(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0025055
(43) 공개일자 2025년02월21일

(51) 국제특허분류(Int. Cl.)
G06N 3/092 (2023.01) *G06N 3/04* (2023.01)
(52) CPC특허분류
G06N 3/092 (2023.01)
G06N 3/04 (2023.01)
(21) 출원번호 10-2023-0106006
(22) 출원일자 2023년08월14일
심사청구일자 없음

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이병준
서울특별시 광진구 아차산로36길 39, 701동 1513호
장지안
경기도 고양시 일산동구 일산로 205, 204동 901호
(74) 대리인
김준석, 박민욱

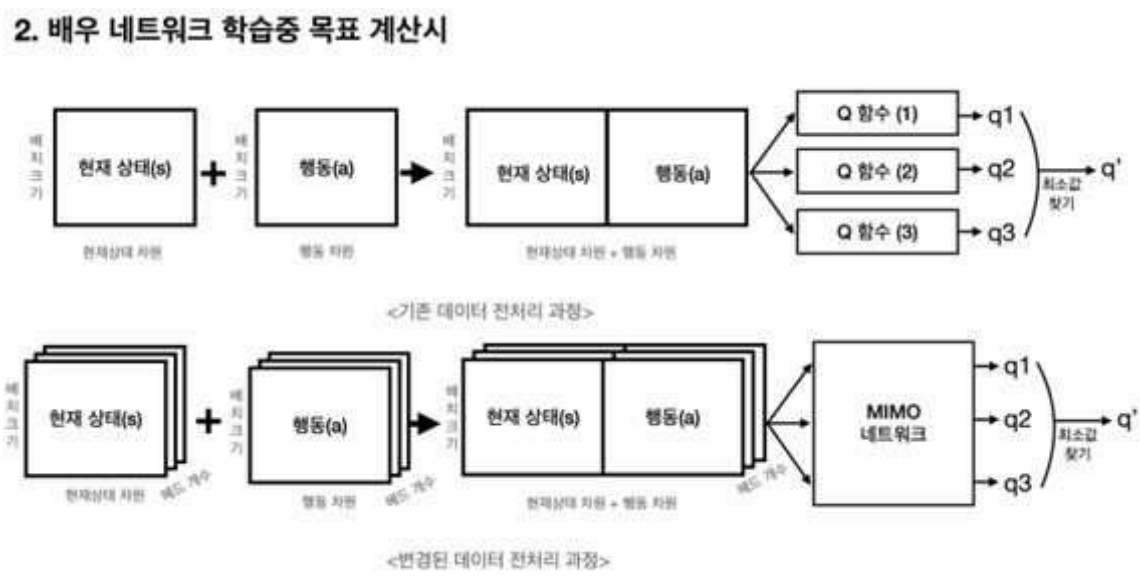
전체 청구항 수 : 총 3 항

(54) 발명의 명칭 **넓은 단일 뉴럴 네트워크로 추정한 불확실성을 활용한 오프라인 강화학습 방법론**

(57) 요 약

본 발명의 일 실시예에 의한 오프라인 강화학습 방법은, MIMO 네트워크를 활용한 오프라인 강화학습 방법에 관한 것으로, 학습 데이터셋을 전처리하는 단계 및 MIMO 네트워크를 이용하여 불확실성을 추정하도록 네트워크를 구성하고, 전처리된 학습 데이터셋을 이용하여 네트워크를 학습시키는 단계를 포함할 수 있다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	1711193973
과제번호	2019-0-00079-005
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(고려대학교)
기 여 율	50/100
과제수행기관명	고려대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193771
과제번호	2022-0-00311-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	일상적 물건들의 다접촉 로봇 조작을 위한 목적지향강화학습 기술 개발
기 여 율	50/100
과제수행기관명	한국과학기술원
연구기간	2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

MIMO 네트워크를 활용한 오프라인 강화학습 방법에 관한 것으로,

학습 데이터셋을 전처리하는 단계; 및

상기 MIMO 네트워크를 이용하여 불확실성을 추정하도록 네트워크를 구성하고, 상기 전처리된 학습 데이터셋을 이용하여 상기 네트워크를 학습시키는 단계; 를 포함하는

오프라인 강화학습 방법.

