(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0076767
(43) 공개일자 2023년05월31일

(51) 국제특허분류(Int. Cl.)
H04L 41/16 (2022.01) *G06N 3/08* (2023.01)
(52) CPC특허분류
H04L 41/16 (2022.05)
G06N 3/088 (2023.01)
(21) 출원번호 10-2022-0155978
(22) 출원일자 2022년11월21일
심사청구일자 2022년11월21일
(30) 우선권주장
1020210163277 2021년11월24일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이상현
서울특별시 동대문구 답십리로56길 105, 104동 1302호(답십리동, 답십리파크자이)
최승일
서울특별시 성북구 북악산로 913, 102동 901호(돈암동, 풍림아파트)
김홍기
경기도 광주시 오포읍 신현로 65-24, 205동 1301호(현대모닝사이드1차아파트)
(74) 대리인
송인호, 윤형근, 최관락

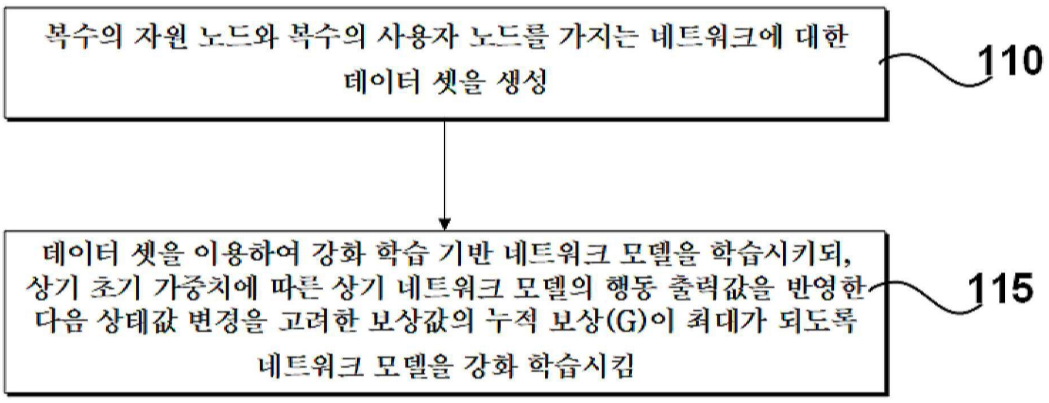
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **강화학습 기반 네트워크 운영 최적화 방법 및 그 장치**

(57) 요 약

강화학습 기반 네트워크 운영 최적화 방법 및 그 장치가 개시된다. 강화 학습 기반 네트워크 운영 최적화 방법은 복수의 자원 노드와 복수의 사용자 노드를 가지는 네트워크에 대한 데이터 셋을 생성하는 단계-상기 데이터 셋은 초기 상태값과 자원 노드와 사용자 노드 사이의 초기 가중치(weight)를 포함함; 및 상기 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 학습시키되, 상기 초기 가중치에 따른 상기 네트워크 모델의 출력인 행동 출력값을 반영한 다음 상태값 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델을 강화 학습시키는 단계를 포함한다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

복수의 자원 노드와 복수의 사용자 노드를 가지는 네트워크에 대한 데이터 셋을 생성하는 단계-상기 데이터 셋은 초기 상태값과 자원 노드와 사용자 노드 사이의 초기 가중치(weight)를 포함함; 및

상기 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 학습시키되, 상기 초기 가중치에 따른 상기 네트워크 모델의 출력인 행동 출력값을 반영한 다음 상태값 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델을 강화 학습시키는 단계를 포함하는 강화학습 기반 네트워크 운영 최적화 방법.

청구항 2

제1 항에 있어서,

상기 상태값은 각 자원 노드 제약 조건과 각 사용자 노드 제약 조건인 것을 특징으로 하는 강화학습 기반 네트워크 운영 최적화 방법.

청구항 3

제1 항에 있어서,

상기 보상값은 각 자원 노드와 각 사용자 노드에 대한 현재 상태에 대한 상기 네트워크 모델의 출력인 행동 출력값을 반영한 다음 상태값과 상기 현재 상태값을 네트워크 최적화를 위한 근사 함수식에 대입한 결과값을 이용하여 도출되는 것을 특징으로 하는 강화학습 기반 네트워크 운영 최적화 방법.

청구항 4

제3 항에 있어서,

상기 누적 보상(G)는 하기 수학식으로 계산되는 것을 특징으로 하는 강화학습 기반 네트워크 운영 최적화 방법.

$$G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$$

여기서, t는 단위 시간을 나타내고, γ은 감가율을 나타내고, r_t는

$$r_t = -\sum_{i,j} \| \rho_{ij}^{(t+1)} - f(\alpha_{ij}^{(t)}) \|^2 - \sum_{i,j} \| \alpha_{ij}^{(t+1)} - g(\rho_{ij}^{(t)}) \|^2$$

보상값을 나타내되, $\rho_{ij}^{(t+1)}$ 과 $\alpha_{ij}^{(t+1)}$ 는 행동 출력값이 반영된 다음 상태값을 나타내고, $\alpha_{ij}^{(t)}$ 와 $\rho_{ij}^{(t)}$ 는 현재 상태값을 나타내고, f()와 g()는 각각 근사 함수식인 것을 특징으로 하는 강화학습 기반 네트워크 운영 최적화 방법.

