정적 평가에 기반하고 있어, 빠르게 변화하는 네트워크 상태에 적응하는 데 한계가 있다. 또한 평판 기반 투표 방식은 일부 노드가 높은 평판을 유지할 경우, 소수의 검증자 노드에 집중되어 블록체인 공정성이 위배될 소지가 있다.

DQN을 기반으로 한 블록체인 합의 알고리즘을 제안하고 IoT(Internet of Things) 네트워크에 적용한 내용을 다루는 이 논문은 RAFT+(Reliable, Replicated, and Fault Tolerant)의 리더 선정 방식을 마코프 결정 과정(MDP)으로 모델링하고, DQN을 활용하여 최적의 리더를 선택하는 방법을 제안한다. 두 개의 신경망(추정 네트워크와 타겟 네트워크)을 이용해 블록 생성 지연을 최소화하는 리더를 선정하고, 네트워크의 상태와 자원 가용성에 따라 최적의 선택을 수행한다. 현재의 리더가 계속해서 성능을 유지하지 못하거나 하드웨어 상태가 악화되는 경우, 중앙 노드는 해당 리더를 교체할 수 있다. 리더의 성능을 지속적으로 모니터링하고, SNR(신호 대 잡음비)와 하드웨어 상태 등 리더의 성능 지표를 기반으로 교체 여부를 결정한다. 리더 선정 과정은 DQN을 이용하여 최적화되고, 중앙 노드는 각 팔로워의 상태 정보를 수집하여 최적의 리더를 학습하고 선택한다. 이 과정에서 경험 재현(Experience Replay)과 타겟 네트워크(Target Network)를 사용하여 학습이 안정적으로 이루어지며, 리더의 선택이 반복적인 평가와 학습을 통해 개선된다. 리더가 좋은 성능을 보일 경우 보상을 통해 그 위치를 유지하게 될 가능성이 높지만, 네트워크는 탐색(Exploration)과 활용(Exploitation)을 균형 있게 유지하여 과도하게 리더 자리를 독점하지 못하도록 방지한다[20].

표 1 처럼, 합의 알고리즘 비교를 통해 각 알고리즘이 가진 장단점과 특화된 적용 환경을 이해할 수 있다. RL-PoCS는 기여도 기반의 공정성을 보장하고, MRL-PoS는 PoS의 보안성을 강화하며, DQN 기반 합의 알고리즘은 자원 제약 환경에서 효율적으로 리더를 선정하는 데 적합하다.

Ⅳ. 동적 검증자 선정 알고리즘

RL-PoCS 알고리즘은 각 노드의 기여도에 따라 공정하게 검증자를 선정하고 보상을 분배하며 DDQN을 PoCS 메커니즘과 결합하여 강화 학습을 통해 네트워크의 효율성과 공정성을 최적화하는 것을 목표로 한다.

표 1. 합의 알고리즘 비교
Table 1. Comparison of consensus algorithms

Item	RL-PoCS	MRL-PoS	DQN-based Consensus Algorithm
Consensus Mechanism	Proof of Contribution Score (PoCS)	Proof of Stake (PoS)	DQN-based Leader Selection
Basic Algorithm	DDQN (Double Deep Q-Network)	Multi-Agent Reinforcement Learning (MARL)	Deep Q-Network (DQN)
Validator/ Leader Selection Criteria	Contribution Score of each node	Trustworthiness assessment based on reputation table	Signal-to-noise ratio (SNR), hardware state, system state
Reward Mechanism	Reward distribution proportional to contribution	Rewards or penalties based on reputation metrics	Rewards based on block generation delay time and consensus time
Fairness Maintenance	Fair selection of validators and reward distribution based on contribution	Leader selection based on reputation table and voting	Preventing leader monopolization through exploration-exploitation balance
Centralization Risk	Low – Contribution score is the main criterion	Reputation-based voting may result in some nodes being continuously selected as leaders	Risk of continued leader retention, but can be prevented through exploration
Dynamic Adaptation	Real-time evaluation of contributions for dynamic adaptation	Slow reputation update may make adaptation difficult	Leader change dynamically performed through state evaluation via DQN
Application Environment	High-dimensional environments like Metaverse	General blockchain based on PoS and secure environments	Resource-constrained environments like IoT networks
Algorithm Learning Method	DDQN used to solve the Q-value overestimation problem	Each node independently learns reputation metrics	Optimal leader selection via a single network
System Stability	High stability – validators are fairly selected based on contribution	Enhanced reliability through leader selection via reputation metrics	Minimizing block generation delay through optimal leader selection

RL-PoCS 의 필요성:

