(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년07월01일
(11) 등록번호 10-2272501
(24) 등록일자 2021년06월28일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06N 3/063* (2006.01)
G06N 7/00 (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/063 (2013.01)
(21) 출원번호 10-2020-0050049
(22) 출원일자 2020년04월24일
심사청구일자 2020년04월24일
(56) 선행기술조사문헌
KR1020190069582 A
KR1020180091842 A
KR1020200037586 A
KR1020190077067 A

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김성륜
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C713호(신촌동)
차한
서울특별시 서대문구 연세로 50, 연세대학교 제3공학관 C707호(신촌동)
(74) 대리인
민영준

전체 청구항 수 : 총 20 항 심사관 : 송근배

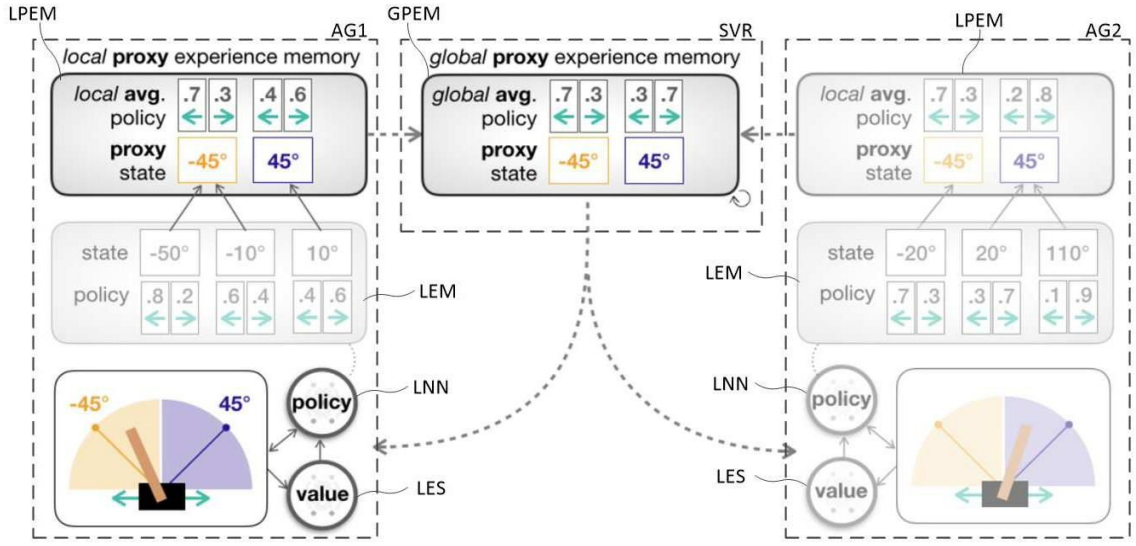
(54) 발명의 명칭 **분산 강화 학습 장치 및 방법**

(57) 요 약

본 발명은 주변 환경에 대한 상태 정보를 인가받고, 이전 학습된 패턴 추정 방식에 따라 상태 정보로부터 수행해야할 동작을 확률적으로 판별하여 나타내는 수행 동작 확률을 추정하는 로컬 신경망, 수행 동작 확률과 중앙 서버로부터 전송된 글로벌 수행 동작 확률로부터 손실값을 추정하여 로컬 신경망을 학습시키는 손실 추정부, 상태

(뒷면에 계속)

대 표 도



정보와 상태 정보에 대응하여 추정된 수행 동작 확률을 맵핑하여 저장하는 로컬 경험 메모리, 로컬 경험 메모리에 저장된 다수의 상태 정보를 미리 지정된 방식으로 클러스터링하여 대표 상태 정보가 미리 설정된 적어도 하나의 상태 클러스터로 구분하고, 구분된 적어도 하나의 상태 클러스터에 포함되는 상태 정보 각각에 맵핑된 수행 동작 확률로부터 상태 클러스터별 대표 수행 동작 확률을 획득하는 클러스터링부 및 중앙 서버로 전송하기 위해 적어도 하나의 상태 클러스터 각각에 대응하는 대표 상태 정보와 대표 수행 동작 확률을 맵핑하여 저장하는 로컬 프록시 메모리를 포함하여, 통신량을 크게 저감할 뿐만 아니라, 각 에이전트의 개별 정보를 보호할 수 있는 분산 강화 학습 장치 및 방법을 제공할 수 있다.

(52) CPC특허분류

G06N 7/005 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711102861
과제번호	2018-0-00923-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원(한국연구재단부설)
연구사업명	정보통신방송연구개발사업
연구과제명	주파수 공유 기반 Beyond 5G 통신 방식 연구
기 여 율	1/1
과제수행기관명	연세대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

주변 환경에 대한 상태 정보를 인가받고, 이전 학습된 패턴 추정 방식에 따라 상기 상태 정보로부터 수행해야할 동작을 확률적으로 판별하여 나타내는 수행 동작 확률을 추정하는 로컬 신경망;

상기 상태 정보와 상태 정보에 대응하여 추정된 수행 동작 확률을 맵핑하여 저장하는 로컬 경험 메모리;

상기 로컬 경험 메모리에 저장된 다수의 상태 정보를 미리 지정된 방식으로 클러스터링하여 대표 상태 정보가 미리 설정된 적어도 하나의 상태 클러스터로 구분하고, 구분된 적어도 하나의 상태 클러스터에 포함되는 상태 정보 각각에 맵핑된 수행 동작 확률로부터 상태 클러스터별 대표 수행 동작 확률을 획득하는 클러스터링부; 및

중앙 서버로 전송하기 위해 상기 적어도 하나의 상태 클러스터 각각에 대응하는 상기 대표 상태 정보와 대표 수행 동작 확률을 맵핑하여 저장하는 로컬 프록시 메모리를 포함하는 분산 강화 학습 장치.

청구항 2

제1항에 있어서, 상기 클러스터링부는

적어도 하나의 상태 클러스터에 각각에 포함되는 적어도 하나의 상태 정보에 각각 맵핑된 수행 동작 확률에 대해 기지정된 방식으로 통계값을 계산하여, 상기 대표 수행 동작 확률을 획득하는 분산 강화 학습 장치.

청구항 3

제1항에 있어서, 상기 분산 강화 학습 장치는

기지정된 조건 구간 동안 저장된 상기 대표 수행 동작 확률과 맵핑된 대표 상태 정보를 함께 상기 중앙 서버로 전송하는 분산 강화 학습 장치.

청구항 4

제1항에 있어서, 상기 분산 강화 학습 장치는

상기 대표 상태 정보를 나타낼 수 있는 적어도 하나의 상태 클러스터의 식별자를 상기 대표 수행 동작 확률과 함께 상기 중앙 서버로 전송하는 분산 강화 학습 장치.

청구항 5

제1항에 있어서, 글로벌 수행 동작 확률은

다수의 분산 강화 학습 장치 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 다수의 분산 강화 학습 장치로 전송되는 분산 강화 학습 장치.

청구항 6

다수의 에이전트로부터 기지정된 적어도 하나의 상태 클러스터에 대응하여 설정된 적어도 하나의 대표 상태 정보와 상기 적어도 하나의 대표 상태 정보 각각에 대응하여 맵핑된 적어도 하나의 대표 수행 동작 확률을 인가받아 저장하는 글로벌 경험 메모리;

상기 글로벌 경험 메모리에 저장된 적어도 하나의 대표 상태 정보 각각에 맵핑된 적어도 하나의 대표 수행 동작 확률로부터 기지정된 방식으로 글로벌 수행 동작 확률을 연산하는 글로벌 연산부; 및

상기 다수의 에이전트 각각으로 재분배 전송하기 위해, 상기 적어도 하나의 대표 상태 정보 각각에 대해 연산된 글로벌 수행 동작 확률을 맵핑하여 저장하는 글로벌 프록시 경험 메모리를 포함하는 분산 강화 학습 장치.

청구항 7

제6항에 있어서, 상기 글로벌 연산부는

적어도 하나의 대표 상태 정보에 각각에 맵핑된 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 통계값을 계산하여, 상기 글로벌 수행 동작 확률을 획득하는 분산 강화 학습 장치.