청구항 2

제1항에 있어서,

상기 학습 데이터셋을 전처리하는 단계는,

상기 학습 데이터셋에서 헤드의 개수만큼 서로 독립적으로 데이터를 추출하여 하나의 텐서로 구성하는 단계;

상기 텐서의 형태를 재성형하는 단계; 및

각 데이터를 상기 헤드의 개수개 만큼 반복하여 변환하는 단계; 를 포함하는

오프라인 강화학습 방법.

청구항 3

제2항에 있어서,

상기 하나의 텐서로 구성하는 단계는,

상기 학습 데이터의 차원이 (배치크기, 헤드의 개수, 각 항목의 차원)이 되도록 구성하고,

상기 텐서의 형태를 재성형하는 단계는,

상기 텐서의 형태를 (배치크기*헤드의 개수, 각 항목의 차원)으로 재성형하는 것을 특징으로 하는

오프라인 강화학습 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 한정된 데이터셋을 이용해 학습하는 오프라인 강화학습 알고리즘에서 넓은 단일 뉴럴 네트워크로 추정한 불확실성을 활용하여 학습하기 위한 데이터 전처리 방법과 네트워크 구성방법 및 커리큘럼에 관한 것이다.

배 경 기 술

[0002] 오프라인 강화학습은 실제 환경과의 상호작용이 불가능함에 따라 에이전트가 환경에 대한 새로운 정보를 알 수 없다. 따라서 오프라인 강화학습 에이전트는 특정 지표를 활용하여 새로운 관찰 데이터의 상태-행동 쌍(state-action pair)이 학습한 데이터 분포 내의 것인지를 파악하고 패널티를 가함으로써 보수적으로 행동하도록 학습한다. 이의 가장 대표적인 알고리즘에는 SAC-N(Soft Actor-Critic-N)이 있다.

[0003] 에이전트가 마주친 관찰 데이터의 상태-행동쌍이 학습한 데이터 분포 내에 존재하는지 파악하는 지표로 많이 사용하는 것이 불확실성(uncertainty)이다. 불확실성을 추정하기 위한 여러 가지 방법론들이 존재하지만, 보다 정확한 불확실성 추정을 위해 앙상블(ensemble) 네트워크를 활용하여 불확실성을 추정한다. 하지만, 이는 동일한

네트워크를 여러번 학습시키기 때문에 많은 컴퓨팅 재화를 필요로 한다. 모델의 규모가 큰 경우에는 정확한 불확실성을 추정할 수 있을 만큼 모델을 학습하는 것이 불가능하며 결국 퍼포먼스에 해를 끼칠 수 있다.

[0004] 근래에는 앙상블 기법보다 더 작은 컴퓨팅 재화를 필요로 하지만, 앙상블 기법에 준하는 성능을 보이는 불확실성 추정방법론들에 대한 연구가 활발히 진행되고 있다. 다중 입력 다중 출력(Multi-Input Multi-Output 이하 MIMO)도 그중 하나이다. MIMO는 서로 독립적으로 추출한 n개의 데이터를 모델의 입력으로 주고 모델에 n개의 헤드를 달아 각각의 헤드가 순차적으로 각각의 입력 데이터를 예측하도록 하는 기법이다.

[0005] 기존 알고리즘은 이런 컴퓨팅 자원 면에서 효율적인 불확실성 추정 방법론들을 사용하지 않고 앙상블 기법을 사용한다. 그러나 다루어야 할 정보가 많은 대규모 환경에서 앙상블 기법을 사용한다면 학습시키는 각 앙상블 네트워크의 규모가 한정적이거나 학습시키는 앙상블 네트워크의 수가 한정적이게 될 수 있으며 이는 정확한 불확실성을 추정할 수 없어 오히려 학습에 악영향을 끼칠 수 있다. 따라서 이러한 문제를 해결하는 것이 오프라인 강화학습 학계에서 큰 이슈가 되고 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 종래 기술의 문제점을 해결하고자 보다 계산 효율적으로 불확실성을 추정하는 방법론을 강화학습에 적용 가능하게 수정하여 효과적으로 불확실성을 추정하고 이를 최적화하는 에이전트를 학습하는 오프라인 강화학습 방법론을 제공하는 데 그 목적이 있다.