- **공정한 검증자 선정**: 메타버스의 복잡한 환경에서는 다수의 검증자 후보가 존재하며, 각 후보의 기여도는 시간에 따라 변한다. RL-PoCS는 강화학습을 통해 이러한 변화에 적응하며 검증자를 공정하게 선정하여 네트워크의 탈중앙성을 촉진한다.
- **보상 분배의 효율성**: 기존의 보상 분배 방식은 네트워크 상황 변화에 적응하지 못하는 한계가 있다. RL-PoCS는 각 노드의 기여도를 실시간으로 평가하고 보상을 분배함으로써, 네트워크의 참여를 장려하고 전체적인 효율성을 높인다.
- **동적 적응성**: 메타버스는 사용자의 참여와 활동이 빠르게 변화하는 동적인 환경이기 때문에, 이를 실시간으로 반영할 수 있는 적응적 알고리즘이 필요하다. RL-PoCS는 강화학습을 활용해 네트워크 상황에 따라 최적의 검증자 선정과 보상 분배 정책을 학습하여 실시간 적응성을 제공한다.

표 2. RL-PoCS 알고리즘
Table 2. RL-PoCS algorithms

Initialization: Initialize each node i with: Contribution score

$CS_i = CSD_i$, Coins $C_i = 0$, Selection count $SC_i = 0$ Set number of selected nodes: $f = (N-1)/3, S = 2f+1$ (according to PBFT rules)

Initialize Double DQN agent with state size N and action size N Initialize replay buffer and target network

for each selected node $r = 1$ to R **do**

1. Node Selection: Sort nodes by their contribution scores CS_i Select the top S nodes with the highest CS_i

While (Number of selected nodes < S):

Choose action a using ϵ-greedy policy:

$$a = \begin{cases} \text{random action}, & \text{with probability } \epsilon \\ \text{argmaxQ(s,a;}\theta\text{)}, & \text{otherwise} \end{cases}$$

Add the selected node to the list

2. Reward Calculation:

for each selected node j **do**

Update coins:

$$C_j \leftarrow C_j + \left(\frac{cs_j}{\sum_{k \in Selectednodes} cs_k}\right) \times R_{total}$$

Increment selection count: $SC_j \leftarrow SC_j + 1$

end for

3. RP Deviation Calculation:

foreach selected node i **do**

Calculate RP deviation: $RP_i = \left| \frac{s_i}{\sum_{j=1}^{S} s_j} - \frac{c_i}{\sum_{j=1}^{S} c_j} \right|$

end for

4. Adjusted Reward Calculation:

Calculate total RP deviation:

$$TotalRPDeviation = \sum_{j=1}^{S} RP_j$$

Adjust the total reward for fairness:

$$AdjustedReward = \frac{R_{total}}{1 + TotalRPDeviation}$$

5. Q-Learning Update:

for each selected node i **do**

Calculate the Q-value update using Double DQN:

$$Y_i^{DDQN} = R_{adjusted} + \gamma Q(s', \arg\max_{a'} Q(s', a'; \theta); \theta')$$

end for

6. Replay Memory Update: Store experience (state, action, reward, next state) in memory Sample a batch of experiences from memory

for each selected node $batch$ **do**

Update Q-values using Double DQN with the target network

end for

7. Target Network Update: Update target network every few steps by copying weights from the main network

end for

Final Output: Selection counts SC_i, Coins C_i, Reasonable percentage RP_i

알고리즘의 주요 과정:

- Initialization: 각 노드는 초기 기여 점수, 코인 수, 선정 횟수를 설정하며, 선정된 노드 수는 PBFT 규칙에 따라 정의한다. DDQN 에이전트도 초기화하여 상태 크기와 행동 크기를 설정한다.
- Node Selection: 기여 점수를 기준으로 상위 2f+1 개의 노드를 선정하며, ε-탐욕적 정책을 사용하여 무작위 또는 Q 값이 최대인 행동을 통해 노드를 추가 한다.
- Reward Calculation: 선정된 노드들에게 기여도에 비례한 코인을 보상으로 지급하며, 선정 횟수도 증가시킨다.
- RP Deviation Calculation: 각 노드의 기여도와 보상 간의 편차를 계산하여 공정성을 평가한다.
- Adjusted Reward Calculation: 전체 RP 편차를 반영하여 총 보상을 조정함으로써 네트워크의 공정성을 유지한다.
- Q-Learning Update: 선정된 노드들의 Q 값을 DDQN으로 업데이트하여 최적의 행동 정책을 학습한다.
- Replay Memory Update: 경험 재현 메모리에 상태, 행동, 보상, 다음 상태를 저장하고, 무작위로 샘플을 추출하여 학습에 사용한다.
- Target Network Update: 메인 네트워크의 가중치를 타겟 네트워크로 복사하여 업데이트하며, 학습의 안정성을 유지한다.
- Final Output: 각 노드의 선정 횟수, 획득한 코인, 합리적 비율을 출력하여 기여도와 보상 분배의 공정성을 평가한다.

