협동 다중 에이전트 강화학습에서 Next Observation 예측을 통한 호기심 기반 탐험 방법

김주봉, 최호빈, 최요한, 허주성, 한연희[1]
한국기술교육대학교 미래융합공학전공

{rlawnqhd, chb3350, yoweif, chill207, yhhan}@koreatech.ac.kr

Curiosity-Based Exploration Method through Next-Observation Prediction for Cooperative Multi-Agent Reinforcement Learning

Ju-Bong Kim, Ho-Bin Choi, Yohan Choi, Joo-Seong Heo, Youn-Hee Han[1]
Future Convergence Engineering
Korea University of Technology and Education

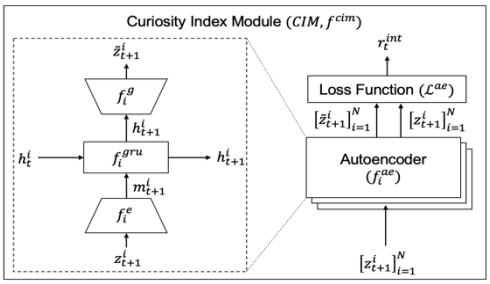
요 약

에이전트들의 복잡한 행동 조정이 필요한 협동 다중 에이전트 강화학습(multi-agent reinforcement learning; MARL)에서 효율적인 탐험 전략은 하나의 해결책이 될 수 있다. 본 연구에서는 중앙 집중 훈련 및 분산 실행(centralized training and decentralized execution; CTDE) 기반 MARL 알고리즘에 쉽게 통합할 수 있는 새로운 호기심 기반 탐험 방법을 제안한다. 제안되는 탐험 방법을 활용하여 다음 관찰(observation)의 호기심에 영향을 받는 탐험 보너스를 생성하는데, 이 때 보너스는 MARL 작업에서 발견되는 에이전트의 역학 및 예측이 어려운 확률적 상태 전환에 의해 크게 영향을 받지 않는다. 연속적으로 복잡한 행동 조정이 요구되는 환경인 M-step payoff matrix game 을 소개하며, 기존에 잘 알려진 다른 탐험 방식과의 비교 평가를 통해 제안된 탐험 방법이 CTDE 기반 MARL 알고리즘에서 상당한 성능 향상을 달성함을 보인다.

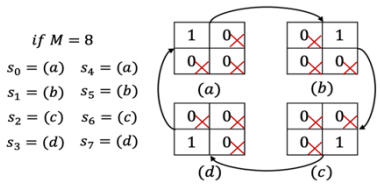
Ⅰ. 서 론

협동 다중 에이전트 강화 학습(MARL)은 자율 주행 자동차 및 로봇 군집 제어와 같은 다중 에이전트 시스템으로 모델링 된 여러 실제 문제를 해결하기 위한 중요한 도구이다. 이러한 시스템에서 에이전트들은 부분적으로만 관찰 가능한 특성으로 인해 각 에이전트의 행동-관찰 기록(action-observation history)을 활용하여 개별 정책을 훈련한다. MARL 에서 에이전트 수가 증가함에 따라 분산된 개별 정책의 공동 행동 공간이 지수적으로 커져 훈련이 어려워진다. 이러한 문제는 개별 수준의 정책 최적화가 공동 정책의 최적화로 이어지는 중앙 집중 훈련 및 분산 실행(centralized training and decentralized execution; CTDE)을 통해 해결된다.

기존 호기심 기반 탐험(exploration) 방법은 현재 관찰-행동 쌍에 대하여 다음 관찰을 예측하는 작업을 통해 상태에 대한 호기심을 도출한다. 하지만 현재 관찰-행동 쌍에 대한 다음 관찰 사이의 직접적인 종속성 때문에 개별 정책의 변화는 공동 정책의 변화를 유발한다. 이러한 종속성으로 인해 개별 에이전트의 탐험이 제한되는 문제를 해결할 필요가 있다. 따라서 본 연구에서는 언급된 문제를 해소할 수 있는 새로운 호기심 기반 탐험 방법을 소개한다. 다음 관찰만을 예측하는 문제를 도입함으로써 현재 관찰-행동 쌍과 다음 관찰 사이의 종속성에 영향을 받지 않는 방법을 제안한다. 이 전까지 관찰된 기록과 다음 관찰을 활용하여 다음 관찰을 직접 예측하는 문제를 통해 호기심을 도출하는 방식이다. 관찰 경험이 적거나 경험이 적은 과거 관찰 기록을 통해 다음 관찰을 방문할 때