청구항 8

제6항에 있어서, 상기 분산 강화 학습 장치는

기지정된 조건 구간 동안 저장된 상기 글로벌 수행 동작 확률과 맵핑된 대표 상태 정보를 함께 상기 다수의 에이전트로 전송하는 분산 강화 학습 장치.

청구항 9

제6항에 있어서, 상기 분산 강화 학습 장치는

상기 적어도 하나의 대표 상태 정보 각각을 나타내는 적어도 하나의 식별자가 맵핑된 적어도 하나의 대표 수행 동작 확률과 함께 전송되면, 상기 글로벌 수행 동작 확률과 맵핑된 식별자를 상기 다수의 에이전트로 전송하는 분산 강화 학습 장치.

청구항 10

제6항에 있어서, 상기 적어도 하나의 대표 상태 정보는

다수의 에이전트 각각에서 주변 환경에 대해 획득된 상태 정보를 적어도 하나의 상태 클러스터로 구분하기 위해 미리 지정된 방식으로 클러스터링되고, 구분된 적어도 하나의 상태 클러스터 각각을 대표하는 상태 정보로 지정되어 획득되고,

상기 적어도 하나의 대표 수행 동작 확률은

다수의 에이전트 각각에서 이전 학습된 패턴 추정 방식에 따라 적어도 하나의 상태 정보로부터 수행해야할 동작이 확률적으로 표현되는 수행 동작 확률이 추정되어 맵핑되면, 상기 적어도 하나의 상태 클러스터에 포함되는 상태 정보 각각에 맵핑된 수행 동작 확률로부터 획득되는 분산 강화 학습 장치.

청구항 11

분산 강화 학습 장치에서 수행되는 분산 강화 학습 방법으로서,

주변 환경에 대한 상태 정보를 인가받고, 이전 학습된 패턴 추정 방식에 따라 상기 상태 정보로부터 수행해야할 동작을 확률적으로 판별하여 나타내는 수행 동작 확률을 추정하는 단계;

상기 상태 정보와 상태 정보에 대응하여 추정된 수행 동작 확률을 맵핑하여 저장하는 단계;

다수의 상태 정보를 미리 지정된 방식으로 클러스터링하여 대표 상태 정보가 미리 설정된 적어도 하나의 상태 클러스터로 구분하고, 구분된 적어도 하나의 상태 클러스터에 포함되는 상태 정보 각각에 맵핑된 수행 동작 확률로부터 상태 클러스터별 대표 수행 동작 확률을 획득하는 단계; 및

중앙 서버로 전송하기 위해 적어도 하나의 상태 클러스터 각각에 대응하는 상기 대표 상태 정보와 대표 수행 동작 확률을 맵핑하여 저장하는 단계를 포함하는 분산 강화 학습 방법.

청구항 12

제11항에 있어서, 상기 대표 수행 동작 확률을 획득하는 단계는

적어도 하나의 상태 클러스터에 각각에 포함되는 적어도 하나의 상태 정보에 각각 맵핑된 수행 동작 확률에 대해 기지정된 방식으로 통계값을 계산하여, 상기 대표 수행 동작 확률을 획득하는 분산 강화 학습 방법.

청구항 13

제11항에 있어서, 상기 분산 강화 학습 방법은

기지정된 조건 구간 동안 저장된 상기 대표 수행 동작 확률과 맵핑된 대표 상태 정보를 함께 상기 중앙 서버로 전송하는 분산 강화 학습 방법.

청구항 14

제11항에 있어서, 상기 분산 강화 학습 방법은

상기 대표 상태 정보를 나타낼 수 있는 적어도 하나의 상태 클러스터의 식별자를 상기 대표 수행 동작 확률과 함께 상기 중앙 서버로 전송하는 분산 강화 학습 방법.

청구항 15

제11항에 있어서, 글로벌 수행 동작 확률은

다수의 분산 강화 학습 방법 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 전송되는 분산 강화 학습 방법.

청구항 16

분산 강화 학습 장치에서 수행되는 분산 강화 학습 방법으로서,

다수의 에이전트로부터 기지정된 적어도 하나의 상태 클러스터에 대응하여 설정된 적어도 하나의 대표 상태 정보와 상기 적어도 하나의 대표 상태 정보 각각에 대응하여 맵핑된 적어도 하나의 대표 수행 동작 확률을 인가받아 저장하는 단계;

저장된 적어도 하나의 대표 상태 정보 각각에 맵핑된 적어도 하나의 대표 수행 동작 확률로부터 기지정된 방식으로 글로벌 수행 동작 확률을 연산하는 단계; 및

상기 다수의 에이전트 각각으로 재분배 전송하기 위해, 상기 적어도 하나의 대표 상태 정보 각각에 대해 연산된 글로벌 수행 동작 확률을 맵핑하여 저장하는 단계를 포함하는 분산 강화 학습 방법.

청구항 17

제16항에 있어서, 상기 글로벌 수행 동작 확률을 연산하는 단계는

적어도 하나의 대표 상태 정보에 각각에 맵핑된 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 통계값을 계산하여, 상기 글로벌 수행 동작 확률을 획득하는 분산 강화 학습 방법.

청구항 18

제16항에 있어서, 상기 분산 강화 학습 방법은

기지정된 조건 구간 동안 저장된 상기 글로벌 수행 동작 확률과 맵핑된 대표 상태 정보를 함께 상기 다수의 에이전트로 전송하는 분산 강화 학습 방법.

청구항 19

제16항에 있어서, 상기 분산 강화 학습 방법은

상기 적어도 하나의 대표 상태 정보 각각을 나타내는 적어도 하나의 식별자가 맵핑된 적어도 하나의 대표 수행 동작 확률과 함께 전송되면, 상기 글로벌 수행 동작 확률과 맵핑된 식별자를 상기 다수의 에이전트로 전송하는 분산 강화 학습 방법.

청구항 20

제16항에 있어서, 상기 적어도 하나의 대표 상태 정보는

다수의 에이전트 각각에서 주변 환경에 대해 획득된 상태 정보를 적어도 하나의 상태 클러스터로 구분하기 위해 미리 지정된 방식으로 클러스터링되고, 구분된 적어도 하나의 상태 클러스터 각각을 대표하는 상태 정보로 지정되어 획득되고,

상기 적어도 하나의 대표 수행 동작 확률은

다수의 에이전트 각각에서 이전 학습된 패턴 추정 방식에 따라 적어도 하나의 상태 정보로부터 수행해야할 동작이 확률적으로 표현되는 수행 동작 확률이 추정되어 맵핑되면, 상기 적어도 하나의 상태 클러스터에 포함되는

상태 정보 각각에 맵핑된 수행 동작 확률로부터 획득되는 분산 강화 학습 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 분산 강화 학습 장치 및 방법에 관한 것으로, 통신 비용을 저감하고 프라이버시를 보호할 수 있는 분산 강화 학습 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근 모바일 기기의 발전으로 인해, 무인 자동차나 드론 또는 스마트 팩토리의 자체 제어 로봇과 같은 지능형 자율 시스템이 적용되는 분야가 확장되고 있다. 이와 같은 지능형 자율 시스템을 사용하는 기기들은 주변 환경과 상호 작용하면서 실시간으로 의사 결정을 수행해야 한다.

[0003] 현재 이러한 지능형 자율 시스템은 인공 신경망으로 구현되는 경우가 많으며, 인공 신경망으로 구현되는 지능형 자율 시스템이 정상적으로 기능하기 위해서는 학습이 수행되어야 한다. 다만 지능형 자율 시스템이 적용되는 기기들이 안정적인 동작을 수행하기 위해서는 다양한 환경에 따른 상호 작용 결과에 대한 대량의 정보가 학습 정보로서 요구된다. 그러나 개별 기기들이 개별적으로 대량의 학습 정보를 획득하는 것은 현실적으로 매우 어렵다. 이러한 한계를 극복하기 위해 분산 강화 학습(Distributed Reinforcement Learning 또는 Distributed Prioritized Experience Replay이라고도 함) 기법이 제안되었다.

