심층강화학습에 은닉 상태 정보 활용을 통한 학습 성능 개선에 대한 고찰

최요한[1], 석영준[2], 김주봉[1], 한연희[1†]
[1]한국기술교육대학교 컴퓨터공학과 미래융합공학전공
[2]한국기술교육대학교 컴퓨터공학과 컴퓨터공학전공
yoweif@koreatech.ac.kr, dsb04163@koreatech.ac.kr,
rlawnqhd@koreatech.ac.kr, yhhan@koreatech.ac.kr,

A Study on Learning Performance Improvement by Using Hidden States in Deep Reinforcement Learning

Yohan Choi[1], Yeong-Jun Seok[2], Ju-Bong Kim[1], Youn-Hee Han[1†]
[1]Future Convergence Engineering, Dept. of Computer Science Engineering, KOREATECH
[2]Dept. of Computer Science Engineering, KOREATECH

요 약

심층강화학습에 완전 연결 신경망과 합성곱 신경망은 잘 활용되는 것에 반해 순환 신경망은 잘 활용되지 않는다. 이는 강화학습이 마르코프 속성을 전제로 하기 때문이다. 지금까지의 강화학습은 환경이 마르코프 속성을 만족하도록 사전 작업이 필요했다, 본 논문에서는 마르코프 속성을 따르지 않는 환경에서 이러한 사전 작업 없이도 순환 신경망의 은닉 상태를 통해 마르코프 속성을 학습함으로써 학습 성능을 개선할 수 있다는 것을 소개한다.

1. 서론

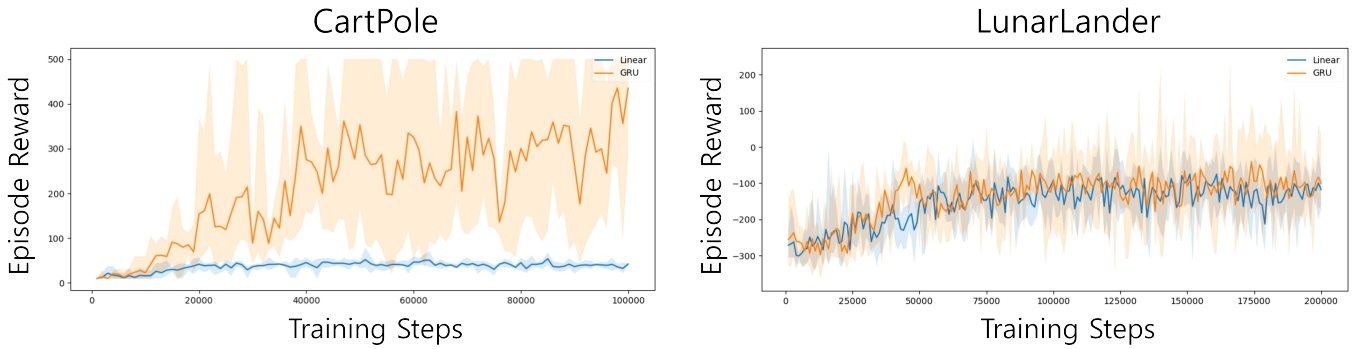
강화학습(Reinforcement Learning)에 심층학습(Deep Learning)이 접목 가능하다는 사실이 알려진 이후 심층강화학습(Deep Reinforcement Learning)이 크게 발전하였고, 심층학습이 발전함에 따라 심층강화학습도 함께 발전하였다 [1]. 이를 통해 게임, 바둑, 로봇제어 등 다양한 문제를 해결했다. 하지만 심층강화학습에서 합성곱 신경망(Convolutional Neural Network)과 완전 연결 신경망(Fully Connected Network)만 주로 사용할 뿐, 순환 신경망(Recurrent Neural Network) 계열들은 자주 사용되지 않았다. 심층강화학습에 순환 신경망을 사용하는 연구들이 있었지만, 한정된 분야에서만 사용하거나 성능이 나아지는 이유를 제대로 설명하지 못하였다 [2, 3]. 본 논문에선 순환 신경망이 강화학습에서 마르코프 속성을 따르는 데에 도움이 된다는 것을 소개한다.

강화학습은 주어진 문제에 대해서 연속된 행동을 제어하여 문제를 해결하므로 시간의 흐름과 밀접한 관련이 있다. 하지만 실제 연구에선 시간의 흐름을 고려하지 않는 경우가 많다. 강화학습은 문제의 환경이 마르코프 속성을 따른다고 가정하기 때문이다. 타입스텝 t일 때 상태 X_t가 주어질 때, 상태가 X_{t+1}로 전이될 확률을 $P(X_{t+1}|X_t)$라고 한다면 마르코프 속성은 다음과 같다.