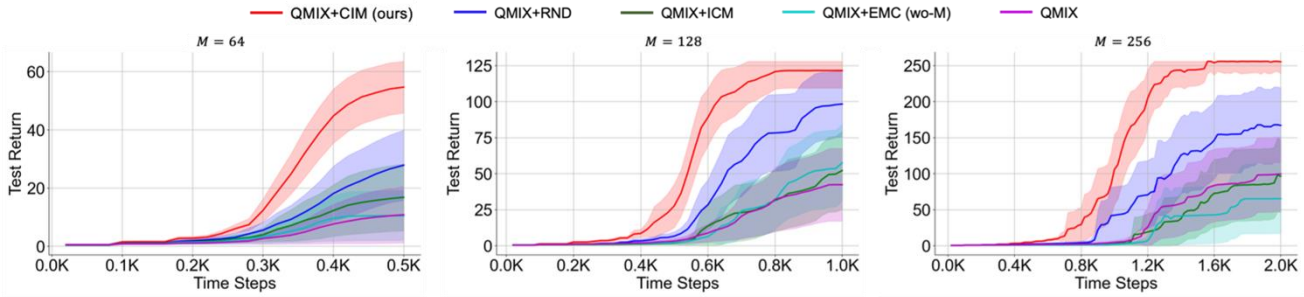


(그림 1) CIM Architecture



(그림 2) M-step payoff matrix game

[1] 한연희(Youn-Hee Han, yhhan@koreatech.ac.kr): 교신저자



(그림 3) M-step payoff matrix game에서 QMIX에 적용된 탐험 기법들의 성능 비교 그래프

호기심이 크게 영향 받는다. 정책 훈련 과정에서 호기심은 탐험 보너스(exploration bonus)로써, 환경으로부터 주어지는 보상에 더해져 훈련에 활용된다.

Ⅱ. 본론

본 장에서는 MARL 에 대한 탐험 보너스를 생성하는 새로운 접근 방식을 소개한다. 에이전트는 탐험을 통해 방문하지 않았던 관찰을 방문함으로써 에이전트의 무지를 줄일 수 있다. 에이전트가 방문한 관찰에 대한 호기심은 해당 관찰에 대한 무지의 정도와 같으며, 호기심은 해당 관찰에 대한 무지를 줄이기 위한 탐험을 에이전트에게 장려한다. 따라서 본 연구에서는 탐험 보너스가 모든 관찰에 대한 호기심을 나타내야 한다고 주장하고, 탐사 보너스를 생성하기 위해 CTDE 기반 MARL 알고리즘에 쉽게 적용할 수 있는 호기심 지수 모듈(curiosity index module; CIM)을 제안한다(그림 1 참고). CIM 은 현재 관찰-행동 쌍에 직접 의존하지 않고, 개별 에이전트의 과거 관찰 기록과 순환 오토인코더(recurrent autoencoder)를 통해 다음 관찰을 재구성한다. CIM 은 각 에이전트 $i \in \mathcal{N} \equiv \{1, 2, \dots, N\}$ 마다 대응되는 오토인코더 f_i^{ae} 와 손실 함수 $\mathcal{L}^{ae}$ 로 구성된다. CIM 은 임의의 시간 t 에서 각 에이전트의 다음 관찰 $[z_{t+1}^i]_{i=1}^N \in \mathcal{Z}^{\mathcal{N}}$ 을 입력으로 받아 탐험 보너스 r_t^{int} 를 출력한다. f_i^{ae} 는 인코더 f_i^e , 순환 모듈 f_i^{gru} 그리고 디코더 f_i^g 로 구성된다. f_i^e 는 z_{t+1}^i 를 입력으로 받아 d 차원의 표현 벡터(representation vector) m_{t+1}^i 를 생성하고, f_i^{gru} 는 m_{t+1}^i 와 이 전 시간 $t-1$ 에 생성된 d 차원의 히든 벡터 h_t^i 를 입력으로 받아 d 차원의 히든 벡터 h_{t+1}^i 를 생성한다. 마지막으로 f_i^g 는 h_{t+1}^i 를 입력으로 받아 다음 관찰에 대한 예측인 $\tilde{z}_{t+1}^i$ 를 출력한다. 손실 함수 $\mathcal{L}^{ae}$ 는 다음의 수식에 의해 r_t^{int} 를 생성한다.