청구항 5

적어도 하나의 명령어를 저장하는 메모리; 및

상기 메모리에 저장된 명령어를 실행하는 프로세서를 포함하되,

상기 프로세서에 의해 실행된 명령어는,

복수의 자원 노드와 복수의 사용자 노드를 가지는 네트워크에 대한 데이터 셋을 생성하는 단계-상기 데이터 셋은 초기 상태값과 자원 노드와 사용자 노드 사이의 초기 가중치(weight)를 포함함; 및

상기 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 학습시키되, 상기 초기 가중치에 따른 상기 네트워크 모델의 행동 출력값을 반영한 다음 상태값 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델을 강화 학습시키는 단계를 수행하는 것을 특징으로 하는 컴퓨팅 장치.

청구항 6

제5 항에 있어서,

상기 네트워크 모델의 강화 학습을 위한 상태값은 각 자원 노드 제약 조건과 각 사용자 노드 제약 조건이되,

상기 네트워크 모델의 행동 출력값을 반영한 다음 상태값과 상기 현재 상태값을 네트워크 최적화 근사 함수식에 대입한 결과값을 이용하여 도출되는 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델의 가중치가 갱신되는 것을 특징으로 하는 컴퓨팅 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 강화학습 기반 네트워크 운영 최적화 방법 및 그 장치에 관한 것이다.

배 경 기 술

[0003] 네트워크 운영 최적화는 무선통신의 자원 할당, 사업체의 물류 관리 등 다양한 분야에서 응용되고 있다. 그러나 네트워크의 규모가 커지면 최적화 계산의 소요 시간이 길어지는 어려움이 있다.

[0004] 정해진 네트워크를 운영할 때 최적화의 경험을 학습하면 특정 순간의 최적화 계산의 결과가 얼마나 최적 상태에 가까운지 수치화 할 수 있다. 이는 과거의 최적화 방식이 좋았는지 나빴는지 알 수 있는 지표가 되므로 학습 기법으로 모형화 할 수 있다.

[0005] 강화학습은 MDP(Markov Decision Process) 기반의 인공지능 기반 최적화 기법으로 다양한 시도에 대한 시행착오(trial-and-error)를 통해 성능을 개선하도록 훈련한다.

[0006] 강화학습은 네트워크의 에너지 효율 개선, 네트워크 캐싱(cache-enabled opportunistic interference) 등 네트

워크 최적화에 많이 이용된다.

[0007] 대규모 사물 통신(massive machine type communication, mMTC)은 초연결(hyper-connectivity)을 목표로 한다. 초연결은 사람, 사물 등 모든 통신기기 간의 통신 연결을 제공하는 서비스이다. 서로 간의 통신이 원활하게 되기 위해서는 네트워크 최적화가 필요하나, 실시간으로 많은 노드들은 한번에 최적화할 수 있는 계산 수단이 없어 실현의 어려움을 겪고 있다.

발명의 내용

해결하려는 과제

[0009] 본 발명은 강화학습 기반 네트워크 운영 최적화 방법 및 그 장치를 제공하기 위한 것이다.

[0010] 또한, 본 발명은 무선 통신의 자원 할당이나 사업체의 물류 관리 등 설정한 목표를 최대로 하도록 네트워크를 운영할 수 있는 강화학습 기반 네트워크 운영 최적화 방법 및 그 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0012] 본 발명의 일 측면에 따르면 강화학습 기반 네트워크 운영 최적화 방법이 제공될 수 있다.

[0013] 본 발명의 일 실시예에 따르면, 복수의 자원 노드와 복수의 사용자 노드를 가지는 네트워크에 대한 데이터 셋을 생성하는 단계-상기 데이터 셋은 초기 상태값과 자원 노드와 사용자 노드 사이의 초기 가중치(weight)를 포함함; 및 상기 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 학습시키되, 상기 초기 가중치에 따른 상기 네트워크 모델의 출력인 행동 출력값을 반영한 다음 상태값 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델을 강화 학습시키는 단계를 포함하는 강화학습 기반 네트워크 운영 최적화 방법이 제공될 수 있다.

[0014] 상태값은 각 자원 노드 제약 조건과 각 사용자 노드 제약 조건이다.

[0015] 상기 보상값은 각 자원 노드와 각 사용자 노드에 대한 현재 상태에 대한 상기 네트워크 모델의 출력인 행동 출력값을 반영한 다음 상태와 상기 현재 상태를 네트워크 최적화 근사 함수식에 대입한 결과값을 이용하여 도출될 수 있다.

