다중 에이전트 강화학습에서의 앙상블 알고리즘

주석훈, 이정우*
서울대학교,

seokhunju@cml.snu.ac.kr, junglee@snu.ac.kr*

Ensemble Algorithm in Multi-agent Reinforcement Learning

Ju Seok Hun, Lee Jung Woo*
Seoul National Univ.

요 약

다중 에이전트 강화학습은 단일 에이전트 강화학습과 달리 에이전트의 수가 늘어남에 따라 탐색해야 할 상태와 행동 공간의 크기가 지수함수적으로 증가한다. 그리고 각 에이전트의 학습에 독립적인 심층 신경망을 사용할 경우 에이전트의 수에 따라 연산량과 학습시간이 증가하는 문제가 존재한다. 이러한 문제로 에이전트간에 심층 신경망을 공유하는 가치기반 다중 에이전트 강화학습 알고리즘들이 제안되었다. 공유된 신경망을 사용하는 경우 에이전트 행동 다양성과 탐색과정의 문제가 존재할 수 있고 이 문제를 완화할 수 있는 방법으로 앙상블 알고리즘을 적용한 다중 에이전트 강화학습을 제안하고 그 성능을 분석한다.

Ⅰ. 서 론

강화학습 환경에서 상태와 행동공간에 대한 효율적인 탐색 알고리즘은 학습속도와 성능에 큰 영향을 미친다. 때문에 효율적인 탐색 알고리즘에 대한 연구가 활발히 진행되고 있고, 다중 에이전트 강화학습의 경우 탐색해야 할 상태 행동 공간이 에이전트의 수가 증가함에 따라 지수적으로 커지기 때문에 효과적인 탐색과 탐색한 상태, 행동을 효과적으로 활용할 수 있는 알고리즘이 필요하다.

최근 많은 연구에서 에이전트의 학습에 심층 신경망을 사용하고 있는데, 다중 에이전트 강화학습의 경우 에이전트 수가 늘어남에 따라 학습시켜야 하는 심층 신경망의 수가 비례해서 증가하므로 학습시간의 증가와 연산량 증가의 문제가 있다. 때문에 기존의 가치 기반 다중 에이전트 강화학습에서는 연산량 증가의 문제로 각 에이전트마다 각각 다른 심층 신경망을 학습시키는 것이 아니라, 단일 심층 신경망을 사용하는 대신 입력 데이터에 에이전트를 식별할 수 있는 벡터를 추가하여 구별해 학습하는 방향으로 진행되어왔다. 이러한 경우 빠른 학습에 도움이 될 수 있으나 에이전트 사이에 매개변수의 공유가 크고, 따라서 행동공간에 대한 탐색이나 에이전트간 행동 유사성이 커지는 문제가 발생할 수 있다.[1] 이에 대한 문제를 완화할 수 있는 방법으로 앙상블 알고리즘을 적용한 다중 에이전트 강화학습을 제안하고 그 성능을 분석한다.

Ⅱ. 본론

가치 기반 다중 에이전트 강화학습에서 대표적인 알고리즘으로는 QMIX 알고리즘을 들 수 있다.[2] QMIX 알고리즘에서는 서론에서 언급한 것 처럼 연산량의 문제로 단일 신경망을 사용하여 모든 에이전트의 상태-행동 가치를 학습한다. 그리고 상태, 행동 공간 탐색에는 입실론 그리디 탐색방법을 이용하고 있다. 본 논문에서는 QMIX 논문에서 제안한 알고리즘을 기반으로 앞서 언급한 두 가지 문제점, 큰 행동 유사성, 효율적인 상태 행동공간 탐색 문제의 해결 방안으로 앙상블을 사용한 알고리즘을 제안하고, QMIX 논문에서 실험한 환경과 동일한 환경에서 실험을 진행하여 그 결과를 분석한다. 또한 앙상블 알고리즘을 적용할 때 여러 개의 상태-행동 가치 심층신경망을 생성하는 방법으로 진행했기 때문에 앙상블 신경망들의 결과값을 평균하여 사용하는 방식인 배깅 알고리즘을 사용하였다.

그림 1 SMAC(2s3z) 환경

Ⅲ. 실험결과

실험 환경은 SMAC[3] 에서 제공하는 2s3z 환경을 사용하였다. 기본이 되는 QMIX 알고리즘과 3 개의 앙상블을 이용한 경우, 5 개의 앙상블을 이용한 경우에

대해 각각 3 개의 시드를 사용하여 실험하였다. 실험결과는 아래 그림 2 와 같다. 실험결과의 가로축은 학습한 스텝 수에 해당하고 세로축은 일정한 주기로 테스트를 진행했을 때 승률을 의미한다.

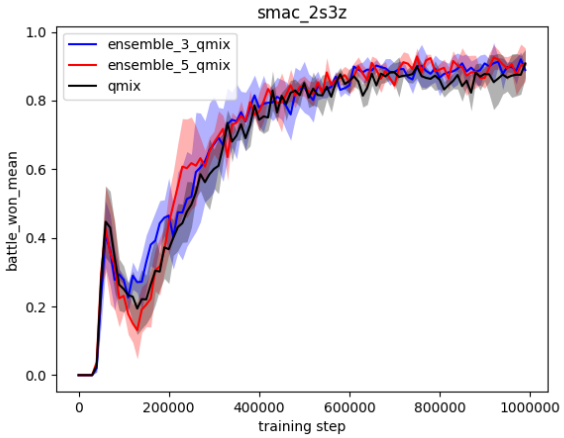

그림 2 실험 결과

실험 결과 학습 초기에는 기존의 QMIX 알고리즘이 좀 더 승률이 높은 것으로 나타나지만 학습이 진행됨에 따라 앙상블을 사용한 알고리즘이 좀 더 높은 승률을 기록함을 확인할 수 있었다. 그러나 3 개의 앙상블을 사용하는 경우와 5 개의 앙상블을 사용하는 경우 사이에는 현저한 차이가 없음을 확인할 수 있었다.

Ⅳ. 결론

다중 에이전트 강화학습에서 에이전트 간의 행동 유사성, 효율적인 상태 행동공간 탐색 문제를 해결하고자 앙상블 알고리즘을 적용해보았고 그 결과를 분석하였다.

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] Chenghao Li et al. 2021. Celebrating Diversity in Shared Multi-Agent Reinforcement Learning. Advances in Neural Information Processing Systems 34 (2021).

[2] Rashid, T., Samvelyan, M et al. QMIX: Monotonic value function factorization for deep multi-agent reinforcement learning. In Proceedings of the 35th International Conference on Machine Learning, 2018.

[3] Mikayel Samvelyan, Tabish Rashid et al. The StarCraft Multi-Agent Challenge. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019.