2-stage 마르코프 의사결정 상황에서 Successor Representation 기반 강화학습 알고리즘 성능 평가

김소현*, 이지항*,+
*상명대학교 지능정보공학과
202131054@smu.ac.kr, jeehang@smu.ac.kr

Evaluating a successor representation-based reinforcement learning algorithm in the 2-stage Markov decision task

So-Hyeon Kim*, Jee Hang Lee*,+
*Department of AI & Informatics, Sangmyung University, Seoul, South Korea
+Corresponding author: Jee Hang Lee

요 약

Successor representation (SR) 은 두뇌 내 해마의 공간 세포가 인지맵을 구성하여 환경을 학습하고, 이를 활용하여 변화하는 환경에서 유연하게 최적 전략을 수립하는 기전을 모사한 강화학습 방법이다. 특히, 학습한 환경 정보를 활용, 환경 구조 안에서 목표가 변화할 때 강인하게 대응하여 일반 model-free 강화학습에 비해 빠르게 보상 변화에 적응하고 최적 전략을 찾는 것으로 알려져 있다. 본 논문에서는 SR 기반 강화학습 알고리즘이 보상의 변화와 더불어 환경 구조, 특히 환경의 상태 천이 확률이 변화하여 보상의 변화를 유발하는 상황에서 어떠한 성능을 보이는 지 확인하였다. 벤치마크 알고리즘으로 SR 의 특성을 목적 기반 강화학습으로 통합한 SR-Dyna 를 사용하였고, 환경 상태 천이 불확실성과 보상 변화가 동시에 나타나는 2-stage 마르코프 의사결정 과제를 실험 환경으로 사용하였다. 시뮬레이션 결과, SR-Dyna 는 환경 내 상태 천이 확률 변화에 따른 보상 변화에는 적절히 대응하지 못하는 결과를 보였다. 본 결과를 통해 두뇌의 강화학습과 알고리즘 강화학습의 차이를 이해하여, 환경 변화에 강인한 강화학습 알고리즘 설계를 기대할 수 있다.

1. 서론

강화학습은 에이전트가 환경과 상호작용하며 획득한 보상을 바탕으로 최적 전략을 학습하는 학습 기전으로, 최근 많은 분야에서 뛰어난 성취를 보이고 있다 [1-3]. 그럼에도 불구하고, 기계의 강화학습은 인간의 강화학습에 비해 유연성과 적응성이 크게 부족한 면을 보이고 있다. 최신 신경과학적 발견에 따르면 인간의 강화학습은 model-based 와 model-free 강화학습 전략을 전두엽 메타 제어를 통해 유연하게 활용하여 환경 변화에 기민하게 대응하고 빠르게 적응하여 높은 성능을 보인다고 알려져 있다 [4,5].

Model-based 강화학습은 환경에 대해 학습하고, 이 정보를 활용하여 계획을 수립한 후 문제를 해결하는 방식이다. 환경을 학습하는 과정을 수반하므로, 환경 변화에 기민하게 대응할 수 있고, 환경에 대해 미리 알고 있기 때문에 더 정확한 예측이 가능하지만 계산량이 많고 복잡하다는 특징이 있다. 반면 model-free 강화학습은 환경에 대한 정보를 학습하는 과정 없이, 행동에 대한 보상 신호만을 통해 문제 해결을 위한 최적 전략을 수립한다. 환경 정보를 고려하지 않기 때문에 한번 학습한 전략은 계획 과정 없이 빠르게 실행이 가능하나, 환경이 변한 경우 새로운 환경에서 최적 전략을 수립하기 위해 학습을 위한 경험이 많이 필요하며, 그만큼 학습 시간도 많이 소요되어 환경 변화에 기민하게 대처하기 어렵다[4].