[0016] 상기 누적 보상(G)는 하기 수학식으로 계산되되,

$$G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$$

[0017]

[0018] 여기서, t는 단위 시간을 나타내고, γ은 감가율을 나타내고, r_t는

$$r_t = -\sum_{i,j} \| \rho_{ij}^{(t+1)} - f(\alpha_{ij}^{(t)}) \|^2 - \sum_{i,j} \| \alpha_{ij}^{(t+1)} - g(\rho_{ij}^{(t)}) \|^2$$

보상값을 나타내되, $\rho_{ij}^{(t+1)}$과 $\alpha_{ij}^{(t+1)}$는 행동 출력값이 반영된 다음 상태값을 나타내고,

$f(\alpha_{ij}^{(t)})$ 와 $g(\rho_{ij}^{(t)})$ 는 네트워크 최적화 근사 함수식이다.

[0020] 본 발명의 다른 측면에 따르면, 강화학습 기반 네트워크 운영 최적화 방법을 수행하기 위한 장치가 제공된다.

[0021] 본 발명의 일 실시예에 따르면, 적어도 하나의 명령어를 저장하는 메모리; 및 상기 메모리에 저장된 명령어를 실행하는 프로세서를 포함하되, 상기 프로세서에 의해 실행된 명령어는, 복수의 자원 노드와 복수의 사용자 노드를 가지는 네트워크에 대한 데이터 셋을 생성하는 단계-상기 데이터 셋은 초기 상태값과 자원 노드와 사용자 노드 사이의 초기 가중치(weight)를 포함함; 및 상기 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 학습시키되, 상기 초기 가중치에 따른 상기 네트워크 모델의 행동 출력값을 반영한 다음 상태값 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델을 강화 학습시키는 단계를 수행하는 것을 특징으로 하는 컴퓨팅 장치가 제공될 수 있다.

[0022] 상기 네트워크 모델의 강화 학습을 위한 상태값은 각 자원 노드 제약 조건과 각 사용자 노드 제약 조건이되, 상기 네트워크 모델의 행동 출력값을 반영한 다음 상태값과 상기 현재 상태값을 네트워크 최적화 근사 함수식에 대입한 결과값을 이용하여 도출되는 보상값의 누적 보상(G)이 최대가 되도록 상기 네트워크 모델의 가중치가 갱신될 수 있다.

발명의 효과

[0024] 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 운영 최적화 방법 및 그 장치를 제공함으로써, 무선 통신의 자원 할당이나 사업체의 물류 관리 등 설정한 목표를 최대로하는 최적해 계산 시간을 단축할 수 있는 이점이 있다.

도면의 간단한 설명

[0026] 도 1은 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 운영 최적화 방법을 나타낸 순서도.

도 2는 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 모델을 도시한 도면.

도 3은 본 발명의 다른 실시예에 따른 강화학습 기반 네트워크 모델을 도시한 도면.

도 4는 본 발명의 일 실시예에 따른 컴퓨팅 장치의 내부 구성을 개략적으로 도시한 블록도.

발명을 실시하기 위한 구체적인 내용

[0027] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0028] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0030] 도 1은 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 운영 최적화 방법을 나타낸 순서도이고, 도 2는 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 모델을 도시한 도면이고, 도 3은 본 발명의 다른 실시예에 따른 강화학습 기반 네트워크 모델을 도시한 도면이다. 이하에서는 무선 통신의 자원 할당을 가정하여 이를 중심으로 설명하나 반드시 자원 할당으로 제한되는 것은 아니다.

[0031] 단계 110에서 컴퓨팅 장치(100)는 복수의 자원 노드와 복수의 사용자 노드를 가지는 네트워크에 대한 데이터 셋

을 생성한다.

[0032] 여기서, 데이터 셋은 각 자원 노드 제약 조건과 각 사용자 노드 제약 조건에 대한 초기 상태($\boldsymbol{\alpha^0, \rho^0}$) 및 각 자원 노드와 각 사용자 노드 사이의 초기 가중치(weight)를 포함할 수 있다. 즉, 상태는 각 노드가 가져야 하는 상태일 수 있다.

[0033] 네트워크 최적화 문제는 NP-hard 문제라서 볼록성(convexity)가 보장되지 않아 인공지능으로 해결하기 어려운 문제점이 있다. 이에 본 발명의 일 실시예에서는 네트워크를 통계역학 시스템으로 해석하여 네트워크의 전체 에너지를 구하는 방식을 이용하기로 한다.

[0034] 네트워크의 에너지 최적화는 네트워크 최적화와 동일하다. 네트워크의 거시적 특성을 다룰 수 있는 통계 물리학(statistical physics) 이론인 복사 대칭 근사(replica symmetric approximation)을 이용하면 네트워크 최적화는 수학식 1 및 수학식 2와 같은 메시지 전달식을 만족하는 α_{ij}, ρ_{ij} 를 구하는 것과 동일하다.

