온라인 강화학습에서의 암묵적 정규화에 의한 특징 상호 적응 현상에 관한 연구

이승민, 이정우
서울대학교

seungmin7792@cml.snu.ac.kr, junglee@snu.ac.kr

A Study on the Feature Co-Adaptation by Implicit Regularization in Online Reinforcement Learning

Seungmin Lee, Jungwoo Lee

Seoul National Univ.

요 약

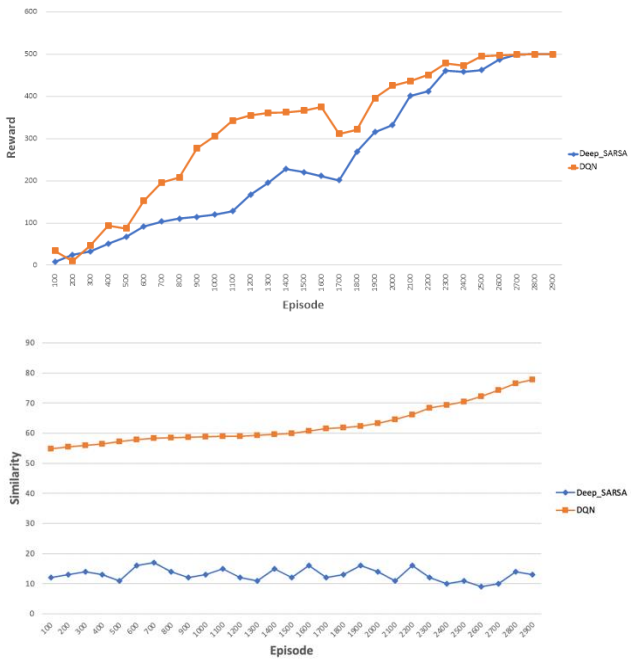
최근 연구에서 암묵적 정규화에 의한 특징 상호 적응 현상이 특정 오프라인 강화학습 알고리즘에서 중대한 영향을 끼침을 보임으로써 새로운 관점의 문제를 제시하였다. 본 연구에서는 특징 상호 적응 현상의 근본적인 원인에 근거하여 오프라인 강화학습 뿐만 아니라 온라인 강화학습에서 특징 상호 적응 현상이 나타남을 보이고, 더 나아가 온라인 강화학습에서 특징 상호 적응 현상에 기여하는 요인을 확인한다.

Ⅰ. 서 론

오프라인 강화학습은 오프라인 데이터 집합으로부터 학습된 정책을 온라인에 적용시키는 상황에서, 학습과정에서 관찰되지 않았던 행동에 대해 가치를 추정해야 하는 "Out-of-Distribution" 문제를 본질적으로 겪을 수밖에 없다. 이에, 기존에 "Out-of-Distribution" 문제에 대한 연구들이 오프라인 강화학습 연구의 주를 이뤘다. 하지만, 최근 연구에서 암묵적 정규화에 의해 연속된 상태-행동 쌍에 대한 심층 신경망의 특징 벡터 간의 유사도가 증가하는 특징 상호 적응 현상이 특정 오프라인 강화학습 알고리즘에서 중대한 영향을 끼침을 보임으로써 새로운 관점의 문제를 제시하였다. 특징 상호 적응 현상의 근본적인 원인은 학습 데이터 집합에서 관측된 적 없는 행동을 학습에 사용하기 때문인데, 이는 오프라인 강화학습 뿐만 아니라 온라인 강화학습에서도 나타나는 현상이다. 본 논문에서는 온라인 강화학습에서 특징 상호 적응 현상이 나타남을 확인하고 각 알고리즘의 차이점에 기반하여 특징 상호 적응 현상에 기여하는 요인에 대해 확인한다.

Ⅱ. 본론

본 논문에서는 OpenAI 에서 제공하는 강화학습 환경인 Gym 환경의 "CartPole-v1" 환경에서 온라인 환경에서 두가지 강화학습 알고리즘을 실행하고 학습 과정 중 심층 신경망의 특징 벡터 간의 유사도를 조사한다. 유사도는 코사인 유사도를 활용해 측정하고, 두가지 알고리즘은 각각 심층 Q 네트워크 학습과 심층 신경망을 활용하는 심층 SARSA 알고리즘이다. 결과는 아래와 같다.



Ⅲ. 결론

본 논문에서는 심층 Q 네트워크 학습에서 보상이 지속적으로 증가하며 학습이 정상적으로 이루어짐에도 불구하고, 학습이 진행됨에 따라 심층 신경망의 특징 벡터 간의 코사인 유사도가 지속적으로 증가함을

확인하였다. 이를 통해, 최근 연구에서 밝혀진 암묵적 정규화에 의한 특징 상호 적응 현상이 오프라인 심층 Q 네트워크 학습에서 뿐만 아니라 온라인 심층 Q 네트워크 학습에서도 나타남을 확인할 수 있었다. 또한, 온라인 심층 Q 네트워크 학습과 달리 온라인 심층 SARSA 알고리즘에서는 심층 신경망의 특징 벡터 간의 코사인 유사도가 비교적 매우 낮은 수준으로 유지되며 이러한 특징 상호 적응 현상이 일어나지 않음을 확인할 수 있었다. 이는 심층 SARSA 알고리즘에서는 항상 학습 데이터 집합에서 관측됐었던 행동을 학습에 사용하기 때문이다. 이처럼 본 논문은 온라인 강화학습에서 특징 상호 적응 현상의 발생유무를 확인하고, 학습 데이터 집합에서 관측된 적 없는 행동을 학습에 사용하는 것이 특징 상호 적응 현상의 근본적인 원인임을 밝혔다. 비록 온라인 강화학습에서 특징 상호 적응 현상이 일어나긴 하지만, 온라인 강화학습은 오프라인 강화학습과 다르게 학습 과정상에서 데이터의 비정상성과 최적 탐색과 관련된 요인들이 추가적으로 존재한다. 또한 비교적 간단한 환경인 “CartPole-v1” 환경에서 진행된 실험이기에, 온라인 강화학습에서의 특징 상호 적응 현상의 발생을 일반화하기엔 어렵다. 이에, 이후 비교 대조 실험을 통해 학습 데이터 집합에서 관측된 적 없는 행동을 학습에 사용하는 것, 데이터의 비정상성, 최적 탐색과 관련된 문제 각각이 특징 상호 적응 현상에 기여하는 정도를 확인하고, 더욱 다양하고 복잡한 환경에서 온라인 강화학습에서의 특징 상호 적응 현상의 유무를 확인하는 것을 후속 연구로 계획한다.

.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(34)), Institute of Information & communications Technology Planning & Evaluation (IITP- 2021-0-01059(33),IITP-2021-0-00106(33)) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21 FOUR program.

참 고 문 헌

[1] V. Mnih. et.al, “Playing atari with deep reinforcement learning,” arXiv:1312.5602, 2013.

[2] V. Mnih. et.al, “Human-level control through deep reinforcement learning,” Nature, 2015.

[3] Kumar, A., Agarwal, R., Ma, T., Courville, A., Tucker, G., and Levine, S. Dr3: Value-based deep reinforcement learning requires explicit regularization. In International Conference on Learning Representations, 2021b.

[4] Zhao D, Wang H, Shao K, Zhu Y. Deep reinforcement learning with experience replay based on SARSA. In: 2016 IEEE symposium series on computational intelligence (SSCI). IEEE; 2016. p. 1e6.