표 2는 PoCS의 공정성과 효율성을 강화하고, 네트워크의 변화에 민첩하게 적응할 수 있는 동적 검증자 노드 선정 메커니즘으로서, 블록체인 기반의 메타버스 환경에서 핵심적인 기능을 할 수 있다.

V. 실험 결과 및 분석

실험 조건은 20개의 노드로 구성된 블록체인 네트워크를 가정하며, 5000 라운드 동안 수행한다. 각 라운드마다 PoCS 알고리즘을 DDQN 강화학습을 통해 검증자 노드를 선정하고, 기여 점수에 비례한 보상을 분배하며, 이를 통해 검증자

노드 선정의 공정성을 지속해서 평가한다. 각 노드의 기여 점수와 보상 간의 비율, 선정 횟수 등의 지표를 사용하여 알고리즘의 성능을 평가하며, ε-탐욕적 정책을 활용해 탐색과 활용 간의 균형을 조절한다.

이 실험을 통해 강화학습 기반의 검증자 노드 선정 및 보상 메커니즘이 기존의 방식보다 네트워크의 효율성과 공정성을 향상시킬 수 있음을 보여준다. 표 3은 본 연구에서 사용하는 하이퍼파라미터이다.

표 3. RL-PoCS 하이퍼파라미터
Table 3. RL-PoCS Hyperparameter

Parameter	Value
state_size	User-defined
action_size	User-defined
memory maxlen	1000
gamma (Discount Factor)	0.95
epsilon (Exploration Rate)	1
epsilon_min	0.01
epsilon_decay	0.995
learning_rate	0.001
hidden_layer_size	24
update_target_frequency	10
optimizer	Adam
loss function	MSELoss
tau (Soft Update Factor)	0.01

그림 2는 각 노드의 합리성 지수(Reasonableness Index)를 비교하고, 각 노드의 기여 점수와 획득한 코인 간의 비율을 기준으로 합리성 지수가 표시되며, 빨간 점선은 기준선(1)을 나타내고, 녹색 실선은 평균 합리성 지수를 나타낸다. 기여도에 비례하여 보상이 지급되고 있는지 평가하는 지표이다. 대부분의 노드가 기준선 1에 가깝게 위치하고 있어, 보상이 공정하게 분배되고 있음을 알 수 있다.

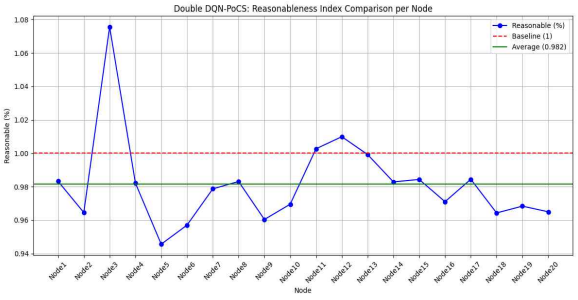

그림 2. RL-PoCS 합리 지수 비교
Fig. 2. RL-PoCS reasonableness index comparison

표 4는 RL-PoCS 의 실험 결과로, 각 노드의 기여 점수, 선정 횟수, 획득한 코인, 코인의 비율, 그리고 합리적 비율(RP)을 요약하여 보여준다.

표 4. RL-PoCS 실험 결과
Table 4. Experimental results

Node	Contribution Score (%)	Selection Count	Total Acquired Coins	Coin (%)	RP
Node1	6.28	3258	319.35	6.39	0.98
Node2	0.09	3211	4.67	0.09	0.96
Node3	17.59	3215	817.64	16.35	1.08
Node4	6.21	3262	316.14	6.32	0.98
Node5	1.98	3314	104.7	2.09	0.95
Node6	1.06	3265	55.38	1.11	0.96
Node7	4.21	3237	215.08	4.3	0.98
Node8	5.98	3258	304.14	6.08	0.98
Node9	1.83	3262	95.28	1.91	0.96
Node10	5.07	3283	261.47	5.23	0.97
Node11	8.52	3250	424.86	8.5	1
Node12	9.99	3255	494.6	9.89	1.01
Node13	8.69	3260	434.84	8.7	1
Node14	3.78	3223	192.3	3.85	0.98
Node15	4.73	3230	240.26	4.81	0.98
Node16	1.29	3209	66.43	1.33	0.97
Node17	4.07	3221	206.71	4.13	0.98
Node18	1.3	3240	67.41	1.35	0.96
Node19	5.81	3312	299.99	6	0.97
Node20	1.52	3235	78.76	1.58	0.96