[0004] 분산 강화 학습은 다수의 기기들 각각이 지능형 자율 시스템 내에서 강화학습을 수행하는 주체인 다수의 에이전트(agent)로 동작하여 자신이 환경에 따른 상호 작용 결과로 획득한 지식(knowledge)을 경험 리플레이 메모리(experience replay memory)라는 지정된 형태의 학습 정보로 상호 교환함으로써, 대량의 학습 정보를 용이하게 획득하여 학습을 수행하는 방식을 나타낸다. 즉 다수의 에이전트 각각이 독립적으로 학습을 수행하는 것이 아니라, 경험 리플레이 메모리 교환을 통해 공동 집단 학습 방식으로 학습을 수행하게 된다.

[0005] 이 때, 분산 강화 학습에서는 다수의 에이전트가 경험 리플레이 메모리에 저장된 정보를 효율적으로 상호 교환할 수 있도록 다수의 에이전트 각각이 획득한 경험 리플레이 메모리에 정보를 수집하고, 수집된 정보를 다시 다수의 에이전트 각각으로 재분배하는 중앙 서버가 함께 이용되는 것이 일반적이다.

[0006] 이와 같이 분산 강화 학습 방식을 이용하는 경우, 다수의 에이전트에서 개별적으로 획득된 지식이 공통의 학습을 위해 이용될 수 있으므로, 용이하게 대량의 학습 정보를 획득하여 우수한 성능을 나타낼 수 있다는 장점이 있다. 그러나 다수의 에이전트가 경험 리플레이 메모리에 저장된 정보를 중앙 서버로 전송하고, 중앙 서버는 전송된 정보를 수집하여 다시 다수의 에이전트 각각으로 재분배해야 하므로, 대량의 경험 리플레이 메모리를 전송에 따른 통신량이 크게 증가되는 문제가 있다. 뿐만 아니라 다수의 에이전트 각각이 전송하는 경험 리플레이 메모리에 저장된 정보에는 각 에이전트의 상태 정보와 이에 대응하여 수행한 각종 동작 정보 등의 다양한 정보가 포함되어 각 에이전트의 개별 정보를 보호하지 못한다는 한계가 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국 공개 특허 제10-2019-0113928호 (2019.10.08 공개)

발명의 내용

해결하려는 과제

[0008] 본 발명의 목적은 경험 리플레이 메모리 크기를 줄여 통신 비용을 저감할 수 있는 분산 강화 학습 장치 및 방법을 제공하는데 있다.

[0009] 본 발명의 다른 목적은 경험 리플레이 메모리 교환 시에 각 에이전트의 개별 정보를 보호할 수 있는 분산 강화 학습 장치 및 방법을 제공하는데 있다.

과제의 해결 수단

[0010] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 분산 강화 학습 장치는 주변 환경에 대한 상태 정보를 인가받고, 이전 학습된 패턴 추정 방식에 따라 상기 상태 정보로부터 수행해야할 동작을 확률적으로 판별하여 나타내는 수행 동작 확률을 추정하는 로컬 신경망; 상기 수행 동작 확률과 중앙 서버로부터 전송된 글로벌 수행 동작 확률로부터 손실값을 추정하여 상기 로컬 신경망을 학습시키는 손실 추정부; 상기 상태 정보와 상태 정보에 대응하여 추정된 수행 동작 확률을 맵핑하여 저장하는 로컬 경험 메모리; 상기 로컬 경험 메모리에 저장된 다수의 상태 정보를 미리 지정된 방식으로 클러스터링하여 대표 상태 정보가 미리 설정된 적어도 하나의 상태 클러스터로 구분하고, 구분된 적어도 하나의 상태 클러스터에 포함되는 상태 정보 각각에 맵핑된 수행 동작 확률로부터 상태 클러스터별 대표 수행 동작 확률을 획득하는 클러스터링부; 및 상기 중앙 서버로 전송하기 위해 상기 적어도 하나의 상태 클러스터 각각에 대응하는 상기 대표 상태 정보와 대표 수행 동작 확률을 맵핑하여 저장하는 로컬 프록시 메모리를 포함한다.

[0011] 상기 클러스터링부는 적어도 하나의 상태 클러스터에 각각에 포함되는 적어도 하나의 상태 정보에 각각 맵핑된 수행 동작 확률에 대해 기지정된 방식으로 통계값을 계산하여, 상기 대표 수행 동작 확률을 획득할 수 있다.

[0012] 상기 분산 강화 학습 장치는 기지정된 조건 구간 동안 저장된 상기 대표 수행 동작 확률과 맵핑된 대표 상태 정보를 함께 상기 중앙 서버로 전송할 수 있다.

[0013] 상기 분산 강화 학습 장치는 상기 대표 상태 정보를 나타낼 수 있는 적어도 하나의 상태 클러스터의 식별자를 상기 대표 수행 동작 확률과 함께 상기 중앙 서버로 전송할 수 있다.

[0014] 상기 글로벌 수행 동작 확률은 다수의 분산 강화 학습 장치 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 다수의 분산 강화 학습 장치로 전송될 수 있다.

[0015] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 분산 강화 학습 장치는 다수의 에이전트로부터 기지정된 적어도 하나의 상태 클러스터에 대응하여 설정된 적어도 하나의 대표 상태 정보와 상기 적어도 하나의 대표 상태 정보 각각에 대응하여 맵핑된 적어도 하나의 대표 수행 동작 확률을 인가받아 저장하는 글로벌 경험 메모리; 상기 글로벌 경험 메모리에 저장된 적어도 하나의 대표 상태 정보 각각에 맵핑된 적어도 하나의 대표 수행 동작 확률로부터 기지정된 방식으로 글로벌 수행 동작 확률을 연산하는 글로벌 연산부; 및 상기 다수의 에이전트 각각으로 재분배 전송하기 위해, 상기 적어도 하나의 대표 상태 정보 각각에 대해 연산된 글로벌 수행 동작 확률을 맵핑하여 저장하는 글로벌 프록시 경험 메모리를 를 포함한다.

[0016] 상기 목적을 달성하기 위한 본 발명의 또 다른 실시예에 따른 분산 강화 학습 방법은 주변 환경에 대한 상태 정보를 인가받고, 이전 학습된 패턴 추정 방식에 따라 상기 상태 정보로부터 수행해야할 동작을 확률적으로 판별하여 나타내는 수행 동작 확률을 추정하는 단계; 상기 수행 동작 확률과 중앙 서버로부터 전송된 글로벌 수행 동작 확률로부터 손실값을 판별하여 상기 수행 동작 확률을 추정하는 패턴 추정 방식을 학습하는 단계; 상기 상태 정보와 상태 정보에 대응하여 추정된 수행 동작 확률을 맵핑하여 저장하는 단계; 다수의 상태 정보를 미리 지정된 방식으로 클러스터링하여 대표 상태 정보가 미리 설정된 적어도 하나의 상태 클러스터로 구분하고, 구분된 적어도 하나의 상태 클러스터에 포함되는 상태 정보 각각에 맵핑된 수행 동작 확률로부터 상태 클러스터별 대표 수행 동작 확률을 획득하는 단계; 및 상기 중앙 서버로 전송하기 위해 적어도 하나의 상태 클러스터 각각에 대응하는 상기 대표 상태 정보와 대표 수행 동작 확률을 맵핑하여 저장하는 단계를 포함한다.

[0017] 상기 목적을 달성하기 위한 본 발명의 또 다른 실시예에 따른 분산 강화 학습 방법은 다수의 에이전트로부터 기지정된 적어도 하나의 상태 클러스터에 대응하여 설정된 적어도 하나의 대표 상태 정보와 상기 적어도 하나의 대표 상태 정보 각각에 대응하여 맵핑된 적어도 하나의 대표 수행 동작 확률을 인가받아 저장하는 단계; 저장된 적어도 하나의 대표 상태 정보 각각에 맵핑된 적어도 하나의 대표 수행 동작 확률로부터 기지정된 방식으로 글로벌 수행 동작 확률을 연산하는 단계; 및 상기 다수의 에이전트 각각으로 재분배 전송하기 위해, 상기 적어도 하나의 대표 상태 정보 각각에 대해 연산된 글로벌 수행 동작 확률을 맵핑하여 저장하는 단계를 포함한다.