두 알고리즘이 가진 특성이 다르기 때문에 기존 강화학습 알고리즘 연구는 model-free 강화학습 위주로 진행되면서, model-based 강화학습을 차용하는 접근을 취하였다. 그러나 최근 신경과학적 발견을 통해 인간과 기계의 강화학습 방법의 차이가 조명되면서, 뇌과학적 발견을 적극적으로 반영한 메타강화학습 알고리즘[6]이나, 강화학습과 연관된 두뇌의 정보처리 기전을 모사한 neuroscience-inspired 알고리즘들이 보다 활

발하게 연구되고 있다[7].

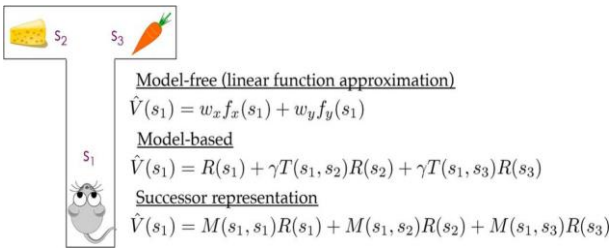
본 논문에서는 최신 신경과학적 발견을 기반으로 구현된 Successor representation (SR) 기반 강화학습 알고리즘을 이용하여 환경 변화에 따른 강화학습 에이전트의 성능을 평가하였다. SR 은 두뇌 내 해마의 공간 세포가 인지맵을 구성하여 환경을 학습하고, 이를 활용하여 변화하는 환경에서 유연하게 최적 전략을 수립하는 기전을 모사한 강화학습 방법이다. 특히, 학습한 환경 정보를 활용, 환경 구조 안에서 목표가 변화할 때 강인하게 대응하여 일반 model-free 강화학습에 비해 빠르게 보상 변화에 적응하고 최적 전략을 찾는 것으로 잘 알려져 있다. 다만, 환경 구조 자체에 대한 변화나 환경 내 상태 천이 불확실성에 대응하여 전략을 수립하는 연구는 상대적으로 부족한 상황이다.

따라서, 본 논문은 환경의 상태 천이 확률이 변화하고, 이와 연계되어 보상 함수 또한 변화하는 상황에서 SR 기반 강화학습 에이전트의 성능을 평가하고, 향후 연구 방향을 도출하고자 한다. 이를 위해, 벤치마크 알고리즘으로 SR 의 특성을 목적 기반 강화학습으로 통합한 SR-Dyna 를 사용하였고, 실험 환경으로 Lee *et al.*이 제안한 2-stage 마르코프 의사결정 과제(이후 2-stage MDT)를 선택하였다[5]. 본 환경은 보상 함수 뿐만 아니라 환경 불확실성을 결정하는 상태 천이 확률 및 목적이 변화하는 환경으로, 보상 함수 변화에 강인한 SR 기반 강화학습 알고리즘이 환경 구조가 변화하는 환경에서 성능에 차이를 보이는지를 확인하는 데 좋은 벤치마크 환경이 될 것으로 보인다.

2. SR 기반 강화학습 모델: SR-Dyna

SR 의 주요 아이디어는 다음과 같다. 해마의 공간 세포가 지니고 있는 정보는 에이전트가 경험한 환경의 공간적 정보인데, 이를 미래 상태를 예측하는 데 사용이 가능하다는 가정에서 출발한다[8]. 미래 상태를 예측하기 위한 정보는 현재 상태에서 액션을 수행하여 (i) 다음 상태로 이동했을 때 획득 가능한 보상의 예측 값과 (ii) 점유가 예측되는 다음 상태 값을 바탕으로 학습한다[9]. 부분적으로 계산된 행동 가치 값과 환경 정보를 모두 고려하여 최적 정책을 수립하므로 model-based 와 model-free 를 모두 고려하여 근사화한 중간계열의 알고리즘이라고 할 수 있다[10]. 그림 1 은 model-free 와 model-based, 그리고 SR 의 가치 업데이트 기전을 통해 SR 의 특성을 보여준다.