그림 3의 그래프는 라운드별 평균 RP 편차를 보여준다. 초기 몇 라운드 동안 편차가 크며, 이후 점차적으로 감소하여 안정화되는 패턴을 보인다. 이는 강화학습을 통해 기여도와 보상 간의 불일치가 점차 감소함을 의미한다. 즉, 시간이 지남에 따라 에이전트가 학습하여 네트워크 내 공정한 보상 분배를 달성하고 있음을 보여준다. 초기에는 불안정성이 크지만, 학습이 진행될수록 안정적인 분배가 이루어지는 것을 확인할 수 있다.

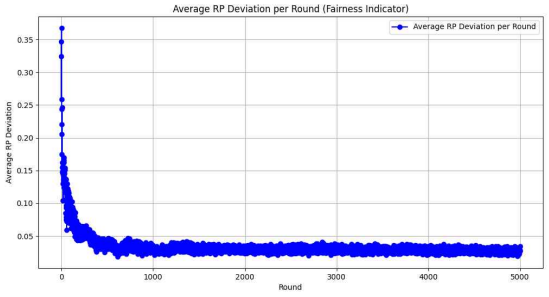

그림 3. 평균 RP 편차
Fig. 3. Average RP deviation

그림 4는 에피소드별 총 보상의 변화를 나타낸 그래프이다. 초기 에피소드에서는 총 보상이 낮으나, 시간이 지남에 따라 점차 증가하고, 이후에는 수렴하는 경향을 보인다. 에이전트가 초기에는 보상을 효율적으로 받지 못하다가, 학습이 진행됨에 따라 더 높은 보상을 받도록 정책을 개선하고 있음을 보여준다. 보상이 수렴한다는 것은 정책이 최적화되었고, 더

이상 큰 변화 없이 안정적으로 최대 보상을 달성하고 있음을 의미한다.

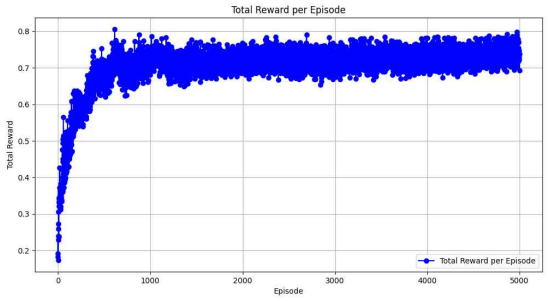

그림 4. 전체 보상
Fig. 4. Total reward

그림 5는 Q-Network 업데이트할 때 발생하는 손실 값의 변화를 나타내는 그래프로, 손실 값이 감소하는 것은 에이전트가 목표값과 실제 예측값 간의 차이를 줄이고 있다는 것을 의미하며, 이는 네트워크가 점점 더 나은 예측을 하고 있음을 나타낸다. 학습 초반에는 손실이 높지만 시간이 지남에 따라 감소하고 있다.

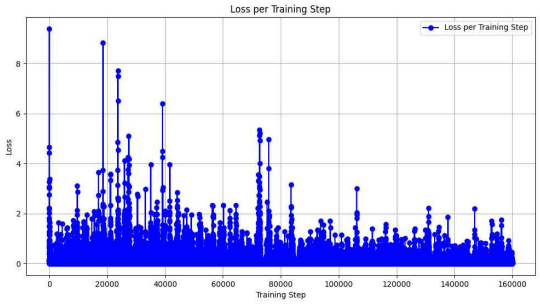

그림 5. 손실
Fig. 5. Loss

그림 6은 에피소드 수를 x축에, ε 값을 y축에 표시하여 탐색률의 변화를 시각화한 그래프로, ε-greedy 전략에서 탐색률 ε의 변화를 보여준다. 탐색률이 점차 감소하는 것을 확인함으로써 에이전트가 학습 후반부에는 탐색보다는 학습된 정책을 활용하여 행동하는 빈도가 증가하는 것을 알 수 있다. 적절한 탐색률 감소는 학습의 중요한 요소이다.