발명의 효과

[0018] 따라서, 본 발명의 실시예에 따른 분산 강화 학습 장치 및 방법은 서로 유사한 다수의 경험 리플레이 메모리를 동일 클러스터로 클러스터링하고, 클러스터를 기반으로 프록시 경험 리플레이 메모리를 생성하여 중앙 서버로 전송하고, 중앙 서버로부터 프록시 경험 리플레이 메모리를 인가받아 학습을 수행함으로써, 통신량을 크게 저감

할 뿐만 아니라, 각 에이전트의 개별 정보를 보호할 수 있다.

도면의 간단한 설명

[0019] 도 1은 분산 강화 학습 시스템의 개략적 구조를 나타낸다.

도 2는 도 1의 분산 강화 학습 시스템이 수행하는 학습 목표의 일 예로 카트폴 게임을 설명하기 위한 도면이다.

도 3은 본 발명의 일 실시예에 따른 분산 강화 학습 시스템의 개략적 구조를 나타낸다.

도 4 및 도 5는 본 발명의 일 실시예에 따른 분산 강화 학습 방법을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0020] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0021] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0022] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0023] 도 1은 분산 강화 학습 시스템의 개략적 구조를 나타내고, 도 2는 도 1의 분산 강화 학습 시스템이 수행하는 학습 목표의 일 예로 카트폴 게임을 설명하기 위한 도면이다.

[0024] 도 1 및 도 2를 참조하면, 분산 강화 학습 시스템은 중앙 서버(SVR)와 다수의 에이전트(AG1, AG2)를 포함한다.

[0025] 다수의 에이전트(AG1, AG2) 각각은 대응하는 기기로부터 주변 환경에 대한 상태 정보(state)를 획득하고, 이전까지 학습된 패턴 추정 방식을 기초로 획득된 상태 정보(state)에 따라 기기가 수행해야할 동작(action)을 확률적으로 판별하여 나타내는 수행 동작 확률을 추정한다. 분산 강화 학습 기법에서는 수행 동작 확률을 정책(policy)이라고도 한다.

[0026] 그리고 다수의 에이전트(AG1, AG2) 각각은 획득된 상태 정보(state)에 따라 추정된 수행 동작 확률(policy)을 기반으로 기기가 대응하는 동작을 수행하도록 하고, 중앙 서버(SVR)로부터 다른 에이전트가 각 상태 정보(state)에 따라 추정한 수행 동작 확률이 수신되면, 획득된 상태 정보(state)와 수행 동작 확률(policy) 그리고 중앙 서버(SVR)로부터 전달된 상태 정보(state)와 수행 동작 확률(policy)을 기반으로 손실값(value)을 추정하여 강화 학습을 수행한다.

[0027] 이때 분산 강화 학습 기법에서는 다수의 에이전트(AG1, AG2) 각각은 자신이 획득한 상태 정보(state)와 상태 정보(state)에 따라 추정한 수행 동작 확률(policy)뿐만 아니라, 다른 에이전트가 획득한 상태 정보(state)와 추정한 수행 동작 확률(policy)을 중앙 서버(SVR)를 통해 인가받아, 이후 다양한 상태 정보(state)에 대응하는 수행 동작 확률(policy)을 추정하기 위해 이용한다. 즉 다른 에이전트의 상태 정보(state)와 추정한 수행 동작 확률(policy) 또한 학습 정보로 이용하여 학습을 수행한다.

[0028] 이에 다수의 에이전트(AG1, AG2) 각각은 로컬 신경망(local neural networks: LNN)과 손실 추정부(LES)를 포함할 수 있다. 로컬 신경망(LNN)은 획득된 상태 정보(state)에 따라 수행 동작 확률(policy)을 추정하는 인공 신경망이고, 손실 추정부(LES)는 로컬 신경망(LNN)이 추정한 수행 동작 확률(policy)과 중앙 서버(SVR)로부터 전달된 상태 정보(state)와 수행 동작 확률(policy)을 기반으로 손실값(value)을 추정하여 로컬 신경망(LNN)에 대한 강화 학습을 수행한다. 여기서 수행 동작 확률(policy)을 추정하는 로컬 신경망(LNN)은 액터(actor)라고도 하며, 학습을 위한 손실값(value)을 추정하는 손실 추정부(LES)는 크리틱(critic)라고도 한다.

[0029] 그리고 다수의 에이전트(AG1, AG2) 각각은 획득된 상태 정보(state)와 이에 대응하여 추정한 수행 동작 확률(policy)을 서로 맵핑하여 저장하는 로컬 경험 메모리(local experiencer memory)(LEM)를 더 포함한다. 로컬 경험 메모리(LEM)는 상기한 경험 리플레이 메모리(experience replay memory)에 대응하는 구성으로, 다수의 상

태 정보(state)와 대응하는 수행 동작 확률(policy)을 저장하고, 저장된 다수의 상태 정보(state)와 수행 동작 확률(policy)을 중앙 서버(SVR)로 전송하기 위해 구비된다. 로컬 경험 메모리(LEM)는 기지정된 조건 구간 동안의 상태 정보(state)와 대응하는 수행 동작 확률(policy)을 서로 맵핑하여 저장할 수 있다. 여기서 지정된 조건 구간은 시간 또는 특정 조건으로 다양하게 설정될 수 있다.

[0030] 일 예로 도 2의 카트폴(cartpole) 게임을 수행하는 기기의 경우를 살펴보면, 카트폴 게임은 카트(cart) 상부면에 수직 방향으로 세워진 폴(pole)이 쓰러지지 않도록 카트(cart)를 이동시키는 게임을 의미한다. 여기서 폴(pole)의 하단은 카트(cart)의 상부면의 피봇 지점(pivot point)에 결합되어 x축 방향으로만 회전하여 쓰러질 수 있으며, 기기는 카트(cart)를 x축 방향으로 기지정된 속도로 이동시킬 수 있는 것으로 가정한다.

[0031] 카트폴 게임을 수행하는 기기에서 기기는 피봇 지점에 결합된 폴(pole)이 카트(cart)의 상부면에 대한 수직 방향으로부터 회전한 각도(Θ)를 상태 정보(state)로 획득하여 대응하는 에이전트로 제공할 수 있다. 그리고 에이전트는 상태 정보(state)에 따른 수행 동작 확률(policy)로 추정하고, 추정된 수행 동작 확률(policy)에 따라 결정되는 동작(action)을 기기로 전달하여 기기가 해당 동작을 수행하도록 한다.

[0032] 도 1의 예에서는 카트폴 게임이 종료될 때까지 기지정된 시간 간격 단위로 획득된 상태 정보(state)와 각 상태 정보(state)에 대응하는 수행 동작 확률(policy)이 맵핑되어 로컬 경험 메모리(LEM)에 저장되는 것으로 가정하였으며, 이에 도 1에서 제1 에이전트(AG1)의 로컬 경험 메모리(LEM)에는 3가지(k = 0, 1, 2) 상태에 대한 상태 정보(state)와 각 상태 정보(state)에 대응하는 수행 동작 확률(policy)이 맵핑되어 저장되어 있다.

[0033] 그리고 수행 동작 확률(policy)은 각 상태 정보(state)에 하나로 추정되어 맵핑되는 것이 아니라, 도 1에 나타난 바와 같이, 다수로 추정되어 맵핑될 수 있다. 일 예로 제1 에이전트(AG1)의 로컬 경험 메모리(LEM)에서 폴(pole)이 -50° 인 경우의 상태 정보($state_{k=0}$) 대해 좌측 방향 이동 확률 0.8와 우측 방향 이동 확률 0.2로 두 가지 수행 동작 확률(policy)이 맵핑되어 있고, 폴(pole)이 10° 인 경우의 상태 정보($state_{k=2}$) 대해서는 좌측 방향 이동 확률 0.4와 우측 방향 이동 확률 0.6로 두 가지 수행 동작 확률(policy)이 맵핑되어 있음을 알 수 있다.