과제의 해결 수단

[0007] 본 발명의 일 실시예에 의한 오프라인 강화학습 방법은, MIMO 네트워크를 활용한 오프라인 강화학습 방법에 관한 것으로, 학습 데이터셋을 전처리하는 단계 및 MIMO 네트워크를 이용하여 불확실성을 추정하도록 네트워크를 구성하고, 전처리된 학습 데이터셋을 이용하여 네트워크를 학습시키는 단계를 포함할 수 있다.

[0008] 또한, 일 실시예에 의하면, 학습 데이터셋을 전처리하는 단계는, 학습 데이터셋에서 헤드의 개수만큼 서로 독립적으로 데이터를 추출하여 하나의 텐서로 구성하는 단계, 텐서의 형태를 재성형하는 단계 및 각 데이터를 상기 헤드의 개수개 만큼 반복하여 변환하는 단계를 포함할 수 있다.

[0009] 또한, 일 실시예에 의하면, 하나의 텐서로 구성하는 단계는, 학습 데이터의 차원이 (배치크기, 헤드의 개수, 각 항목의 차원)이 되도록 구성하고, 텐서의 형태를 재성형하는 단계는, 텐서의 형태를 (배치크기*헤드의 개수, 각 항목의 차원)으로 재성형하는 것을 특징으로 할 수 있다.

발명의 효과

[0010] 본 발명에 따른 오프라인 강화학습 방법은 주어진 컴퓨팅 자원이 한정된 환경에서 더 강건한 강화학습 에이전트 학습이 가능하다.

[0011] 또한, 본 발명에 따른 오프라인 강화학습 방법은 앙상블 기법을 사용하는 종래의 기술들에 비해 더 적은 시간으로 비슷한 성능을 낼 수 있다.

[0012] 뿐만 아니라, 본 발명에서 사용하는 단일 신경망을 이용한 불확실성 추정 방법론은 본 발명에서 제안하는 오프라인 강화학습 알고리즘뿐만 아니라 다양한 기존의 알고리즘에 적용 가능하며, 적용시 이상에서 언급한 효과와 유사한 효과를 기대할 수 있다.

도면의 간단한 설명

[0013] 도 1은 일 실시예에 따른 입력 형태 변환 과정을 도시한 도면이다.

도 2는 일 실시예에 따른 네트워크 구조를 도시한 도면이다.

도 3은 일 실시예에 따른 MIMO 네트워크를 활용한 Q함수를 사용하는 알고리즘이다.

발명을 실시하기 위한 구체적인 내용

[0014] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 서로 다른 다

양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0015] 본 명세서에서 사용되는 용어에 대해 간략히 설명하고, 본 발명에 대해 구체적으로 설명하기로 한다.

[0016] 본 발명에서 사용되는 용어는 본 발명에서의 기능을 고려하면서 가능한 현재 널리 사용되는 일반적인 용어들을 선택하였으나, 이는 당 분야에 종사하는 기술자의 의도 또는 판례, 새로운 기술의 출현 등에 따라 달라질 수 있다. 또한, 특정한 경우는 출원인이 임의로 선정한 용어도 있으며, 이 경우 해당되는 발명의 설명 부분에서 상세히 그 의미를 기재할 것이다. 따라서 본 발명에서 사용되는 용어는 단순한 용어의 명칭이 아닌, 그 용어가 가지는 의미와 본 발명의 전반에 걸친 내용을 토대로 정의되어야 한다.

[0017] 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 있음을 의미한다. 또한, 명세서에서 사용되는 "부", "모듈", "유닛" 등의 용어는 적어도 하나의 기능 또는 동작을 처리하는 단위를 의미하며, 소프트웨어, FPGA 또는 ASIC과 같은 하드웨어 구성요소, 또는 소프트웨어와 하드웨어의 결합으로 구현될 수 있다. 그렇지만 "부", "모듈", "유닛" 등의 용어가 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. "부", "모듈", "유닛" 등은 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 ""부", "모듈", "유닛" 등의 용어는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로 코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다.