그림 7은 에피소드별 평균 Q 값을 보여주는 그래프이다. 초기에는 Q 값이 낮으며, 이후 학습이 진행됨에 따라 Q 값이 점차 증가하고 안정화된다. Q 값은 특정 상태에서 기대되는 총 보상을 나타내며, 평균 Q 값의 증가는 에이전트가 보다 나은 정책을 학습하고 있음을 의미한다. Q 값이 안정화된다는 것은 에이전트가 최적의 행동을 선택하는 정책을 거의 확립했음을 나타낸다.

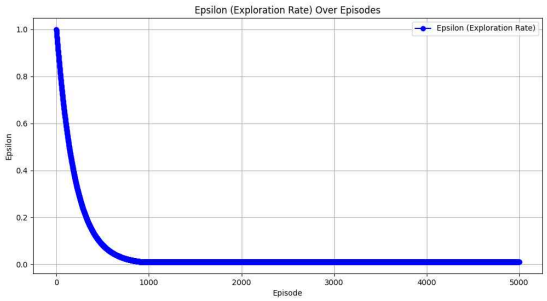

그림 6. 엡실론
Fig. 6. Epsilon

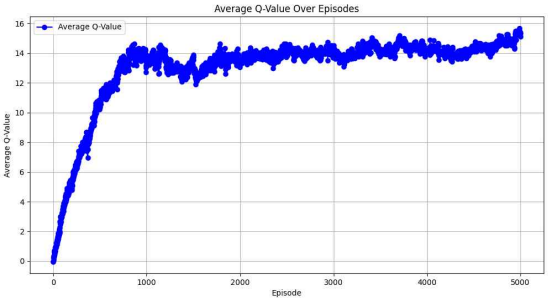

그림 7. 평균 Q-값
Fig. 7. Average Q-value

그림 8은 에피소드에서 각 노드가 선정된 비율을 나타내는 그래프이다. 초기에는 큰 변동이 있으나, 에피소드가 진행될수록 모든 노드의 선택 비율이 비슷한 수준으로 수렴하고 있다. 이는 에이전트가 학습을 통해 각 노드의 기여도를 고려하여 공정하게 선택하도록 정책을 조정한 결과이다. 초기에는 무작위로 선택되면서 비율의 변동이 크지만, 이후 기여도가 높은 노드를 선택하는 비율이 점차적으로 안정화되며, 전체적으로 네트워크의 공정성을 유지하고있다.

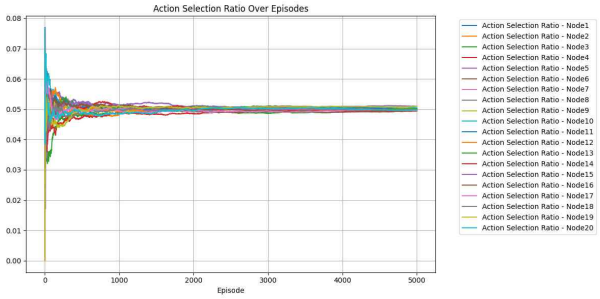

그림 8. 액션 선택 비율
Fig. 8. Action selection ratio

결론적으로, RL-PoCS 알고리즘은 검증자 노드 선정의 공정성을 달성하는 데 있어 유망한 결과를 보인다. 보상 분배의 일부 변동성은 여전히 남아 있지만, 강화 학습을 사용한 알고리즘은 지속적인 개선을 가능하게 하며, 시간이 지남에 따라

노드의 실제 기여도에 따른 보상이 더욱 정확해질 것으로 예상된다. 기여도 점수 계산의 세분화 및 추가적인 공정성 매개변수 도입과 같은 최적화를 통해 관찰된 편차를 줄이고, 보다 공정한 보상 분배를 보장할 수 있다. 이 분석은 RL-PoCS 접근 방식이 동적 블록체인 환경, 특히 메타버스 환경처럼 분산형 환경에서 검증자 선정에 강력하고 확장 가능하며 공정한 솔루션을 제공할 수 있음을 보여준다.

Ⅵ. 논의 및 향후 연구 방향

이 논문에서 제안하는 RL-PoCS 알고리즘은 메타버스 의 동적 환경에서 블록체인 검증자 노드 선정을 위한 인사이트를 제공한다. DDQN과 PoCS를 결합한 이 접근 방식은 네트워크의 변화를 동적으로 반영하여 검증자가 기여도를 기준으로 선정되도록 하며, 공정성을 유지하는 데 중점을 둔다. 특히 합리적 비율(RP)을 공정성 측정 지표로 도입한 점이 주요 기여점이고, 이를 통해 노드의 지분과 보상 간의 차이를 정량적으로 평가하고, 학습 과정에 반영함으로써 보상 분배의 형평성을 향상시킨다.