[0034] 중앙 서버(SVR)는 다수의 에이전트(AG1, AG2) 각각의 로컬 경험 메모리(LEM)에 저장된 다수의 상태 정보(state)와 대응하는 수행 동작 확률(policy)을 인가받아 글로벌 경험 메모리(gloval experiencer memory)(GEM)에 저장한다. 그리고 글로벌 경험 메모리(GEM)에 저장된 다수의 상태 정보(state)와 수행 동작 확률(policy)을 다시 다수의 에이전트(AG1, AG2) 각각으로 재분배한다.

[0035] 즉 중앙 서버(SVR)는 다수의 에이전트(AG1, AG2) 각각에서 획득된 상태 정보(state)와 수행 동작 확률(policy)을 취합하고, 취합된 상태 정보(state)와 수행 동작 확률(policy)을 다시 다수의 에이전트(AG1, AG2)로 재분배함으로써, 다수의 에이전트(AG1, AG2)가 학습 정보인 상태 정보(state)와 수행 동작 확률(policy)을 상호 공유하여 공동으로 이용할 수 있도록 한다.

[0036] 그러나 이 경우 다수의 에이전트(AG1, AG2) 각각이 획득한 많은 양의 상태 정보(state)와 수행 동작 확률(policy)을 중앙 서버(SVR)로 전송하고, 중앙 서버(SVR)는 다수의 에이전트(AG1, AG2)로부터 취합된 대량의 상태 정보(state)와 수행 동작 확률(policy)을 다시 다수의 에이전트(AG1, AG2) 각각으로 전송해야 하므로, 전송해야 하는 통신량이 매우 크게 발생된다. 비록 도 1에서는 설명의 편의를 위하여 2개의 에이전트(AG1, AG2)만을 표시하였으나, 분산 강화 학습 시스템에는 대량의 에이전트가 포함될 수 있다. 따라서 통신 비용이 크게 발생된다. 뿐만 아니라, 다수의 에이전트(AG1, AG2) 각각이 자신이 획득한 상태 정보(state)와 이에 따른 수행 동작 확률(policy)을 중앙 서버(SVR)로 그대로 전송함에 따라 다수의 에이전트(AG1, AG2) 각각에 대한 개별 정보가 보호되지 않는다.

[0037] 여기서는 간단하게 카트폴 게임을 수행하는 기기를 예로 들어 설명하였으나, 에이전트가 포함되는 기기는 용도에 따라 보호되어야 하는 중요 정보를 포함할 수도 있으므로, 각 에이전트(AG1, AG2)가 획득한 개별 정보를 보호할 필요성이 있다.

[0038] 도 3은 본 발명의 일 실시예에 따른 분산 강화 학습 시스템의 개략적 구조를 나타낸다.

[0039] 도 3을 참조하면, 본 실시예에 따른 분산 강화 학습 시스템은 도 1의 분산 강화 학습 시스템과 마찬가지로 중앙 서버(SVR)와 다수의 에이전트(AG1, AG2)를 포함한다. 그리고 다수의 에이전트(AG1, AG2) 각각은 획득된 상태 정보(state)에 따라 수행 동작 확률(policy)을 추정하는 로컬 신경망(LNN)과 클러스터링부(미도시), 손실 추정부(LES) 및 획득된 상태 정보(state)와 이에 대응하는 수행 동작 확률(policy)을 서로 맵핑하여 저장하는 로컬

경험 메모리 (LEM)를 포함한다.

[0040] 도 3의 에이전트(AG1, AG2) 또한 도 1의 에이전트(AG1, AG2)와 마찬가지로, 2개의 로컬 신경망(LNN, LNN2)과 로컬 경험 메모리(LEM)를 포함하며, 2개의 로컬 신경망(LNN, LNN2)과 로컬 경험 메모리(LEM)의 동작은 도 1과 동일하므로, 여기서는 상세하게 설명하지 않는다.

[0041] 그러나 본 실시예에 따른 다수의 에이전트(AG1, AG2) 각각은 클러스터링부(미도시)와 로컬 프록시 경험 메모리(local proxy experiencer memory)(LPEM)를 더 포함한다.

[0042] 클러스터링부는 로컬 경험 메모리(LEM)에 저장된 다수의 상태 정보(state)를 분석하여, 기지정된 기준 범위 단위로 다수의 상태 정보(state)를 구분하고, 동일한 범위에 포함되는 것으로 구분된 상태 정보(state)들을 상태 클러스터로 클러스터링하고, 각 상태 클러스터에 대한 대표 상태 정보(proxy state)를 설정한다.

[0043] 그리고 클러스터링부는 각 클러스터에 포함된 상태 정보(state)에 대응하는 수행 동작 확률(policy)로부터 기지정된 방식으로 연산하여 대표 상태 정보(proxy state) 각각에 대응하는 대표 수행 동작 확률(proxy policy)을 획득하고, 획득된 대표 상태 정보(proxy state)와 대응하는 대표 수행 동작 확률(proxy policy)을 맵핑하여 로컬 프록시 경험 메모리(LPEM)에 저장한다.

[0044] 여기서 대표 상태 정보(proxy state)는 일 예로 클러스터에 포함된 상태 정보(state)에 대응하는 수행 동작 확률(policy)에 대해 평균값과 같이 통계적 방식으로 연산하여 획득될 수 있으나, 각 상태 정보(state)에 따른 가중 평균값 등의 방식으로 획득되거나, 다른 기지정된 방식으로 연산되어 획득될 수도 있다. 이하에서는 수행 동작 확률(policy)에 대한 평균값을 계산하여 대표 수행 동작 확률(proxy policy)을 획득하는 것으로 가정하여 설명한다.

[0045] 도 3의 예에서는 클러스터링부는 카트폴 게임에서 폴(pole)이 카트 상부면으로부터 수직 방향(0°)을 기준으로 다수의 상태 정보(state)를 좌측과 우측으로 구분하여 2개의 상태 클러스터로 클러스터링하고, 각 상태 클러스터에 대한 대표 상태 정보(proxy state)를 -45° 와 45° 로 설정한 경우를 도시하였다.

[0046] 이에 로컬 경험 메모리(LEM)에 저장된 -50°, -10°, 10° 의 3가지(k = 0, 1, 2) 상태에 대한 상태 정보(state)는 로컬 프록시 경험 메모리(LPEM)에서 2개의 상태 클러스터로 구분되어, -45° 와 45° 의 2개의 대표 상태 정보(proxy state)로 저장되었음을 알 수 있다. 여기서는 설명의 편의를 위하여 폴(pole)의 회전 각도만으로 상태 정보(state)로 설정하였으나, 카트 위치, 카트 속도, 폴 각도, 폴 팁 속도(velocity of pole tip) 등과 같은 서로 다른 다양한 상태 조건을 상태 정보로 포함할 수도 있다. 즉 상태 정보에는 다양한 환경 정보가 포함될 수도 있다.

[0047] 그리고 2개의 대표 상태 정보(proxy state) 중 제1 대표 상태 정보(-45°)에 대응하는 대표 수행 동작 확률(proxy policy)은 -50°, -10° 의 2가지 상태에 대한 수행 동작 확률(policy)의 평균값을 계산하여, 좌측 방향 이동 확률 0.7(=(0.8+0.6)/2)와 우측 방향 이동 확률 0.3(=(0.2+0.4)/2)로 두 가지 수행 동작 확률이 맵핑되어 저장되고, 제2 대표 상태 정보(45°)에 대응하는 수행 동작 확률(policy)은 10° 상태에 대한 상태 정보(state)만이 존재하므로, 대표 수행 동작 확률(proxy policy)은 그대로 좌측 방향 이동 확률 0.4와 우측 방향 이동 확률 0.6로 두 가지 수행 동작 확률이 맵핑되어 저장되었음을 알 수 있다.

[0048] 그리고 다수의 에이전트(AG1, AG2) 각각은 로컬 프록시 경험 메모리(LPEM)는 저장된 대표 상태 정보(proxy state)와 대응하는 대표 수행 동작 확률(proxy policy)을 중앙 서버(SVR)로 전송한다.