수학식 1

[0035] $$\rho_{ij} = f(\alpha_{ij})$$

수학식 2

[0036] $$\alpha_{ij} = g(\rho_{ij})$$

[0037] 예를 들어, α_{ij}, ρ_{ij} 는 사용자 노드 제약 조건과 자원 노드 제약 조건을 나타낸다. 여기서, i와 j는 사용자 노드와 자원 노드 인덱스를 나타낸다.

[0038] 일반적으로 반복적인 계산을 통해 α_{ij}, ρ_{ij} 를 구한다. 그러나, 반복 계산 시간이 길어 실시간 네트워크에 적용이 어렵다. 따라서, 본 발명의 일 실시예에서는 강화 학습 기반 네트워크 모델을 사전 학습하여 실제 해에 유사한 근사에 빠른 속도로 도달하도록 하여 지연 시간을 줄일 수 있다.

[0039] 따라서, 본 발명의 일 실시예에서는 각 노드(즉, 자원 노드, 사용자 노드) 제약 조건 α_{ij}, ρ_{ij} 을 강화 학습을 위한 상태로 설정하기로 한다.

[0040] 단계 115에서 컴퓨팅 장치(100)는 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 학습시키되, 상기 초기 가중치에 따른 상기 네트워크 모델의 행동 출력값을 반영한 다음 상태값 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 네트워크 모델을 강화 학습시킨다.

[0041] 이에 대해 도 2를 참조하여 보다 상세히 설명하기로 한다.

[0042] 우선, 본 발명의 일 실시예에 따르면, 강화 학습 기반 네트워크 모델은 도 2에 도시된 바와 같다. 네트워크 모델은 강화 학습을 통해 네트워크 최적화(예를 들어, 자원 할당)를 위한 최적해를 도출하도록 학습될 수 있다.

[0043] 네트워크 모델의 강화 학습을 위한 상태는 α_{ij}, ρ_{ij} 이다. 여기서, α_{ij}, ρ_{ij} 는 각각 사용자 노드 제약

조건과 자원 노드 제약 조건을 나타낸다. 즉, α_{ij}, ρ_{ij}는 사용자 노드 i가 가질 수 있는 상태(예를 들어, 자원 노드)와 자원 노드 j가 가질 수 있는 상태를 각각 나타낸다. 여기서, $1 \le i \le M,\ 1 \le j \le N$이다.

[0044] 따라서, 강화 학습 기반 네트워크 모델이 강화 학습을 통해 $\rho_{ij}^{(t+1)} = f(\alpha_{ij}^{(t)})$와 $\alpha_{ij}^{(t+1)} = g(\rho_{ij}^{(t)})$를 만족하는 상태 $\alpha_{ij}^{(t)}, \rho_{ij}^{(t)}$를 도출하면, 네트워크 최적화에 성공하게 되며, 해당 조건은 누적 보상에서 반영될 수 있다.

[0045] 강화 학습 기반 네트워크 모델은 인공 신경망으로 구성될 수 있다. 강화 학습 기반 네트워크 모델은 데이터 셋으로 현재 상태, 초기 가중치 및 누적 보상 G_t를 입력받을 수 있다. 강화 학습 기반 네트워크 모델은 초기 가중치에 따라 현재 상태에 대한 행동 출력값($m_{ij}^{(t)}, n_{ij}^{(t)}$)을 출력하되, 해당 행동 출력값을 반영하여 다음 상태값이 결정될 수 있다. 예를 들어, 무선 자원 할당을 가정하는 경우, 상태 α_{ij}, ρ_{ij}가 각 노드(즉, 자원 노드, 사용자 노드)의 상태라고 가정하면, 행동 출력값은 각 노드간의 할당 결과일 수 있다.

[0046] 상태값 $\alpha_{ij}^{(t)}, \rho_{ij}^{(t)}$이 행동을 반복해 업데이트되면 최적화된 값으로 수렴하여 결과적으로 네트워크 최적화에 성공하게 된다.

[0047] 결과적으로, $\rho_{ij}^{(t+1)} = f(\alpha_{ij}^{(t)})$와 $\alpha_{ij}^{(t+1)} = g(\rho_{ij}^{(t)})$를 만족하는 상태 $\alpha_{ij}^{(t)}, \rho_{ij}^{(t)}$를 구하면, 네트워크 최적화에 성공하게 된다. 따라서, 본 발명의 일 실시예에 따른 강화 학습 기반 네트워크 모델은 $\alpha_{ij}^{(t)}, \rho_{ij}^{(t)}$를 강화 학습하여 $\alpha_{ij}^{(\infty)}, \rho_{ij}^{(\infty)}$를 구하는 것을 목표로 한다.