동일한 환경 구조에서 보상 함수가 변화하는 grid world 환경에서 SR 의 특성을 잘 확인할 수 있다[12]. 에이전트가 미로를 탐색할 때 설치류의 여러 기본 능력과 유사한 형태로 동작하는데, 이 능력은 인지맵에 기인한다고 볼 수 있다. 인지맵은 해마에서 인코딩되며 이는 latent learning task 와 shortcut, detour task 에 모두 빠르게 적응할 수 있게 도와준다. 그 중에서도 model-based 강화학습의 한 종류로써 experience replay 를 하는 dyna[11]와 결합한 SR (이후 SR-Dyna)는 이 능력을 모두 보여준다[12].

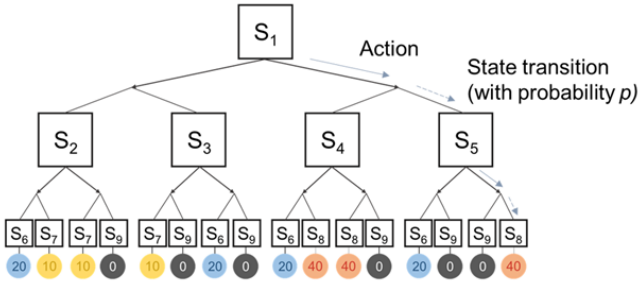


(그림 1) 간단한 T-maze 미로에서 model-free, model-based 및 SR 기반 강화학습 알고리즘들이 value 를 계산하는 방식(© Gershman, 2018 [9])

이를 고려하여, 본 논문에서는 보상 함수가 변화하는 환경에서도 구조를 잘 배우며 model-based 와 model-free 의 행동이 분리되는 시나리오에서도 성능을 보일 수 있을 것이라고 기대되는 SR-Dyna 를 본 연구의 모델로 삼았다.

3. 실험 환경

보상 함수 변화에 강인한 SR-Dyna 알고리즘이 환경 구조가 변화하는 환경에서 성능에 차이를 보이는지를 확인하기 위해 2-stage MDT [5] (그림 2)에서 시뮬레이션을 수행하였다.



(그림 2) 2-stage 마르코프 의사결정 환경[5]

신경과학에서 사용한 2-Stage 마르코프 의사결정 환경은 초기상태 S_1 에서 에이전트가 두 번의 Left/Right 이동 액션을 통해 최종 상태로 이동하고, 최종 상태 (S_6-S_9)와 연계된 동전에 기입된 보상을 받는 환경이다. goal condition 에 따라 상태 천이 확률이 조정되는데, specific goal condition 의 경우 에이전트가 Left/Right 선택 시 상태 천이는 (0.9, 0.1)의 확률로 결정된다. 다시 말해, 에이전트가 Left 를 선택했을 경우, 선택한 행동은 90%의 확률로 수행되고, 10%의 확률로 반대로 수행됨을 뜻한다. Specific goal condition 에서는 수집할 동전의 색깔이 제시된다. 이 때, 에이전트는 반드시 제시된 색깔의 동전을 수집해야 동전에 기록

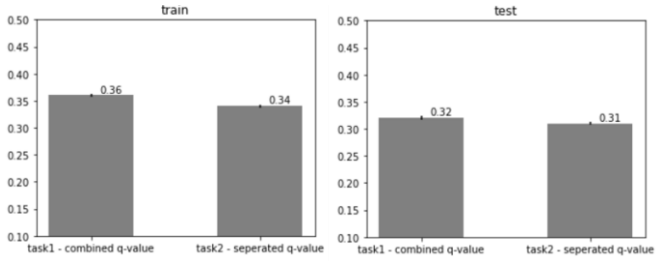
된 보상을 받을 수 있다. 만일 색깔이 다른 동전을 수집하면 동전에 기록된 보상은 주어지지 않고 보상으로 0 을 받게 된다. 반면, flexible goal condition 에서는 어떤 색깔의 동전을 수집하더라도 동전에 기록된 보상을 받을 수 있다. 다만, 상태 천이 확률이 (0.5, 0.5)로 설정되어, Left/Right 선택은 에이전트의 선택을 따르지 않고 무작위로 선택된다.