[0049] 상기한 바와 같이, 도 3에서는 다수의 에이전트(AG1, AG2) 각각이 로컬 경험 메모리(LEM)에 저장된 다수의 상태 정보(state)와 이에 대응하는 다수의 수행 동작 확률(policy)을 그대로 중앙 서버(SVR)로 전송하지 않고, 다수의 상태 정보(state)를 다수의 클러스터로 클러스터링하여, 각 클러스터에 대한 대표 상태 정보(proxy state)를 획득하고, 각 클러스터에 포함된 상태 정보(state)에 대응하는 다수의 수행 동작 확률(policy)에 대한 대표 수행 동작 확률(proxy policy)을 기지정된 방식으로 계산하여 로컬 프록시 경험 메모리(LPEM)에 저장하며, 로컬 프록시 경험 메모리(LPEM)에 저장된 대표 상태 정보(proxy state)와 대표 수행 동작 확률(proxy policy)을 중앙 서버(SVR)로 전송한다. 따라서 설정된 클러스터 개수에 대응하는 개수의 대표 상태 정보(proxy state)와 대표 수행 동작 확률(proxy policy)만을 중앙 서버(SVR)로 전송하므로 통신량을 크게 저감시킬 수 있다.

[0050] 또한 획득된 상태 정보(state)와 이에 대응하는 수행 동작 확률(policy)을 그대로 전달하지 않고, 대표 상태 정보(proxy state)와 대표 수행 동작 확률(proxy policy)로 변환하여 전달하므로 다수의 에이전트(AG1, AG2) 각각에 대한 개별 정보를 보호할 수 있도록 한다. 즉 다수의 에이전트(AG1, AG2) 각각의 실질적 상태와 이에 따

른 실질적 대응 동작을 그대로 중앙 서버(SVR)로 전송하지 않으므로, 다수의 에이전트(AG1, AG2)의 개별 정보를 보호할 수 있다.

[0051] 한편, 도 3에서 중앙 서버(SVR)는 글로벌 연산부(미도시) 및 글로벌 경험 메모리(GEM)가 아닌 글로벌 프록시 경험 메모리(global proxy experiencer memory)(GPEM)를 포함한다. 도 1에 도시된 글로벌 경험 메모리(GEM)의 경우, 다수의 에이전트(AG1, AG2) 각각에서 전송된 상태 정보(state)와 수행 동작 확률(policy)을 저장하고, 저장된 상태 정보(state)와 수행 동작 확률(policy)을 단순히 다수의 에이전트(AG1, AG2)로 재분배하였다.

[0052] 그에 비해 도 3의 중앙 서버(SVR)에서는 다수의 에이전트(AG1, AG2) 각각으로부터 클러스터 개수에 대응하는 개수의 대표 상태 정보(proxy state)와 대표 수행 동작 확률(proxy policy)이 전달되면, 글로벌 연산부가 다수의 에이전트(AG1, AG2)로부터 전달된 대표 상태 정보(proxy state) 각각에 대응하는 대표 수행 동작 확률(proxy policy)로부터 기지정된 방식으로 글로벌 수행 동작 확률(global proxy policy)을 연산하여 획득된 글로벌 프록시 경험 메모리(GPEM)에 저장한다. 그리고 글로벌 프록시 경험 메모리(GPEM)에 저장된 대표 상태 정보(proxy state)와 대응하는 글로벌 수행 동작 확률(global proxy policy)을 함께 다수의 에이전트(AG1, AG2)로 분배한다.

[0053] 즉 중앙 서버(SVR) 또한 다수의 에이전트(AG1, AG2)로부터 전달되는 대표 상태 정보(proxy state)에 대한 다수의 대표 수행 동작 확률(proxy policy)로부터 글로벌 수행 동작 확률(global proxy policy)을 획득하여 글로벌 프록시 경험 메모리(GPEM)에 저장하고, 저장된 대표 상태 정보(proxy state)와 대응하는 글로벌 수행 동작 확률(global proxy policy)만을 다수의 에이전트(AG1, AG2)로 전달함으로써, 중앙 서버(SVR)에서 다수의 에이전트(AG1, AG2)로 전송하는 통신량을 크게 저감시킬 수 있다.

[0054] 상기에서는 다수의 에이전트(AG1, AG2)와 중앙 서버(SVR)가 대표 상태 정보(proxy state)를 전달하는 것으로 설명하였으나, 대표 상태 정보(proxy state)는 미리 지정될 수 있다. 즉 다수의 에이전트(AG1, AG2)는 모두 동일한 방식으로 다수의 상태 정보(state)를 구분하여 클러스터링해야 하므로, 각 클러스터를 대표하는 대표 상태 정보(proxy state)는 다수의 에이전트(AG1, AG2)와 중앙 서버(SVR)에 미리 지정될 수 있다.

[0055] 본 실시예에서 다수의 에이전트(AG1, AG2)와 중앙 서버(SVR)가 대표 상태 정보(proxy state)를 전송하는 것은 단순히 대표 수행 동작 확률(proxy policy)과 글로벌 수행 동작 확률(global proxy policy)이 포함되는 클러스터를 식별하기 위한 식별자로 이용하기 위해서이다. 그러므로 경우에 따라서는 다수의 에이전트(AG1, AG2)와 중앙 서버(SVR)가 대표 상태 정보(proxy state)가 나타내는 클러스터에 따른 식별자를 대표 수행 동작 확률(proxy policy)과 글로벌 수행 동작 확률(global proxy policy)에 맵핑하여 함께 전송하도록 구성될 수도 있다. 즉 대표 상태 정보(proxy state)가 아닌 클러스터 식별자를 전송하여 통신량을 더욱 줄일 수 있을 뿐만 아니라, 상태 정보의 의미가 노출되지 않도록 하여 보안성을 더욱 향상시킬 수 있다.

[0056] 한편 대표 상태 정보(proxy state)와 대응하는 글로벌 수행 동작 확률(global proxy policy)을 인가받은 다수의 에이전트(AG1, AG2)는 로컬 경험 메모리(LEM)에 저장된 다수의 상태 정보(state)와 대응하는 수행 동작 확률(policy)과 함께 인가된 대표 상태 정보(proxy state)와 대응하는 글로벌 수행 동작 확률(global proxy policy)을 기반으로 로컬 신경망(LNN)을 강화 학습시킨다.

[0057] 손실 추정부(LES)는 로컬 신경망(LNN)이 상태 정보(state)에 따라 추정한 수행 동작 확률(policy)과 중앙 서버(SVR)로부터 전달된 대표 상태 정보(proxy state)와 대응하는 글로벌 수행 동작 확률(global proxy policy)를 기반으로 손실값(value)을 계산하여 강화 학습을 수행할 수 있다. 여기서 손실값은 상태 정보에 따른 수행 동작 확률(policy)과 글로벌 수행 동작 확률(global proxy policy) 사이의 교차 엔트로피 손실로 계산될 수 있다.

[0058] 본 실시예의 분산 강화 학습 시스템에서 다수의 에이전트(AG1, AG2) 각각은 개별적으로 지정된 기능을 수행하는 각종 기기에 포함될 수 있으며, 분산 강화 학습 방식으로 학습되어 기기가 주변 환경과 상호 작용하여 수행할 동작을 결정하는 분산 강화 학습 장치로 볼 수 있으며, 중앙 서버(SVR) 또한 다수의 에이전트(AG1, AG2)에서 전송된 대표 수행 동작 확률(proxy policy)로부터 글로벌 수행 동작 확률(global proxy policy)를 획득하여 다수의 에이전트(AG1, AG2)로 분배하는 분산 강화 학습 장치로 볼 수 있다.

[0059] 도 4 및 도 5는 본 발명의 일 실시예에 따른 분산 강화 학습 방법을 설명하기 위한 도면으로 도 4는 에이전트의 동작을 설명하기 위한 도면이고, 도 5는 중앙 서버의 동작을 설명하기 위한 도면이다.