[0048] 초기값(즉, 초기 상태값)은 $\alpha_{ij}^{(0)}, \rho_{ij}^{(0)}$로 표기하기로 한다. 종래의 경우, $\alpha_{ij}^{(\infty)}, \rho_{ij}^{(\infty)}$를 구하기 위해 메시지 전달 식(근사 함수식) $f(x), g(x)$에 $\alpha_{ij}^{(0)}, \rho_{ij}^{(0)}$를 대입하여 $\alpha_{ij}^{(1)}, \rho_{ij}^{(1)}$를 구하고, 동일하게 복수회 순차적으로 반복하여 $\alpha_{ij}^{(t)}, \rho_{ij}^{(t)}$를 도출한 후 최종적으로 수렴하는

$\alpha_{ij}^{(\infty)}$, $\rho_{ij}^{(\infty)}$ 를 얻을 수 있다.

[0049] 본 발명의 일 실시예에서는 이를 강화 학습을 통해 학습하여 순차적으로 진행하는 과정을 줄여 전체 계산 과정을 단축시킬 수 있다.

[0050] 따라서, 본 발명의 일 실시예에서는 누적 보상(G)를 수학식 3과 같이 정의한다.

수학식 3

[0051]
$$G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$$

[0052] 여기서, t는 단위 시간을 나타내고, γ 은 감가율을 나타낸다. 또한, $r_{t'}$ 는 보상값을 나타낸다.

[0053] 여기서, $r_{t'}$ 는 수학식 4와 같이 계산될 수 있다.

수학식 4

[0054]
$$r_t = -\sum_{i,j} \| \rho_{ij}^{(t+1)} - f(\alpha_{ij}^{(t)}) \|^2 - \sum_{i,j} \| \alpha_{ij}^{(t+1)} - g(\rho_{ij}^{(t)}) \|^2$$

[0055] 만일 $\rho_{ij}^{(t+1)} = f(\alpha_{ij}^{(t)})$, $\alpha_{ij}^{(t+1)} = g(\rho_{ij}^{(t)})$ 를 만족하는 경우, 누적 보상은 더 이상 감소하지 않으며 수렴하게 된다. 따라서, 강화 학습을 통해 누적 보상(G_t)가 최대가 되도록 하면, $\alpha_{ij}^{(\infty)}$, $\rho_{ij}^{(\infty)}$ 를 구할 수 있게 된다.

[0056] 이에, 컴퓨팅 장치(100)는 데이터 셋을 이용하여 강화 학습 기반 네트워크 모델을 강화 학습시키되, 각 사용자 노드 제약 조건과 각 자원 노드 제약 조건을 각각 상태값 $\alpha_{ij}^{(t)}$, $\rho_{ij}^{(t)}$ 으로 설정하고, 해당 상태값이 강화 학습 기반 네트워크 모델의 출력인 행동 출력값에 의해 갱신될 수 있다.

[0057] 즉, $\rho_{ij}^{(t+1)} = \alpha_{ij}^{(t)} + m_{ij}^{(t)}$, $\alpha_{ij}^{(t)} = \rho_{ij}^{(t)} + n_{ij}^{(t)}$ 에서, $m_{ij}^{(t)}$, $n_{ij}^{(t)}$ 이 행동 출력값이며, 해당 행동은 보상에 따라 갱신될 수 있다.

$$\theta^{\mu} \leftarrow \theta^{\mu} + \alpha \frac{\partial J(\theta)}{\partial \theta}\Big|_{\theta=\theta^{\mu}}$$

[0058] 강화 학습 기반 네트워크 모델의 가중치는 로 표현되며, 이는 수학식 5와 같이 나타낼 수 있다.

수학식 5

$$\theta^{\mu} \leftarrow \theta^{\mu} + \alpha G_t \nabla_{\theta} \pi_{\theta}(m, n | \alpha, \rho)$$

[0059]

[0060] 그래디언트는 최적화 기법을 이용해 탐색하는 방식으로 구배를 추정해 모델의 가중치를 갱신할 수 있다. 예를 들어, Adam 알고리즘이 이용될 수 있다.

[0061] 본 발명의 일 실시예에 따르면, 컴퓨팅 장치(100)는 누적 보상을 이용한 손실을 역전파하여 그래디언트를 구한 후 가중치를 갱신할 수 있다.

[0062] 상술한 바와 같이, 본 발명의 일 실시예에 따른 컴퓨팅 장치는 초기 가중치를 강화 학습 기반 네트워크 모델에 적용하여 행동 출력값을 출력하고, 해당 행동 출력값이 반영된 다음 상태값과 근사 함수식을 적용한 결과값을 이용하여 도출되는 보상값을 누적한 누적 보상(G)이 최대가 되도록 네트워크 모델을 강화 학습시킬 수 있다.

[0063] 도 3에는 본 발명의 다른 실시예에 따른 강화 학습 기반 네트워크 모델이 도시되어 있다. 도 3에 도시된 바와 같이, 본 발명의 일 실시예에 따른 강화 학습 기반 네트워크 모델은 행동가-비평가 기반 네트워크 모델로 구현될 수도 있다. 도 3에 도시된 바와 같이, 비평가 네트워크 모델이 지역 변수로 네트워크 가중치를 가지도록 구현하여 행동가 네트워크 모델이 과거 모든 데이터를 저장하는 버퍼(replay buffer)를 구비하지 않도록 구현될 수 있다.

