심층 Q-네트워크 학습에서 보상 및 학습 진행률 기반 적응형 탐색

이승민, 이정우
서울대학교

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

Reward and learning progress rate based adaptive exploration for Deep Q-Network learning

Seungmin Lee , Jungwoo Lee
Seoul National Univ.

요 약

심층 신경망 학습 분야의 발전과 더불어 강화 학습이 발전함에 따라, 심층 Q-네트워크 학습을 시작으로 심층 신경망 학습과 강화 학습을 결합하여 높은 성능을 달성한 심층 강화 학습이 등장하였다. 본 논문에서는 강화 학습의 본질적인 문제였던 이용과 탐색 사이의 상충 관계를 보상과 학습 진행률을 기반으로 한 적응형 탐색을 통해 해결하고자 하였고, 시뮬레이션을 통해 성능을 입증하였다.

Ⅰ. 서 론

심층 신경망 학습 분야의 발전과 더불어 강화 학습이 발전함에 따라, 심층 Q-네트워크 학습을 시작으로 심층 신경망 학습과 강화 학습을 결합하여 높은 성능을 달성한 심층 강화 학습이 등장하였다. 하지만, 강화 학습은 현재까지의 경험을 기반으로 보상을 최대화하는 행동을 수행하는 이용과 다양한 경험을 쌓기 위한 새로운 행동을 시도하는 탐색이 서로 상충하는 본질적인 문제를 갖고 있다. 본 논문에서는 이용과 탐색 사이의 균형을 찾는 보상 및 학습 진행률 기반 적응형 탐색 모델(Adaptive Reward and Learning progress based Exploration, ARLE)을 제안한다.

Ⅱ. 본론

가. 보상 및 학습 진행률 기반 탐색 모델 제안

1-ε의 확률로 현재 상태에서 보상을 최대화하는 행동을 선택하고, ε의 확률로 무작위 행동을 선택하는 ε-탐욕 정책을 사용한다. 본 논문에서는 아래 3 가지 아이디어를 기반으로 ε을 조절하는 보상 및 학습 진행률 기반 탐색 모델을 제안한다.

1) 강화 학습은 학습이 진행될수록 보상이 커져야 한다.
2) 학습이 진행될수록 보상이 커진다면, 탐색의 비중을 줄이고 이용의 비중을 늘려야한다.
3) 목표한 보상과 현재 상태에서 취한 행동이 얻은 보상의 차이가 클수록 탐색의 비중을 크게 늘려야한다.

본 논문에서 제안하는 ARLE 모델은 현재 상태에서 취한 행동이 얻은 보상이 지정한 보상 임계값보다 작을 경우 ε의 값을 감소시키지 않고 ε의 확률로 탐색한다. 반대로, 현재 상태에서 취한 행동이 얻은 보상이 지정한 보상 임계값보다 큰 경우 그 차이만큼 보상 임계값을 증가시키고 ε을 현재 상태에서 취한 행동이 얻은 보상과 목표한 보상의 차이에 반비례하는 값만큼 감소시킨 후 ε의 확률로 탐색한다.

나. 시뮬레이션

OpenAI Gym 의 CartPole-v1 환경에서 진행하였다.

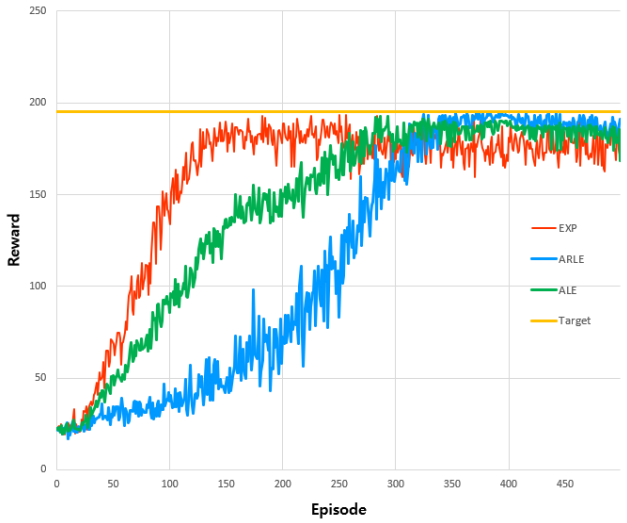

그림 1. 탐색 방법에 따른 성능 비교 분석

그림 1 에서 EXP 는 현재 상태에서 취한 보상이 보상 임계값보다 큰 경우 0.99 로 지정한 ε 감소율을 반복적으로 곱해서 ε를 감소시키며, 보상 임계값을 증가시킨다. ALE 는 보상 임계값에 상관없이 행동을 취할 때마다 현재 상태에서 취한 행동이 얻은 보상과 목표한 보상의 차이에 반비례하는 값만큼 감소시킨 후 ε의 확률로 탐색한다.

Ⅲ. 결론

EXP 와 ARLE 를 분석해보았을 때, 둘 다 현재 상태에서 취한 행동이 얻은 보상과 보상 임계값의 대소관계를 기반으로 ε을 감소시킨다는 점은 같다. 하지만, EXP 는 ε을 지수적으로 감소시키므로 ARLE 보다 훨씬 빠른 속도로 탐색의 비중을 줄이고 이용의 비중을 높인다. 이 때문에 EXP 는 목표한 보상에 훨씬 빠른 속도로 도달하지만, 충분한 탐색을 거치지 않았기 때문에 학습 후반부에 일정한 보상 값을 유지하지 못하고 불안정한 모습을 보이며 더 낮은 최대 보상값을 가진다. ALE 와 ARLE 를 분석해보았을 때, 둘 다 현재 상태에서 취한 행동이 얻은 보상과 목표한 보상의 차이에 반비례하는 값만큼 ε을 감소시킨다는 점은 같다. 즉, ALE 와 ARLE 모두 학습 진행률에 기반한 적응형 탐색을 하기 때문에 학습 후반부에 EXP 보다 훨씬 안정된 보상 값을 유지함을 확인할 수 있다. 하지만, ALE 는 보상 임계값에 상관없이 ε을 감소시키기 때문에 ARLE 보다 비교적 빠른 속도로 탐색의 비중을 줄이고 이용의 비중을 높인다. 따라서 ARLE 에 비해 충분한 탐색을 거치지 않아, ARLE 보다 수렴 속도는 빠르지만 조금 더 낮은 최대값을 가짐을 알 수 있다. 정리하면, 보상 임계값에 따른 보상 기반 탐색은 학습의 성능(최대치)을 높이는데 기여하고, 학습 진행률에 따른 적응형 탐색은 학습의 수렴 안정성에 기여함을 확인할 수 있다.

ACKNOWLEDGMENT

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504(30)), Institute of Information & communications Technology Planning & Evaluation (IITP- 2021-0-00106(40),IITP-2021-0-02068(40)) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21-plus

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] L. P. Kaelbling. et.al, “Reinforcement Learning: A Survey,” Journal of Artificial Intelligence Research, vol. 4, 1996.