[0060] 도 4를 참조하면, 다수의 에이전트(AG1, AG2) 각각은 기기로부터 주변 환경에 대한 상태 정보(state)를 획득한다(S11). 그리고 이전까지 학습된 패턴 추정 방식을 기초로 획득된 상태 정보(state)에 따라 기기가 수행해야 할 동작(action)을 확률적으로 판별하여 나타내는 수행 동작 확률(policy)을 추정하고, 추정된 상태 정보

(state)와 수행 동작 확률(policy)을 맵핑하여 함께 저장한다(S12). 이에 기기는 추정된 수행 동작 확률(policy)에 기반하여 동작을 수행하고, 에이전트(AG1, AG2)는 동작 수행 결과로부터 손실값(value)을 추정하여 기지정된 방식으로 강화 학습을 수행한다(S13).

[0061] 그리고 기지정된 조건 구간 동안 획득된 상태 정보(state)를 미리 지정된 방식으로 클러스터링하여 다수의 상태 클러스터로 구분한다(S14). 이때 다수의 상태 클러스터 각각에 대한 대표 상태 정보(proxy state)가 미리 설정될 수 있다.

[0062] 다수의 상태 클러스터로 구분되면, 구분된 상태 클러스터에 포함된 적어도 하나의 상태 정보(state)에 맵핑된 적어도 하나의 수행 동작 확률(policy)에 대해 기지정된 방식으로 연산을 수행하여 대표 수행 동작 확률(proxy policy)을 획득한다(S15).

[0063] 그리고 획득된 대표 수행 동작 확률(proxy policy)을 중앙 서버(SVR)로 전송한다(S16). 이때, 에이전트(AG1, AG2)는 대표 수행 동작 확률(proxy policy)에 대응하는 대표 상태 정보(proxy state) 또는 대표 상태 정보(proxy state)를 식별할 수 있는 식별자를 함께 중앙 서버(SVR)로 전송할 수 있다.

[0064] 이후 중앙 서버(SVR)로부터 다수의 에이전트에서 전송된 다수의 대표 수행 동작 확률(proxy policy)로부터 기지정된 방식으로 연산되어 획득된 글로벌 수행 동작 확률(global proxy policy)이 수신되는지 판별한다(S17). 만일 글로벌 수행 동작 확률(global proxy policy)이 수신되면, 수신된 글로벌 수행 동작 확률(global proxy policy)과 이전 획득된 수행 동작 확률(policy)을 기반으로 강화 학습을 수행한다(S18).

[0065] 한편 도 5를 참조하면, 중앙 서버(SVR)는 다수의 에이전트(AG1, AG2) 중 적어도 하나의 에이전트로부터 대표 수행 동작 확률(proxy policy)이 수신되는지 판별한다(S21). 그리고 적어도 하나의 에이전트로부터 대표 수행 동작 확률(proxy policy)이 수신되면, 대표 수행 동작 확률(proxy policy)을 저장한다(S22). 그리고 이전 적어도 하나의 에이전트들로부터 수신된 대표 수행 동작 확률(proxy policy)과 함께 기지정된 방식으로 연산을 수행하여, 글로벌 수행 동작 확률(global proxy policy)을 획득한다(S23). 글로벌 수행 동작 확률(global proxy policy)이 획득되면, 획득된 글로벌 수행 동작 확률(global proxy policy)을 분산 강화 학습에 참여한 다수의 에이전트로 전송한다(S24). 이때 중앙 서버(SVR)는 글로벌 수행 동작 확률(global proxy policy)에 대응하는 대표 상태 정보(proxy state) 또는 대표 상태 정보(proxy state)를 식별할 수 있는 식별자를 함께 에이전트로 전송할 수 있다.

[0066] 본 발명에 따른 방법은 컴퓨터에서 실행시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0067] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0068] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

부호의 설명

[0069] AG1, AG2: 에이전트 SVR: 중앙 서버

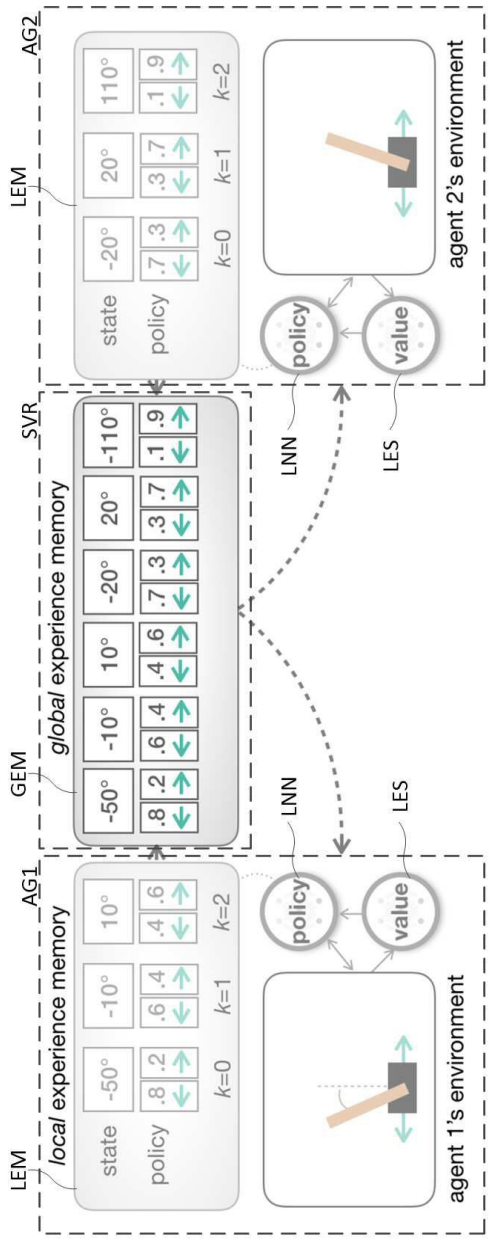
LNN: 로컬 신경망 LES: 손실 추정부

LEM: 로컬 경험 메모리 GEM: 글로벌 경험 메모리

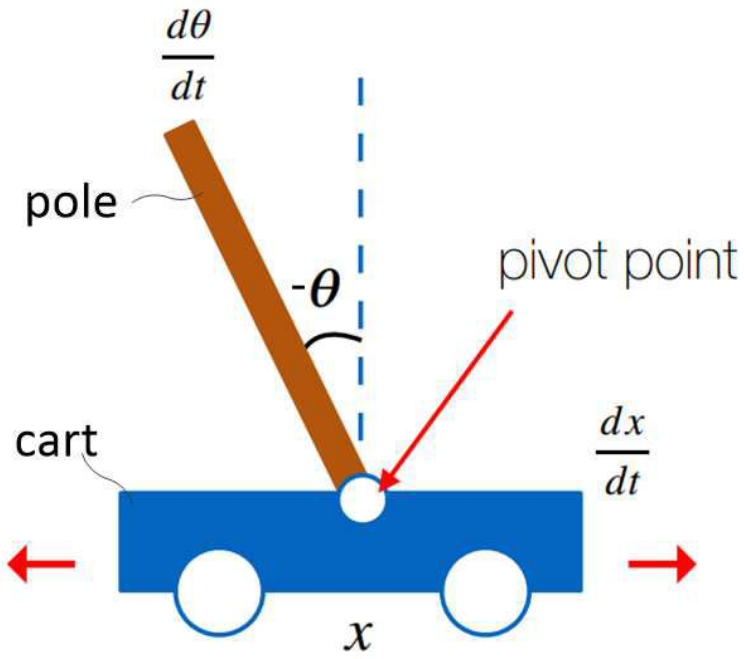
LPEM: 로컬 프록시 경험 메모리 GPEM: 글로벌 프록시 경험 메모리