[0064] 도 3에 도시된 바와 같이, 정책 네트워크 블록과 평가 네트워크 블록을 복제하여 사용하는 것이다. 감쇠 상수 (τ)를 이용하여 네트워크 블록가 급변하여 발산하는 것을 방지할 수 있다.

[0065] 도 3에 도시된 바와 같이, 강화 학습 기반 네트워크 모델은 행동가 네트워크 블록과 비평가 네트워크 블록을 가지며, 행동가 네트워크 블록은 정책 네트워크 블록과 이를 복제한 목표 정책 네트워크 블록을 가질 수 있다. 또한, 비평가 네트워크 블록은 평가 네트워크 블록과 이를 복제한 목표 평가 네트워크 블록을 가질 수 있다.

[0066] 행동($m_{ij}^{(t)}$, $n_{ij}^{(t)}$)은 크기가 M x N인 실수 행렬일 수 있다. 정책 네트워크 블록은 상태와 가중치를 이용하여 행동의 분포($\pi(\boldsymbol{A}|\boldsymbol{s},\boldsymbol{\theta}^{\boldsymbol{\mu}})$)를 도출한다. 여기서, $\boldsymbol{s}$ 는 $\alpha_{ij}^{(t)}$, $\rho_{ij}^{(t)}$ 를 모두 포함하는 상태 행렬을 나타내고, $\boldsymbol{\theta}^{\mu}$ 는 정책 네트워크 블록의 가중치를 나타내며, $\boldsymbol{A}$ 는 $+m_{ij}^{(t)}$, $+n_{ij}^{(t)}$ 를 모두 포함하는 행동 행렬을 나타낸다. $\pi(\boldsymbol{A}|\boldsymbol{s},\boldsymbol{\theta}^{\boldsymbol{\mu}})$ 는

$+m_{ij}^{(t)}$, $+n_{ij}^{(t)}$의 사후 확률 분포이다. $\pi(\boldsymbol{A}|\boldsymbol{s},\boldsymbol{\theta}^{\boldsymbol{\mu}})$를 구하면 현재 상태 ($\alpha_{ij}^{(t)}$, $\rho_{ij}^{(t)}$)를 다음 상태($\alpha_{ij}^{(t+1)}$, $\rho_{ij}^{(t+1)}$)로 갱신할 수 있다.

[0067] 상술한 바와 같이, 강화 학습 기반 네트워크 모델을 이용하여 각 자원 노드와 사용자 노드의 최적해(상태, 할당)를 도출할 수 있는 이점이 있다.

[0069] 도 4는 본 발명의 일 실시예에 따른 컴퓨팅 장치의 내부 구성을 개략적으로 도시한 블록도이다.

[0070] 도 4를 참조하면, 본 발명의 일 실시예에 따른 컴퓨팅 장치(100)는 초기화부(410), 학습부(420), 메모리(430) 및 프로세서(440)를 포함하여 구성된다.

[0071] 초기화부(410)는 복수의 자원 노드와 복수의 사용자 노드를 포함하는 네트워크에 대한 데이터 셋을 생성하기 위한 수단이다. 여기서, 데이터 셋은 복수의 자원 노드와 사용자 노드 각각에 대한 초기 상태와 초기 가중치(weight) 및 누적 보상(G)을 포함할 수 있다. 또한, 초기 상태는 각 자원 노드 제약 조건과 각 사용자 노드 제약 조건으로 구성될 수 있다.

[0072] 학습부(420)는 데이터 셋을 기초로 강화 학습 기반 네트워크 모델을 학습시키기 위한 수단이다. 강화 학습 기반 네트워크 모델은 데이터 셋을 입력받아 각 노드(즉, 자원 노드, 사용자 노드)간의 행동 출력값을 출력할 수 있다. 학습부(420)는 행동 출력값을 반영한 다음 상태값의 변경을 고려한 보상값의 누적 보상(G)이 최대가 되도록 네트워크 모델을 강화 학습시킬 수 있다. 네트워크 모델로 입력되는 상태값은 해당 네트워크 모델의 출력인 행동 출력값에 의해 갱신되되, 해당 행동 출력값은 보상에 따라 갱신될 수 있다.

[0073] 메모리(430)는 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 운영 최적화 방법을 수행하기 위한 프로그램 코드를 저장하기 위한 수단이다.

[0074] 프로세서(450)는 본 발명의 일 실시예에 따른 강화학습 기반 네트워크 운영 최적화를 위한 컴퓨팅 장치(100)의 내부 구성 요소들(예를 들어, 초기화부(410), 학습부(420), 메모리(430) 등)을 제어하기 위한 수단이다.

[0075] 프로세서(440)는 메모리(430)에 저장된 프로그램 코드(명령어)를 실행할 수 있으며, 해당 프로그램 코드는 도 1에서 설명한 바와 같은 각각의 단계를 수행할 수 있다.