일반적으로 specific goal condition 은 환경 불확실성이 낮아 환경 구조를 배우기 용이하고, 상태 예측이 가능한 상황이다. 따라서, 이 상황에서는 model-based 강화학습 전략이 선호된다. 반면, flexible goal condition 은 환경 불확실성이 매우 높아 환경 구조를 배우기 어렵고, 따라서 상태 예측이 매우 어렵다. 따라서, 이 상황에서는 행동에 따른 보상을 추구하는 방식, 즉 model-free 강화학습 전략이 더 적합한 학습 전략이 될 수 있다. 이렇게 환경 변수에 따라 환경 구조가 변화하고 보상 값이 변하는 환경을 통해 SR 기반 강화학습 알고리즘의 성능 평가를 시도하였다.

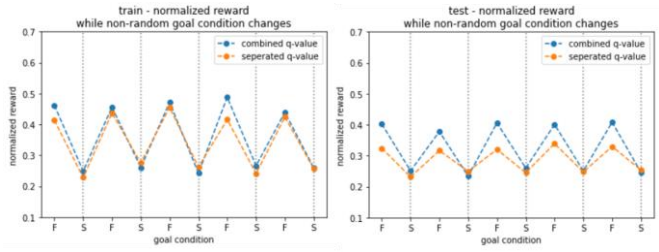
4. 실험 및 결과

본 논문의 SR-Dyna 가 상태 천이 확률 변화와 보상 함수 변화가 동시에 일어나는 환경을 학습할 수 있는지를 확인하고자 한다. 이를 위해, 두 가지 버전의 SR-Dyna 를 구성하고, 각각을 실험군/대조군으로 사용하였다. 실험은 SR 기반 강화학습 에이전트가 2-stage MDT 를 총 500 게임 수행하도록 하였다. 각 goal condition 은 매 50 game 마다 바뀌도록 설정하였다. 즉 500 game 을 수행하는 동안 10 번의 goal condition 변화가 발생한다 (이를 1 session 이라 정의한다). 에이전트가 획득한 보상은 0-1 사이의 값으로 normalize 하였다.

기존 SR-Dyna 알고리즘은 tabular 방식으로 하나의 SR-matrix 에서 상태 가치 값이 계산되고, 보상이 주어졌을 때 이 SR-matrix 를 기반으로 전략을 수립한다. 이를 실험군으로 설정하고, 2-stage MDT 를 학습하도록 하였다. 대조군으로는 2-stage MDT 를 수행할 때, 각 goal condition 별로 SR-matrix 를 학습할 수 있는 에이전트를 구현하였다. 이 에이전트는 모든 환경 변화를 관찰할 수 있고, 그에 맞추어 가장 최적화된 결정을 수행하는 이상적 에이전트라고 가정하였다. 따라서, 2-stage MDT 에서 goal condition 에 따라 보상의 지급 방식과 상태 천이 확률이 다른 것을 모두 알고 이에 최적으로 대응하고 학습할 수 있도록 SR-matrix 를 두 개(specific matrix/flexible matrix)로 분리한 에이전트를 사용하였다. 마지막으로 2-stage MDT 에 최적화되어 학습된 forward-backward 알고리즘 기반 model-based 강화학습 에이전트[5]를 이용하여 최적 보상 값을 구하고, 이 값을 실험군/대조군 성능 평가를 위한 upper bound 로 사용하였다.



(그림 3) average normalized reward



(그림 4) average normalized reward while goal condition changes

그림 3 은 단일 matrix 기준 1 session 을 30 번 수행(=500 game * 30 session = 15000 games)한 것에 대한 정규화된 평균 보상 값이다. 그림 4 는 goal condition 별로 실험군/대조군의 보상 차이를 보여준다.