그림 2, 그림 9, 그림 10은 RL-PoCS, PoCS, PoS 알고리즘에서의 각 노드별 합리성 지수를 보여준다. 이들 그래프에서 주요하게 관찰된 점들은 다음과 같다:

- RL-PoCS: 그림 2는 합리성 지수의 평균 값이 0.982로 대부분의 노드에서 합리성 지수가 1에 가까운 값을 보였으며, 기여도와 보상 간의 비율이 상대적으로 공정하게 분배되고 있음을 확인할 수 있다. 특정 노드(Node3)의 경우 다른 노드들보다 높은 지수를 보이지만, 전반적으로 수치가 평균적인 수준에 머물러 있어 보상 분배의 안정성이 높음을 알 수 있다.
- PoCS: 그림 9는 합리성 지수의 평균 값이 1.243으로 특정 노드(Node3)의 합리성 지수가 비교적 높게 나타났다. 이는 초기 보상 분배가 특정 노드에 집중되는 경향을 보이며, 보상 분배의 불균형이 존재함을 의미한다. 그러나 대부분의 다른 노드들 에서는 기준선(1)에 근접한 수치를 보여주고 있어, 일부 노드를 제외하고는 비교적 공정한 보상 분배가 이루어졌음을 확인할 수 있다.
- PoS: 그림 10의 경우, 합리성 지수의 평균 값이 6.672로 특정 노드(Node3)의 합리성 지수가 80% 이상으로 매우 높은 편차를 보여준다. 이는 검증자 노드의 선정에서 특정 노드에 대한 편향적인 보상 분배가 이루어지고 있음을 의미하며, 네트워크의 공정성 확보가 어려운 상황을 보여준다. 다른 노드들의 지수는 기준선에 근접하고 있으나, 여전히 분배의 불균형이 발생하고 있다.

결과적으로, 세 알고리즘 간의 비교를 통해 DDQN 방식의 RL-PoCS가 다른 방법들보다 상대적으로 공정한 보상 분배를 이루고 있음을 확인할 수 있다. Double DQN은 Q-learning에서 발생할 수 있는 과도한 행동 가치 평가 문제(overestimation)를 줄여주는 알고리즘이다. 이를 통해 각 노드의 기여도에 대한 보다 신뢰할 수 있는 평가를 가능하게 하여, 불공평한 보상 분배를 방지하고 네트워크 내 공평성을 유지한다. 이는 심층 신경망을 활용하여 복잡한 상태 공간에서의 최적 행동을 학습할 수 있어, 각 노드의 기여도를 더욱 세밀하게 평가하고, 네트워크의 상태 변화에 적응할 수 있다.

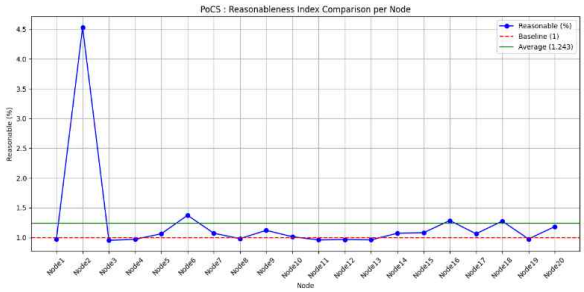

그림 9. PoCS 합리지수 비교
Fig. 9. PoCS Reasonableness index comparison

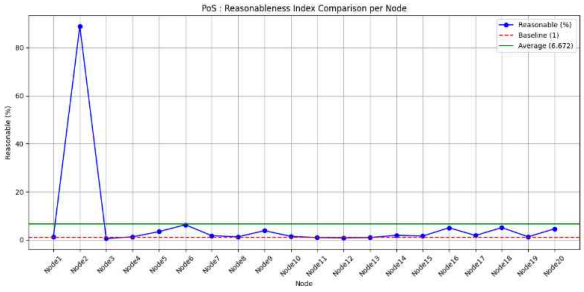

그림 10. PoS 합리 지수 비교
Fig. 10. PoS Reasonableness index comparison

본 연구의 주요 결과는, 강화학습이 분산되고 빠르게 진화하는 환경에서 검증자 선정을 효과적으로 해결할 수 있음을 보여준다. 시뮬레이션 결과, RL-PoCS 알고리즘이 블록체인 기반 시스템에 확장 가능한 솔루션을 제공할 뿐만 아니라, 참여 노드 간의 보상 분배에서 공정성을 크게 향상시키는 것을 확인할 수 있다. 이는 특히 메타버스 환경에서 탈 중앙화된 상호작용과 거래가 높은 빈도로 발생하기 때문에, 효율적이고 공정한 합의 메커니즘이 필수적인 상황에서 핵심적인 역할을 할 수 있다. 그러나, 연구는 몇 가지 도전 과제도 존재한다. 주된 제한 사항 중 하나는 동적 시스템인 메타버스에서 기여도 점수가 자주 변동될 수 있다는 점이다. 이로 인해 실시간 기여 활동을 정확하게 반영할 수 있는 효율적이고 신뢰할 수 있는 데이터 수집 메커니즘이 필요하다. 또한, 실시간으로 DDQN 모델을 학습하는 데 필요한 계산 오버헤드는 블록체

인 네트워크에서 확장성 문제를 야기할 수 있다는 점도 개선 과제로 제기된다.