도면

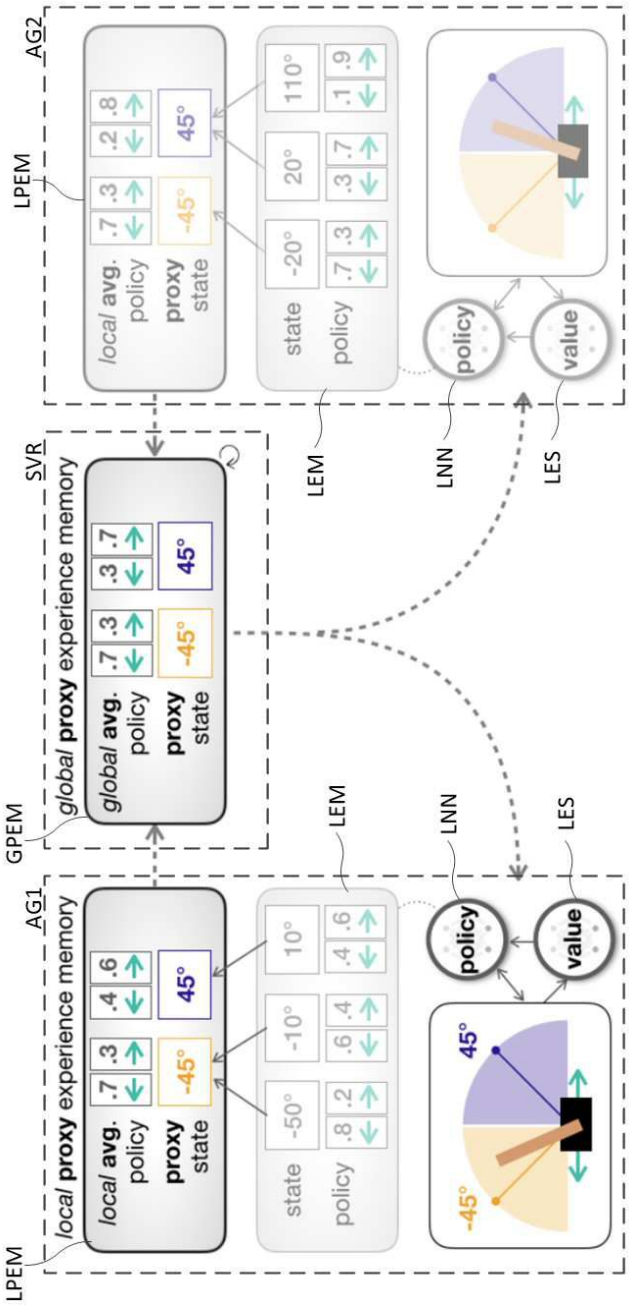
도면1


AG2
LEM
state
policy
-20°
20°
110°
k=0
k=1
k=2
agent 2's environment
policy
value
SVR
GEM
global experience memory
-50°
-10°
10°
-20°
20°
-110°
LNN
LES
AG1
local experience memory
agent 1's environment

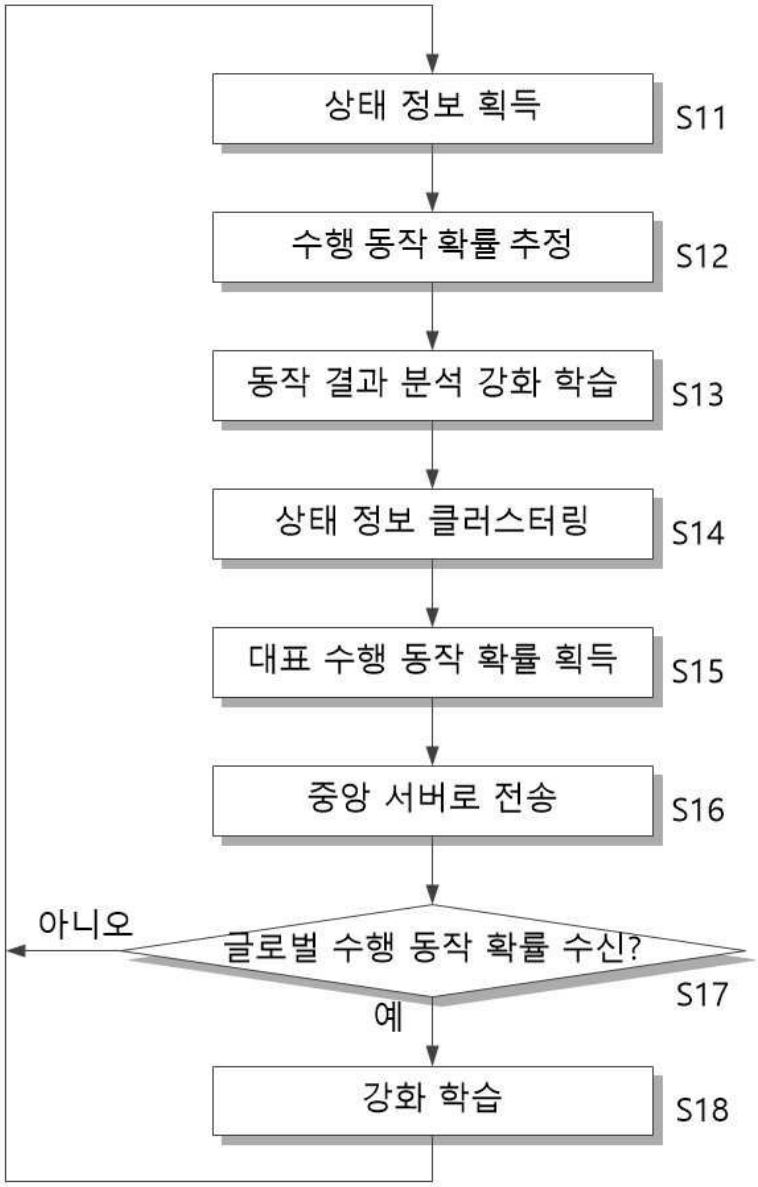
도면2



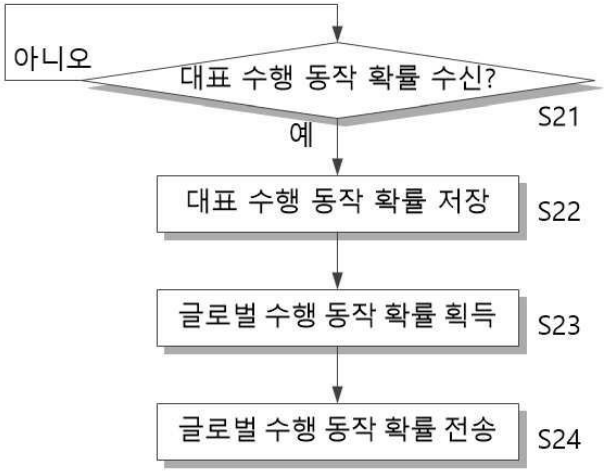
도면3


AG2
LPEM
local avg. policy
proxy state
.2 .8
.7 .3
45°
-45°
state
policy
-20°
20°
110°
.7 .3
.3 .7
.1 .9
LEM
LNN
LES
policy
value
SVR
GPEM
global proxy experience memory
global avg. policy
proxy state
AG1
local proxy experience memory
-50°
-10°
10°
.8 .2
.6 .4
.4 .6

도면4



도면5



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 5

【변경전】

제1항에 있어서, 상기 글로벌 수행 동작 확률은

다수의 분산 강화 학습 장치 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 다수의 분산 강화 학습 장치로 전송되는 분산 강화 학습 장치.

【변경후】

제1항에 있어서, 글로벌 수행 동작 확률은

다수의 분산 강화 학습 장치 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 다수의 분산 강화 학습 장치로 전송되는 분산 강화 학습 장치.

【직권보정 2】

【보정항목】 청구범위

【보정세부항목】 청구항 15

【변경전】

제11항에 있어서, 상기 글로벌 수행 동작 확률은

다수의 분산 강화 학습 방법 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 전송되는 분산 강화 학습 방법.

【변경후】

제11항에 있어서, 글로벌 수행 동작 확률은

다수의 분산 강화 학습 방법 각각이 적어도 하나의 클러스터 각각에 대응하여 획득한 다수의 대표 수행 동작 확률에 대해 기지정된 방식으로 계산된 통계값으로 상기 중앙 서버에 의해 획득되어 전송되는 분산 강화 학습 방법.