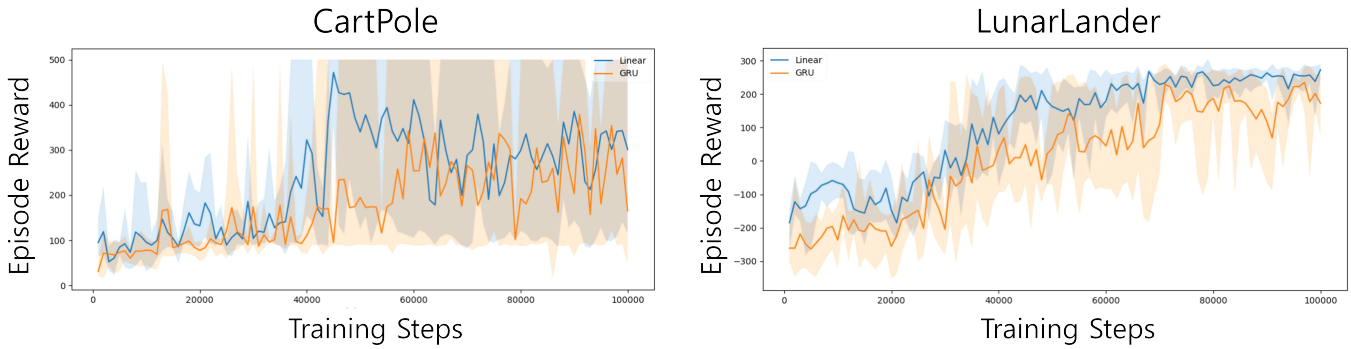
$$P(X_{t+1}|X_t) = P(X_{t+1}|X_t, X_{t-1}, \dots, X_0) \quad (1)$$

과거의 상태 정보들은 의미 없고, 오직 현재의 상태 정보로만 다음 스텝의 상태가 결정된다는 것이다. 실제로도 강화학습에서 환경 모델을 만들 때 이러한 속성을 따르도록 설계한다. 이러한 속성은 다른 의미로도 해석될 수 있다. 현재의 정보가 미래가 결정되는 데에 필요한 과거의 정보를 내포하고 있다는 의미이다. Atari Games에서는 본래의 상태 정보는 현재 타입 스텝의 게임 화면이지만 이 정보만으로는 현재 게임 내의 각각의 객체나 물체들이 어느 방향으로 움직이고 있는지 알 수 없다. 따라서 강화학습에서 Atari Games 환경을 모델링할 때, 해당 모델이 마르코프 속성을 따른다고 가정하기 힘들다. Mnih et al. [1]에서 이를 해결하기 위해 현재 타임 스텝부터 가장 최신의 4개의 게임 화면을 현재의 상태 정보로 활용한다. 이렇듯 강화학습에서 환경을 활용하기

† 한연희: 교신저자



(그림 1) 속도를 제거한 환경에서 GRU 비교 실험 결과



(그림 2) 본래 환경에서의 GRU 비교 실험 결과

위해 마르코프 속성을 따르도록 사전 작업이 필요하다. 본 논문에선 순환 신경망이 마르코프 속성을 따르지 않는 환경에서 도움이 된다는 것을 소개한다.

2. 순환 신경망을 활용한 강화학습 제안

앞서 보았듯 강화학습은 문제 특성상 시간의 흐름과 밀접한 관련이 있지만, 마르코프 속성을 만족하게 함으로써 시간과의 연관성을 분리했다. 이를 위해 최신 4개의 정보를 하나의 정보로 취급하는 등 별도의 작업이 필요했다. 또는 이러한 사전 작업이 수행된 환경이라 할지라도 마르코프 속성을 따른다고 보장하기 힘들다. 그리고 실제 환경도 마르코프 속성을 보장하기 힘들다. 본 논문에선 이를 해결하기 위해 순환 신경망을 사용한다. 순환 신경망은 완전 연결 신경망과 유사한 구조로 되어 있지만, 신경망의 출력이 다시 자신의 입력으로 들어온다는 점에서 차이가 있다. 이러한 구조 때문에 신경망이 메모리 역할을 하게 되고 신경망에 입력이 들어오면 과거에 학습한 내용을 현재의 입력과 연계하여 학습하는 효과를 가진다. 이러한 특징 때문에 순환 신경망은 시계열 데이터를 다루는 기계학습에 자주 사용된다.

사람의 경우 마르코프 속성이 보장되지 않은 현실 세계에 살지만 당장 다음 상태를 예측하는 것에 어려움이 없다. 예를 들어 공이 공중에 떠 있다면, 사람은 당장 공의 위치 정보만을 통해 공의 다음 상태를 예측하는 것이 아닌 기억 속에서 공의 이전 위치와 현재 위치를 함께 고려하여 공의 속력과 방향을 생각하고 공의 다음 상태를 예측한다. 이렇듯 사람은 기억을 통해 마르코프 속성을 보완해 나간다.