[0077] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0078] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0079] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모

든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0081] 100: 컴퓨팅 장치

410: 초기화부

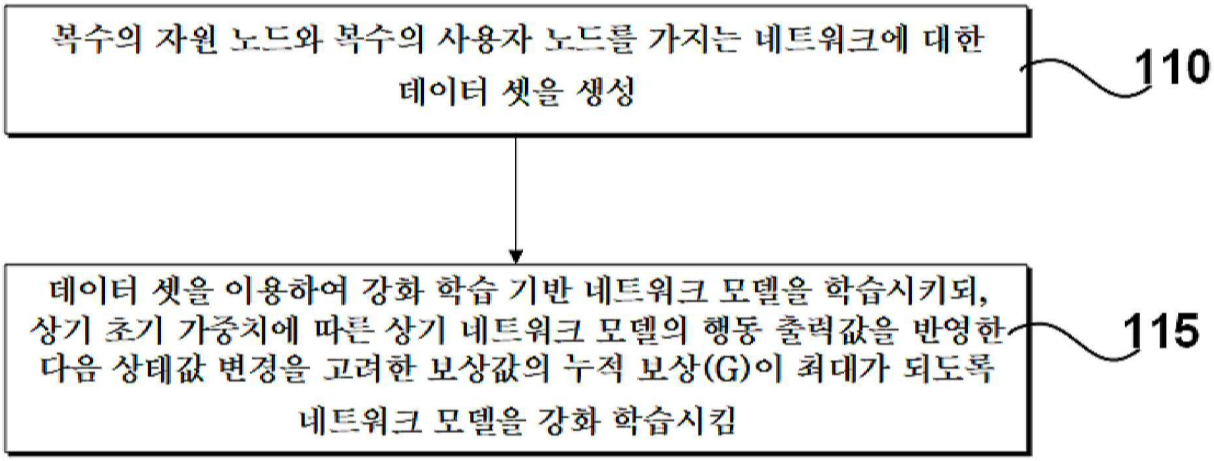
420: 학습부

430: 메모리

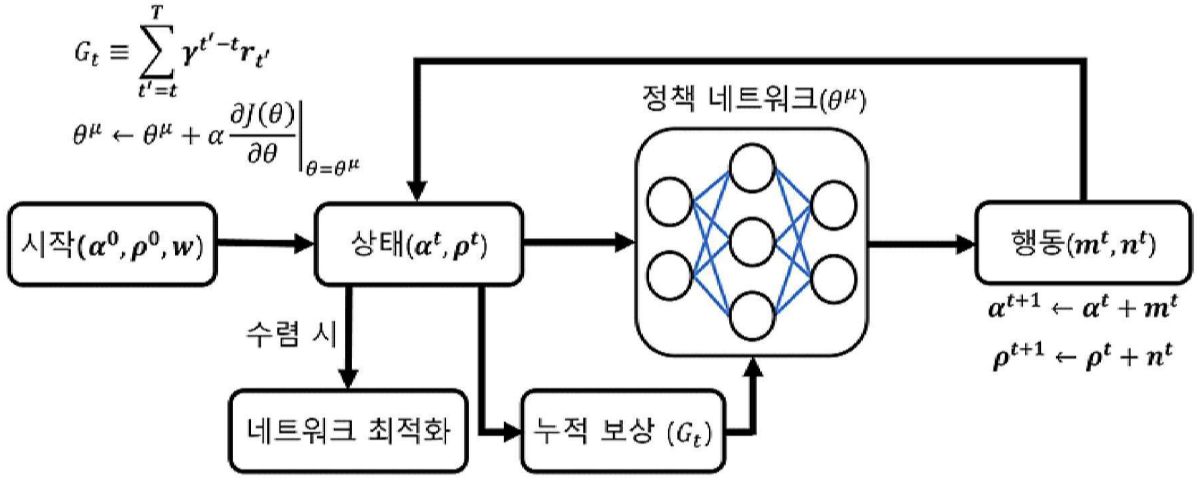
440: 프로세서

도면

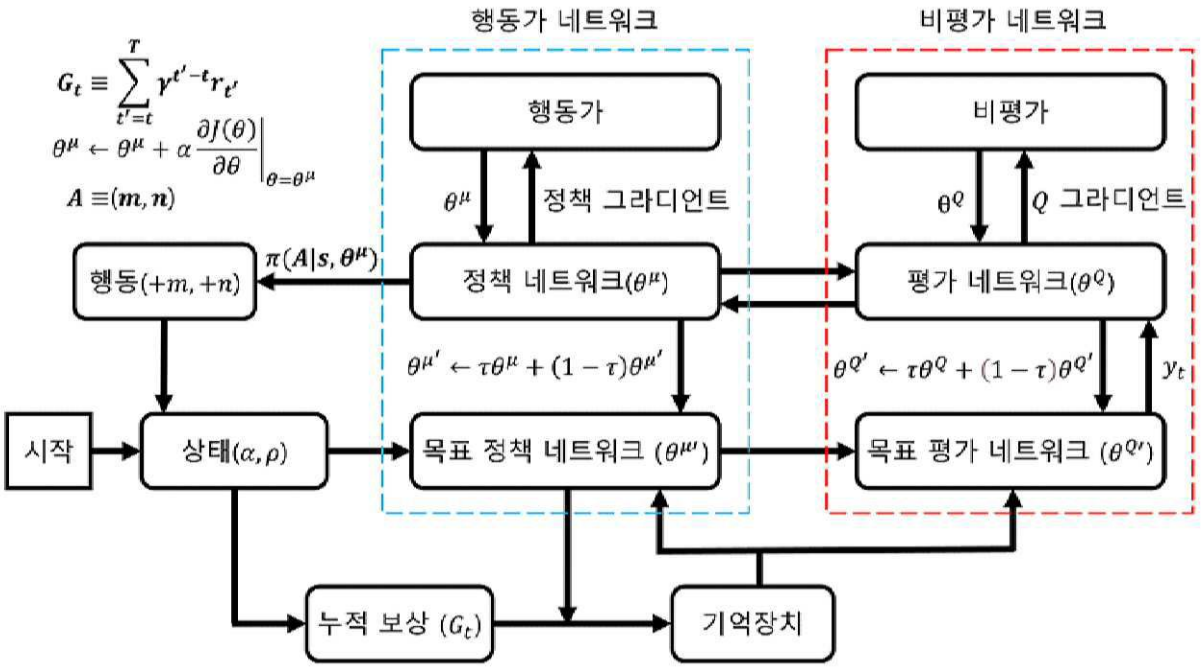
도면1



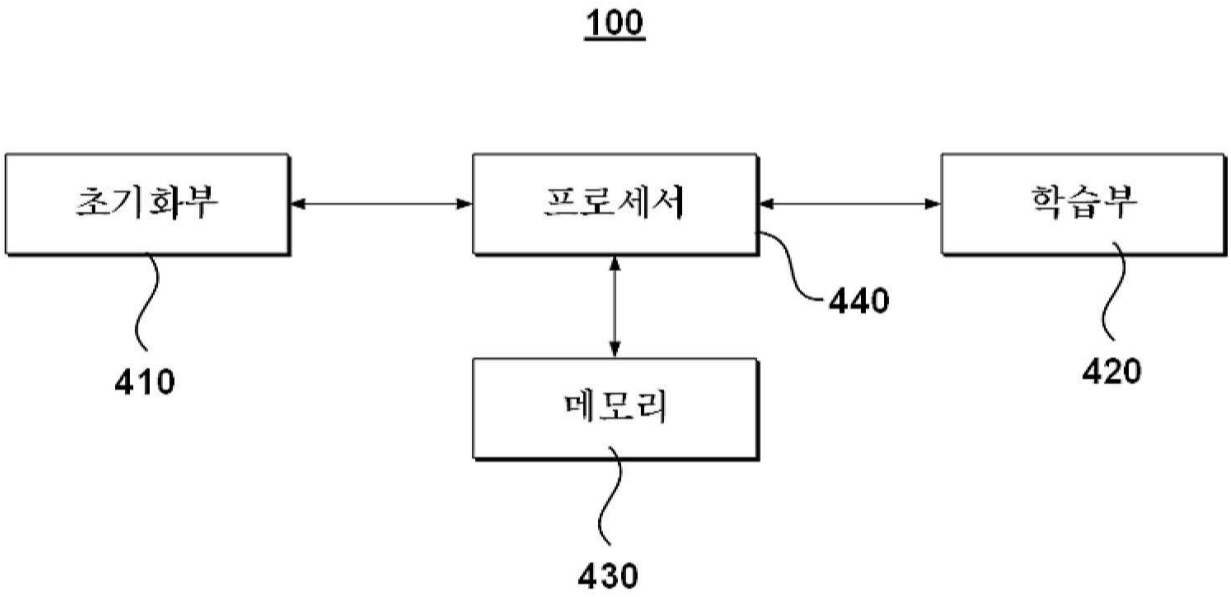
도면2



도면3


$G_t \equiv \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$
$\theta^{\mu} \leftarrow \theta^{\mu} + \alpha \frac{\partial J(\theta)}{\partial \theta}\Big|_{\theta=\theta^{\mu}}$
$A \equiv (m, n)$
행동가 네트워크
비평가 네트워크
행동가
비평가
θ^{μ}
정책 그라디언트
θ^{Q}
Q 그라디언트
행동$(+m, +n)$
$\pi(A|s, \theta^{\mu})$
정책 네트워크(θ^{μ})
평가 네트워크(θ^{Q})
$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$
$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$
y_t
시작
상태(α, ρ)
목표 정책 네트워크 $(\theta^{\mu'})$
목표 평가 네트워크 $(\theta^{Q'})$
누적 보상 (G_t)
기억장치

도면4


100
초기화부
프로세서
학습부
410
440
420
메모리
430