향후 연구 방향:

- **동적 기여도 평가 모델의 개선**: RL-PoCS 알고리즘은 Double DQN을 사용하여 기여도를 평가하고 있지만, 메타버스와 같은 복잡한 환경에서 더욱 정밀한 기여도를 평가하기 위한 다중 에이전트 강화학습(MARL)[21] 기술을 적용하는 것이 유망할 수 있다. 여러 에이전트들이 공동의 목표를 달성하기 위해 협력하거나 경쟁하는 환경에서 기여도를 평가하는 방식으로, 네트워크의 효율성을 높일 수 있다.
- **메타버스 블록체인의 실 세계 응용**: 시뮬레이션이 유망한 결과를 제공했지만, RL-PoCS 알고리즘을 실 세계 메타버스 블록체인 환경에 배포하는 것이 향후 연구의 중요한 초점이 되어야 한다. 이러한 배포는 사용자 행동과 거래량을 포함한 실제 조건에서 알고리즘의 성능을 실증적으로 검증할 수 있는 기회를 제공한다.
- **보안 강화:** 공격자가 다수의 가짜 노드(Sybil 노드)를 생성하여 기여 점수를 인위적으로 높여 검증자로 반복 선정될 수 있다. 특정 그룹(카르텔)이 조직적으로 기여 점수를 조작하여 검증 프로세스를 독점할 수 있다. 그래서, 샤미르 비밀 공유 방식을 도입하여, 검증자가 선정되더라도, 최소 t 개 이상의 검증자가 DSS(Dynamic Secret Sharing) 서명을 제공해야 블록이 승인되도록 설계한다.

요약하면, RL-PoCS 알고리즘은 메타버스의 동적이고 탈중앙화 된 환경에서 검증자 선정 문제를 해결하는 유망한 솔루션을 제시한다. 이 접근 방식은 공정성, 확장성, 효율성의 균형을 맞추는 데 잠재력을 보여주지만, 대규모 블록체인 시스템에서 실용적으로 적용하려면 계산 및 구조적 과제를 해결해야 한다. 향후 연구를 통해 다중 에이전트 시스템, 분산 학습, 실세계 테스트를 통합함으로써 RL-PoCS 알고리즘은 메타버스의 차세대 블록체인 합의 메커니즘의 중요한 구성 요소가 될 수 있다.

Ⅷ. 결 론

본 논문에서는 RL-PoCS 알고리즘을 통해 메타버스 블록체인 환경에서 공정하고 확장 가능한 검증자 노드 선정 방식을 제안한다. 제안된 알고리즘은 강화학습과 PoCS 메커니즘을 결합하여, 네트워크의 동적 변화에 적응하고, 검증자 노드의 기여도에 따라 보상을 공정하게 분배한다. 특히, RP 개념을 도입하여 노드의 지분율과 보상 간의 공정성을 정량적으로 측정하고 학습 과정에 반영함으로써, 검증자 선정의 공정성을 향상시킨다. 시뮬레이션 결과, RL-PoCS 알고리즘은 메타버스 블록체인 환경에서 실시간 적응성과 공정한 자원 분배가 가능 하다는 것이 입증되었다. 이는 대규모 분산 시스템에서의 검증자 노드 선정 및 보상 분배의 효율성을 높이는 데 기여 할 것이다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원의 연구의 (IITP-2023-RS-2022-00156318)와 문화체육관광부 및 한국콘텐츠 진흥원의 2023년도 문화기술 연구개발사업(RS-2023-00219237)으로 수행된 연구로서, 관계부처에 감사드립니다.

참고문헌

[1] L.-H. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, ... and P. Hui, “All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda,” arXiv:2110.05352, 2021. https://doi.org/10.48550/arXiv.2110.05352

[2] N. Stephenson, *Snow Crash*, New York, NY: Bantam Books, 1992.