[0018] 아래에서는 첨부한 도면을 참고하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략한다. “제1", "제2" 등과 같이 서수를 포함하는 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 구성 요소들은 용어들에 의해 한정되지는 않는다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. "및/또는" 이라는 용어는 복수의 관련된 항목들의 조합 또는 복수의 관련된 항목들 중의 어느 하나의 항목을 포함한다.

[0019] 본 명세서에 설명된 다양한 시스템, 장치, 저장 매체, 모듈 및 기구는 전용 연산 디바이스, 또는 하나 이상의 전용 연산 디바이스의 하나 이상의 연산 칩에서 구현될 수도 있다.

[0021] 이하, 본 발명에 따른 오프라인 강화학습 방법의 실시예들을 수학식 및 도면을 참조하여 설명한다.

[0022] 강화학습은 마르코프 결정 과정(Markov Decision Process)을 따르며, 이는 의사결정과정을 모델링하는 수학적인 기본 배경이다. 마르코프 결정 과정은 의사결정 과정에 의한 다음 상태의 확률이 오로지 전 상태에만 영향을 받는 상황을 가정한다. 강화학습 알고리즘은 이러한 마르코프 결정과정을 따르고 있다고 가정한, 학습하고자 하는 환경으로부터 관찰 정보를 입력으로 받으면 학습 목표(goal)을 달성하기 위한 행동(action)을 출력하는 정책(policy)을 학습하는 것을 주요 목적으로 하고 있다. 일반적으로, 강화학습의 학습 목표는 에이전트가 환경과의 상호작용을 통하여 얻는 보상(reward)의 할인된 총합(discounted sum)을 최대화하는 것을 뜻한다.

[0023] 본 발명이 다루고 있는 오프라인 강화학습은 보상(reward)의 할인된 총합 최대화를 환경과의 추가 상호작용 없이 사전에 확보되어 있는 한정된 수량의 데이터셋만을 사용하여 이루는 것을 목적으로 한다. 이를 위해 본 발명은 데이터셋의 상태에서 특정 행동을 실행하였을 때 미래에 예상되는 할인된 총합을 예측하는 네트워크인 행동-가치 함수(action value function)에 해당하는 Q 함수(Q function)를 학습한다. 이러한 Q함수를 사용하여, 배우 네트워크(actor network)가 데이터셋의 상태-행동 쌍(state-action pair)중 미래 가치가 높은 행동을 취한 데이터(미래의 행동-가치 함수의 값이 높은 데이터, Q값이 큰 데이터)에 대해서 높은 확률을 부여할 수 있도록 해준다.

[0024] 오프라인 강화학습에서 Q함수는 한정된 데이터셋을 이용하여 학습하기 때문에 데이터셋이 다루고 있는 상태-행동 분포가 넓지 않을 경우, 학습된 Q 함수가 데이터셋 분포 밖 샘플(Out of distribution)에 대해서 예측하는 가치가 실제의 가치를 모사하지 못할 가능성이 높아진다. 따라서, 오프라인 강화학습 연구자들은 Q-함수가 보수

적(conservative)으로 학습되도록 하는 방법을 연구했었다. 여기서 보수적인 Q 함수는 데이터셋 분포내의 상태-행동에 대한 가치를 높게 보는 Q함수를 의미한다. 상기한 Q-함수의 보수성(conservatism)을 올리기 위해, 기존의 오프라인 강화학습 알고리즘은 학습을 위해 Q-함수에 앙상블 기법을 적용하고 각 네트워크에서 출력으로 나오는 미래 가치 값(Q값)중 가장 낮은 값을 활용한 정규화(regularization) 기법을 적용해 왔었다.

[0025] 앙상블 기법은 동일한 구조의 네트워크를 처음 네트워크 파라미터의 초기화만을 다르게 하여 학습시킨 후, 이후 동일한 데이터를 각 앙상블 네트워크에 입력으로 주고 나오는 각 출력 값들 간의 분산 혹은 표준편차를 계산하여 불확실성을 추정하는 방법이다. 즉, 분산(혹은 표준편차)는 편차제곱 합의 기댓값으므로 각 네트워크들의 출력 값들 간의 편차가 클수록 불확실성이 높다고 볼 수 있다. 상기의 방법론에서 앙상블 네트워크들 간의 불확실성(각 네트워크의 출력 값들 간의 분산 혹은 표준편차)를 직접적으로 사용하지 않았으나, 가장 작은 출력 값을 사용함으로써 이를 간접적으로 사용하고 있다.