강화학습에도 이러한 메모리 역할을 하는 순환 신경망을 사용함으로써 강화학습 모델을 과거의 정보와 현재의 정보를 연계하여 학습시킬 수 있다. 순환 신경망이 마르코프 속성을 학습하여 순환 신경망의 은닉 상태가 식 (1)의 $X_t, X_{t-1}, \ldots, X_0$의 역할, 즉 과거 정보를 하길 기대하는 것이다. 이를 통해 식 (1)을 따르지 않는 환경에서 순환 신경망을 이용해 식 (1)을 따르는 것처럼 학습하는 것에 도움이 될 수 있다.

3. 마르코프 속성 학습을 위한 GRU 비교 실험

실험 환경으로 OpenAI Gym의 CartPole과 LunarLander를 사용한다 [4]. 두 환경의 관찰

정보로는 오브젝트의 위치와 속도 등이 있다. 명백하게 마르코프 속성을 따르지 않는 환경을 만들기 위해 관찰 정보 중 오브젝트의 속도 정보를 제거하고 위치 정보만을 활용한다. 속도 정보 없이 현재 위치 정보만을 통해 다음 상태를 예측하는 것은 매우 어려우므로 식 (1)을 만족하기 힘들다.

실험에 사용된 네트워크의 구조는 다음과 같다. 비교군으로써 128개의 유닛을 가진 완전 연결 층 3개로 이루어진 네트워크가 사용되고, 해당 네트워크 앞에 재표현 층으로써 환경에서 나오는 입력 정보의 크기에 2배의 유닛 수를 가진 완전 연결 층과 GRU [5]층을 하나씩 추가한 네트워크을 사용한다. 학습에 사용되는 강화학습 알고리즘은 Double DQN [6]이다.

(그림 1, 2)는 학습 횟수에 따른 Episode Reward를 나타낸 그래프이다. 실험 5번의 평균치를 선으로 나타내었고 실험 중 최고점과 최저점을 흐린 영역으로 나타냈다. 파란색이 완전 연결 층으로만 이루어진 네트워크, 주황색이 재표현 층으로써 GRU를 활용한 네트워크의 실험 결과이다.

(그림 1)은 마르코프 속성을 따르지 않도록 속도를 제거한 환경에서의 실험 결과이다. 속도를 제거한 환경의 경우 완전 연결 신경망으로는 전혀 학습되지 않지만, 순환 신경망인 GRU를 활용했을 때는 CartPole에서 받을 수 있는 최고점수인 500점과 LunarLander에서 학습이 잘 되었을 때 받을 수 있는 200점을 달성하며 완전 연결 신경망의 학습보다 월등히 좋게 학습되었다. 이는 흐린 영역으로 표시된 부분에서 최고점을 보면 알 수 있다. (그림 2)는 마르코프 속성을 정확히 따르는 본래의 환경에서 학습한 결과이다. 이 경우에는 순환 신경망을 활용하면 학습 성능이 유지되거나 오히려 떨어지는 모습을 보인다. 이를 통해 (그림 1)에서 순환 신경망을 활용했을 때, 학습 성능이 개선된 이유가 순환 신경망이 완전 연결 신경망보다 뛰어나기 때문이 아니고, 마르코프 속성을 보장할 수 없기 때문임을 알 수 있다.

4. 결론

강화학습에서 마르코프 속성을 보장할 수 없는 환경을 학습할 때, 순환 신경망을 활용하면 학습이 거듭될수록 순환 신경망의 은닉 상태는 과거 정보를 더 효율적으로 나타내게 되고, 이는 순환 신경망이 마르코프 속성을 학습하는 것과 같은 효과를 낸다. 따라서 순환 신경망을 활용한 재표현 층을 거치면 마르코프 속성을 보장할 수 없는 환경에서도 별도의 사전 작업 없이 마르코프 속성을 만족하는 것에 가깝게 학습할 수 있다. 이는 일반적인 실제 환경과 같이, 마르코프 속성을 보장할 수 없을뿐더러 임의의 작업으로 마르코프 속성을 만족하게 만들기 힘든 환경에서 순환 신경망이 유용하게 쓰일 수 있음을 보여준다.

ACKNOWLEDGMENT

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2020R1I1A3065610).

참고문헌

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M., “Playing Atari with Deep Reinforcement Learning,” NIPS, 2013.

[2] Li, X., Li, L., Gao, J., He, X., Chen, J., Deng, L., & He, J., “Recurrent reinforcement learning: a hybrid approach,” 2015, arXiv preprint arXiv:1509.03044.

[3] Hausknecht, M., & Stone, P., "Deep recurrent Q-learning for Partially Observable mdps," 2015, AAAI.

[4] https://www.gymlibrary.ml/

[5] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y., “Empirical Evaluation of Gated Recurrent neural Networks on Sequence Modeling,” NIPS, 2014.

[6] Van Hasselt, H., Guez, A., & Silver, D., “Deep Reinforcement Learning with Double Q-learning,” In Proceedings of the AAAI conference on artificial intelligence, Vol. 30, No. 1, 2016.