[3] S. Hollensen, P. Kotler, and M. O. Opresnik, “Metaverse – The New Marketing Universe,” *Journal of Business Strategy*, Vol. 44, No. 3, pp. 119-125, April 2023. https://doi.org/10.1108/JBS-01-2022-0014

[4] Q. Yang, Y. Zhao, H. Huang, Z. Xiong, J. Kang, and Z. Zheng, “Fusing Blockchain and AI with Metaverse: A Survey,” *IEEE Open Journal of the Computer Society*, Vol. 3, pp. 122-136, 2022. https://doi.org/10.1109/OJCS.2022.3188249

[5] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, “A Survey on Metaverse: Fundamentals, Security, and Privacy,” *IEEE Communications Surveys & Tutorials*, Vol. 25, No. 1, pp. 319-352, 2023. https://doi.org/10.1109/COMST.2022.3202047

[6] S. King and S. Nadal, PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake, Self-Published Paper 19, August 2012.

[7] P. Sethi, T. Nguyen, M. R. Chowdhury, S. Pirttikangas, and A. P. da Silva, “Fair Consensus in Blockchain with Heterogeneous Miners Using Reinforcement Learning Aided Adaptive Proof-of-Work,” in *Proceedings of 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, Las Vegas: NV, pp. 937-942, January 2024. https://doi.org/10.1109/CCNC51664.2024.10454844

[8] S. B. Thrun, Efficient Exploration in Reinforcement

Learning, Carnegie Mellon University, Pittsburgh: PA, Technical Report CMU-CS-92-102, January 1992.

[9] D. Han, B. Mulyana, V. Stankovic, and S. Cheng, "A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation," *Sensors*, Vol. 23, No. 7, 3762, April 2023. https://doi.org/10.3390/s23073762

[10] T. R. Gadekallu, T. Huynh-The, W. Wang, G. Yenduri, P. Ranaweera, Q.-V. Pham, ... and M. Liyanage, "Blockchain for the Metaverse: A Review," arXiv:2203.09738, March 2022. https://doi.org/10.48550/arXiv.2203.09738

[11] A. Salau, R. Dantu, K. Morozov, S. Badruddoja, and K. Upadhyay, "Making Blockchain Validators Honest," in *Proceedings of the 4th International Conference on Blockchain Computing and Applications (BCCA)*, San Antonio: TX, pp. 267-273, September 2022. https://doi.org/10.1109/BCCA55292.2022.9921952

[12] L. Marchesi, M. Marchesi, R. Tonelli, and M. I. Lunesu, "A Blockchain Architecture for Industrial Applications," *Blockchain: Research and Applications*, Vol. 3, No. 4, 100088, December 2022. https://doi.org/10.1016/j.bcra.2022.100088

[13] K. Hyun, "PoCS: Proof of Contribution Score Consensus Algorithm for Blockchain Using Metaverse Clients," *Journal of Virtual Convergence Research*, Vol. 1, No. 1, pp. 114-147, January 2025.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: The MIT Press, 1998.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., Cambridge, MA: The MIT Press, 2018.

[16] C. J. C. H. Watkins, Learning from Delayed Rewards, Ph.D. Dissertation, King's College, London, UK, May 1989.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, ... and D. Hassabis, "Human-Level Control through Deep Reinforcement Learning," *Nature*, Vol. 518, No. 7540, pp. 529-533, February 2015. https://doi.org/10.1038/nature14236

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," arXiv:1312.5602, December 2013. https://doi.org/10.48550/arXiv.1312.5602

[19] H. van Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-Learning," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix: AZ, pp. 2094-2100, February 2016. https://doi.org/10.1609/aaai.v30i1.10295

[20] Z. Liu, L. Hou, K. Zheng, Q. Zhou, and S. Mao, "A DQN-Based Consensus Mechanism for Blockchain in IoT Networks," *IEEE Internet of Things Journal*, Vol. 9, No. 14, pp. 11962-11973, July 2022. https://doi.org/10.1109/JIOT.2021.3132420

[21] T. Islam, F. H. Bappy, T. S. Zaman, M. S. I. Sajid, and M. M. A. Pritom, "MRL-PoS: A Multi-Agent Reinforcement Learning Based Proof of Stake Consensus Algorithm for Blockchain," in *Proceedings of 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas: NV, pp. 409-413, January 2024. https://doi.org/10.1109/CCWC60891.2024.10427777

현기정(Ki-Jeong Hyun)

2005년 : 서강대학교 정보통신대학원 (공학석사)

2004년~현 재: 엔씨소프트 팀장

2024년~현 재: 서강대학교 메타버스전문대학원 박사과정

※관심분야 : 메타버스 게임, 지능형 블록체인, 인공지능