[0026] 앙상블 기법을 이용하면 보다 정확한 불확실성을 추정하는 것이 가능하다고 이전 연구들에서 알려졌지만, 이는 컴퓨팅 재화가 한정되지 않았을 때를 가정한다. 컴퓨팅 재화가 한정됨에 따라 학습시키고자 하는 모델(상기 방법론에서는 Q함수를 지칭)의 규모 혹은 그 개수를 학습시키는데 제약이 생기게 되며 이는 결국 강화학습 에이전트의 퍼포먼스에 해를 끼칠 수 있다. 오프라인 강화학습 방법론을 적용하는 대부분의 실제 상황에서는 컴퓨팅 재화가 한정되어 있으며 학습시키고자 하는 모델(상기 방법론에서는 Q함수를 지칭)의 규모가 큰 경우에 더 치명적인 문제로 작용한다. 따라서, 기존에 사용하던 앙상블 기법대신 강화학습 에이전트 학습에서도 잘 작동할 수 있는 컴퓨팅 재화면에서 효과적인 불확실성 추정방법에 대한 연구가 필요한 상황이다.

[0027] 본 발명은 SAC-N 알고리즘에서 오프라인 강화학습 에이전트가 앙상블 기법을 사용한 것과는 다르게 MIMO네트워크를 활용하여 보다 효과적으로 불확실성을 추정하도록 네트워크를 구성하고 학습시키기 위한 데이터 전처리 과정 및 네트워크 구성하고 학습시키는 알고리즘을 개발하였고 이는 도 1~3에 표현되어 있다.

[0028] 본 발명은 학습에 사용한 MIMO네트워크는 복권 가설(lottery hypothesis)에 입각해 네트워크를 구상하였다. 복권 가설은 일반적으로 심층 뉴럴 네트워크가 파라미터적으로 과적합된 모델이라고 보고 무작위로 초기화된 밀집된 심층 뉴럴 네트워크에서 특정 부분을 가져와 따로 학습하더라도, 기존의 심층 뉴럴 네트워크와 동일한 성능을 같은 학습 횟수 내에 달성할 수 있다는 것이다. 이때, 심층 뉴럴 네트워크의 특정 부분은 기존 심층 뉴럴 네트워크의 최대 20%까지만을 사용해도 기존의 네트워크와 동일한 성능을 가질 수 있다.

[0029] 이러한 가설에 의하면 일반적인 심층 뉴럴 네트워크는 하나의 심층 뉴럴 네트워크는 3~4개의 서로 독립적인 하위 네트워크들을 동시에 학습시킬 수 있는 역량이 충분하다는 것을 의미한다. 따라서 MIMO는 단일 심층 네트워크에 여러 입력을 주고 각 입력에 대한 출력을 네트워크가 내뱉도록 학습시킴으로써 서로 독립적인 하위 네트워크들을 동시에 학습시킨다. 이후 모델을 평가하거나 특정 입력에 대한 불확실성을 추정하기 위해서는 동일한 데이터를 반복하여 입력으로 사용하고 모델의 각 하위 네트워크가 예측한 출력값들을 바탕으로 모델의 성능을 평가하거나 불확실성을 추정한다.

[0030] MIMO가 기존의 다른 효과적인 불확실성 추정 방법론들과의 차별화된 점은 불학실성을 추정하는데 있어 기존의 다른 방법들은 4~10개의 추가 단계(forward pass)들이 필요한 반면에 MIMO는 추가 단계 없이 가능하다는 점이다. 또한, 기존의 MIMO 네트워크는 이미지 분류 작업(image classification task)에서만 사용하였으나, 본 논문은 MIMO 네트워크를 강화학습에도 적용가능하도록 하였다. 이를 위해 데이터 전처리 과정과 MIMO 네트워크 구성 방법을 제안한다.