$$\mathcal{L}^{ae}\left([z_{t+1}^i]_{i=1}^N, [\tilde{z}_{t+1}^i]_{i=1}^N\right) = r_t^{int} = \frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{z}_{t+1}^i - z_{t+1}^i\right\|_2^2 \quad (1)$$

수식 (1)에 의해 계산된 r_t^{int} 에 탐험 보너스 적용 가중치 β 가 곱해지며, 환경으로부터 주어지는 보상 r_t^{ext} 에 더해져 보상 $r = r_t^{ext} + \beta r_t^{int}$ 을 대체한다.

제안된 CIM 아키텍쳐에서 다음 경우들에 대하여 호기심이 높게 산출되는 것을 의미하며, 즉 높은 탐험 보너스가 산출된다.

i. 다음 관찰이 많이 경험되지 않았을 때
ii. 많이 경험되지 않은 과거 관찰 기록을 통해 다음 관찰을 방문할 때

Ⅲ. 실험 평가

그림 2 는 두 에이전트 ($\mathcal{N} = \{1, 2\}$)가 있는 M-step payoff matrix game 을 묘사한다. 행과 열은 각각 에이전트의 전략을 설명하며 행과 열의 교차점의 숫자는 두 에이전트에게 주어지는 공동 보상을 의미한다. 에피소드는 최대 M 스텝 진행할 수 있고, 매 스텝마다 공동 보상이 1일 경우 다음 스텝으로 진행할 수 있으나 그렇지 않은 경우에는 에피소드를 종료한다.

실험에서는 기존에 잘 알려져 있는 CTDE 기반 MARL 알고리즘인 QMIX [1]에 제안하는 기법인 CIM 을 비롯하여 RND [2], ICM [3] 그리고 EMC (wo-M) [4]을 각각 적용한 후 (64, 128, 256-step) payoff matrix game 에서의 성능을 비교 평가한다. 그림 3 은 각 환경에서 모든 알고리즘을 각각 20 번 수행한 결과의 평균과 95% 신뢰 구간을 나타낸 성능 그래프이다. QMIX+ CIM 은 다른 방법들과 비교하였을 때 각 환경에서 얻을 수 있는 최적의 성능을 달성함으로써 가장 뛰어난 성능을 보였다.

Ⅳ. 결론

본 논문에서는 MARL 환경의 동적 특성에 크게 영향을 받지 않는 탐색 방법을 제안한다. 개별 에이전트의 과거 관찰 기록과 다음 관찰을 통해서 다음 관찰에 대한 예측 오차를 계산하여 탐험 보너스를 생성한다. 제안된 탐색 방법은 잘 알려진 CTDE 기반 MARL 알고리즘에 적용하기 용이하며 다른 탐험 알고리즘들과 비교 실험 평가를 통해 대표적인 알고리즘인 QMIX 의 높은 성능 향상을 이끌어 낼 수 있음을 보인다.

ACKNOWLEDGMENT

이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2018R1A6A1A03025526 & No. 2020R1I1A3065610).

참 고 문 헌

[1] Rashid, T. et al., “QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning,” ICML, 4295-4304, 2018.

[2] Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O., “Exploration by random network distillation,” ICLR, 2019.

[3] Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T., “Curiosity-Driven Exploration by Self-Supervised Prediction,” ICML, 2778-2787, 2017.

[4] Zheng, L. et al., “Episodic Multi-agent Re- inforcement Learning with Curiosity-driven Exploration,” NIPS, 3757-3769, 2021.