우선 그림 3 에서 보다시피 실험군(단일 SR-matrix)와 대조군 (이중 SR-matrix)의 평균 보상에 큰 차이를 보이지 않았다. 특히 본 논문의 주요 포인트는 specific goal condition 에서의 두 matrix 비교였는데, 그림 4 와 같이 specific goal condition 인 경우 두 matrix 간 차이는 거의 나지 않았으며, 두 matrix 모두 flexible goal condition 에 비해 리턴이 매우 낮은 것을 볼 수 있었다.

2-stage MDT 에서 chance level 보상 값은 0.35 로 분석되는데[5], 두 matrix 모두 기준 값 보다는 평균적으로 높은 보상을 보였다. 그러나 upper bound 에 해당하는 model-based 강화학습 알고리즘의 보상 0.63 보다 낮은 수치를 보이는 것을 고려했을 때, SR-Dyna 는 상태 천이 확률이 변화하여 보상 함수가 변화하는 환경은 효과적으로 학습하지 못한 것으로 보인다.

5. 결론

본 연구에서는 model-based 와 model-free 강화학습을 고루 근사화 한 SR-기반 강화학습 에이전트가 model-based 와 model-free 를 분리하는 환경에서 어떠한 성능을 내는지 보고자 하였다. 성능 평가를 위해 환경 변화에 따른 보상 변화에 강인한 SR-Dyna 모델을 사용하였다. 실험결과 SR-Dyna 는 환경의 상태 천이 확률을 학습하지 못하였으며, 이에 따른 보상 변화도 효과적으로 학습하지 못하는 것을 확인하였다. 이는

goal condition 에 따라 matrix 를 분리하여 학습한 이중 matrix 의 경우에서도 동일하였다. 추후 연구에서는 이를 극복하기 위해 SR-기반 강화학습 알고리즘을 개선하여 환경 상태 천이 확률 변화 및 복잡한 환경에도 강인한 성능을 낼 수 있도록 하고자 한다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2020R1G1A1102683). 본 연구는 삼성미래기술육성센터의 지원을 받아 수행하였음 (No. SRFC-TC1603-52). 본 결과물은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 사회맞춤형 산학협력 선도대학 (LINC+) 육성사업의 연구결과임.

참고문헌

[1] SILVER, David, et al. Mastering the game of Go with deep neural networks and tree search. nature, 2016, 529.7587: 484-489.

[2] SILVER, David, et al. Mastering the game of go without human knowledge. nature, 2017, 550.7676: 354-359.

[3] SCHRITTWIESER, Julian, et al. Mastering atari, go, chess and shogi by planning with a learned model. Nature, 2020, 588.7839: 604-609.

[4] LEE, Jee Hang, et al. Toward high-performance, memory-efficient, and fast reinforcement learning-Lessons from decision neuroscience. 2019.

[5] LEE, Sang Wan; SHIMOJO, Shinsuke; O'DOHERTY, John P. Neural computations underlying arbitration between model-based and model-free learning. Neuron, 2014, 81.3: 687-699.

[6] WANG, Jane X., et al. Prefrontal cortex as a meta-reinforcement learning system. Nature neuroscience, 2018, 21.6: 860-868.

[7] HASSABIS, Demis, et al. Neuroscience-inspired artificial intelligence. Neuron, 2017, 95.2: 245-258.

[8] STACHENFELD, Kimberly L.; BOTVINICK, Matthew M.; GERSHMAN, Samuel J. The hippocampus as a predictive map. Nature neuroscience, 2017, 20.11: 1643-1653.

[9] GERSHMAN, Samuel J. The successor representation: its computational logic and neural substrates. Journal of Neuroscence, 2018, 38.33: 7193-7200.

[10] MOMENNEJAD, Ida, et al. The successor representation in human reinforcement learning. Nature human behaviour, 2017, 1.9: 680-692.

[11] SUTTON, Richard S. Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bulletin, 1991, 2.4: 160-163.

[12] RUSSEK, Evan M., et al. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS computational biology, 2017, 13.9: e1005768.