[0031] 오프라인 강화학습에서 MIMO 네트워크를 학습하기 위한 데이터 셋을 구성할 때 학습 데이터 셋에서 헤드의 개수(기존 앙상블 기법에서 학습시키는 앙상블 네트워크의 개수와 동일)만큼 서로 독립적으로 데이터를 추출(sampling)하여 하나의 텐서로 구성하여(학습 데이터의 차원이 (배치크기, 헤드의 개수, 각 항목의 차원)이 되도록 구성) 모델이 각 독립적인 데이터를 동시에 입력으로 받을 수 있도록 한다. 이후 불확실성을 추정하기 위해서 해당 텐서의 형태를 (배치크기*헤드의 개수, 각 항목의 차원)으로 재성형한 후 각 데이터를 헤드의 개수개만큼 반복하여 (배치크기*헤드의 개수, 헤드의개수, 각 항목의 차원)으로 변환한다. 주어진 데이터를 이렇게 변환함으로써 오프라인 강화학습 에이전트는 행동(action)을 결정하는데 있어 동일한 입력에 대한 앙상블 네트워크의 효과를 기대할 수 있으며 그중 가장 작은 값을 선택하여 보수적으로 행동한다. 입력 데이터의 형태 변환 과정은 도 1에 표현되어 있다.

[0032] 오프라인 강화학습의 Q 함수에 MIMIO 네트워크를 적용하기 위해 추가로 필요한 매개변수에는 네트워크의 폭을 조정해주는 상수 값(β)와 헤드의 개수(N)이 있다. 네트워크의 폭을 조정해주는 상수 값(β)은 네트워크의 폭을

기존 네트워크보다 늘림으로써 단일 네트워크 안에 더 많은 하위 네트워크들을 학습하기 위함이다. 또한, 기존 네트워크와는 달리 각 입력값들이 서로 독립적으로 네트워크에 작용하도록 하기 위한 인코더가 존재한다. 독립적으로 추출된 각 데이터들은 각각의 인코더를 통과하여 해당 데이터에 대한 특징을 뽑은 후 이를 하나의 벡터로 합쳐 네트워크의 인풋으로 넣어준 후 각 데이터에 해당하는 특징들에 대해 그에 대응되는 각각의 헤드들이 예측값을 내뱉도록 하는 구조이다. Q 함수 앙상블을 위한 MIMO 네트워크 구조는 도 2에 표현되어 있다. 이러한 MIMO 네트워크를 학습시키기 위해 기존의 Q 함수의 손실 함수와 동일한 함수를 사용하며 이는 [수학식 1]에 나타나 있다.

수학식 1

[0033] $$Loss = E_{s,\,a,\,s' \sim D}\left[M(s,a) - \left(r(s,a) + \gamma\, E_{a' \sim \pi(\cdot|s')}[\min M(s',a')] \right) \right]$$

[0034] 위에서 언급한 손실함수를 이용하여 학습한 Q 함수를 통해 에이전트의 정책(policy)을 학습하며 정책의 손실함수는 기존 알고리즘인 SAC-N를 벤치마크한 [수학식 2]를 따른다. 에이전트의 정책의 경우 앙상블 기법을 사용하지 않았다.

수학식 2

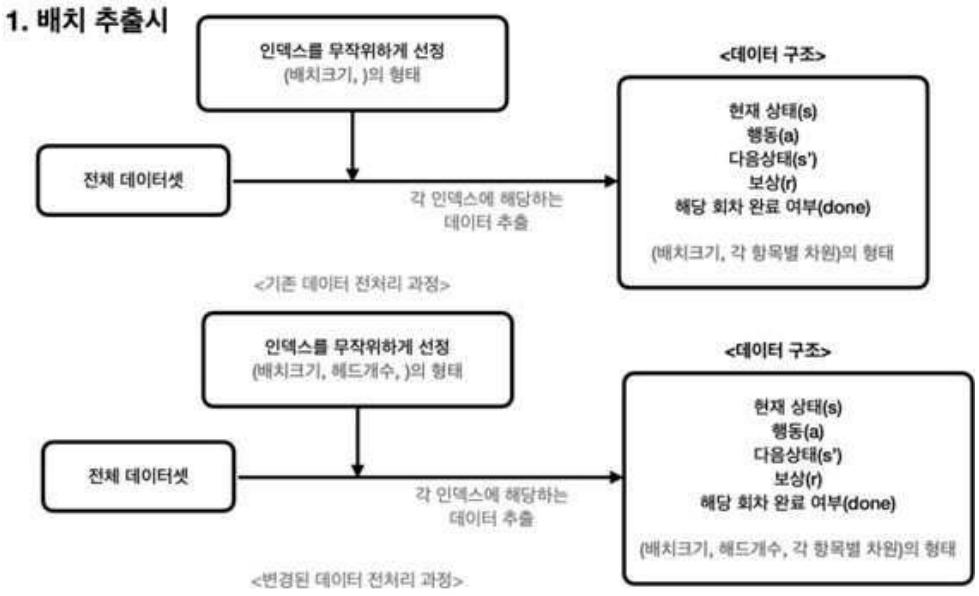
[0035] $$Loss = E_{a \sim \pi(\cdot|s),\, s \sim D}\left[\alpha \log(a \mid s) - M(s,a) \right]$$

[0036] 도 3은 앞에서 제시한 MIMO 네트워크를 활용한 Q 함수를 사용하는 알고리즘을 표현하고 있다. 매 iteration마다 상기 언급한 손실함수를 계산하기 위해 도 1과 이전에 언급한 데이터 전처리 과정을 거쳐 배치를 추출을 한 후 해당 배치 샘플들에 대한 손실함수의 평균을 구하여 각 네트워크를 업데이트 한다.

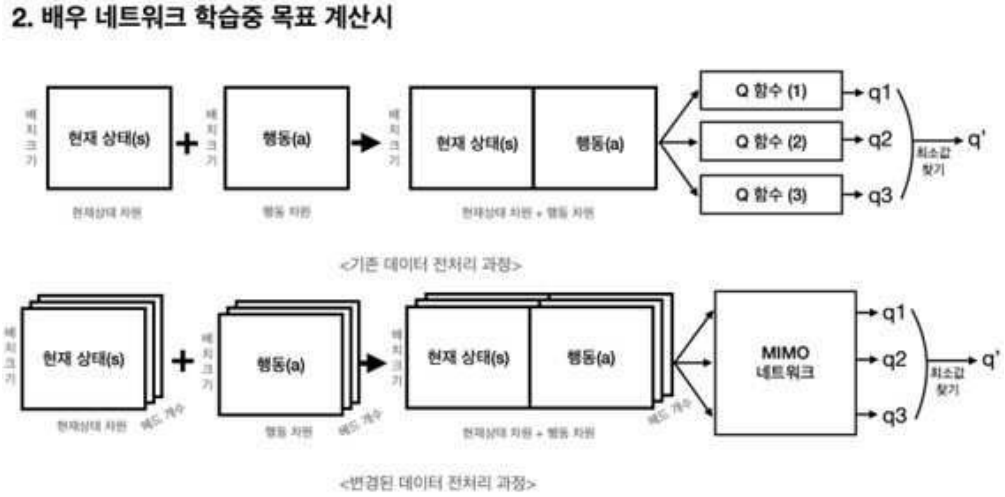
[0038] 본원 발명의 실시예 들과 관련된 기술 분야에서 통상의 지식을 가진 자는 상기 기재의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로, 개시된 방법들은 한정적인 관점이 아닌 설명적 관점에서 고려되어야 한다. 본 발명의 범위는 발명의 상세한 설명이 아닌 특허청구 범위에 나타나며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면

도면1



도면2



도면3

Algorithm 1 학습알고리즘

Require: 데이터셋 $\mathcal{D}$, 배치크기 B, MIMO-Q함수 M, 헤드개수 H, 정책네트워크 π_θ, 총 이터레이션 수 t

모든 네트워크 초기화
for $i = 1$ to t **do**
  B, H 의 형태 및 크기로 데이터셋 $\mathcal{D}$에서 데이터 샘플링
  for 데이터 샘플링 **do**
    MIMO-Q함수 M를 [수학식 2]의 손실함수를 이용하여 업데이트
    정책네트워크 π_θ를 [수학식 1]의 손실함수를 이용하여 업데이트
  end for
end for

return M, π_θ