https://doi.org/10.7840/kics.2023.48.1.81

강화학습의 트리 탐색 기반 종단간 소프트 Q러닝

한 승 엽◆, 조 태 현*, 한 형 근*,
이 희 수*, 김 형 진*, 이 정 우°

Tree Search Based End-to-End Soft Q-Learning in Reinforcement Learning

Seungyub Han◆, Taehyun Cho*,
Hyeonggeun Han*, Heesoo Lee*,
Hyungjin Kim*, Jungwoo Lee°

요 약

강화학습을 다루는 환경 중에서도 희소 보상 (sparse reward) 환경 체계에서 특히 기존의 시간 차이 학습 (temporal difference learning, TD) 기반 탐험법의 경우 보상을 주는 상태-행동 쌍을 발견하기 힘들 뿐만 아니라, 발견하더라도 해당 궤적에 있는 상태-행동 쌍에 대한 가치 함수에 반영이 빠르게 되지 못한다. 본 논문에서는 희소 보상체계 궤적에서 얻어지는 보상 값의 합이 종단 간 학습 과정에서 효율적으로 학습될 수 있는 Boltzmann guided sparse sampling (BGSS) 기법을 제안한다. 실험결과에서도 BGSS가 기존의 트리 탐색법과 시간 차이 학습법 모두 비교했을 때 훨씬 빠른 속도로 학습됨을 보였다.

키워드 : 몬테카를로 트리 탐색, 볼츠만 희소 샘플링, soft Q 학습

Key Words : Monte Carlo tree search, Boltzmann guided sparse sampling, soft Q-learning

ABSTRACT

In the sparse reward system of reinforcement learning, existing temporal difference learning based exploration is especially hard to both find state-action pairs which give non-zero reward and update the value function on those state-action pairs even if a reinforcement learning agent finds a reward giving state-action pair. In this paper, we propose Boltzmann guided sparse sampling (BGSS), which is an end-to-end reinforcement learning that can efficiently learn to find trajectories with high return from sparse-reward system. In the experiments, we can also demonstrate the proposed method, BGSS, is a faster reinforcement learning algorithm than tree search-based method and temporal difference learning.

Ⅰ. 서 론

강화학습은 주어진 환경에서 정의된 agent가 현재의 상태 (state)와 자신이 할 수 있는 행동 (action)을 인식하여 선택 가능한 행동 중에서 상태별로 주어지는 보상 값의 합이 정의된 마르코프 결정 과정 (Markov decision process, MDP)에서 최대가 되는 궤적을 찾는 기계 학습법이다. 최근 인공신경망을 사용하는 심층 강화학습 연구에서는 시간 차이 학습 기반으로 MuJoco 보행 제어 임무와 같은 도전적인 시뮬레이션 환경에서 우수한 성능을 보여주었다[1,2].

강화학습의 경우 기존의 최적 제어 기법과 달리 시스템 전이 (system transition)가 알려지지 않은 상황에서 직접 탐험 (exploration)을 하며 각 상태-행동 쌍에 대해 주어지는 보상 값을 주어진 초기 상태에서부터 보상 값의 합이 최대가 될 수 있는 궤적을 찾아낼 수 있도록 하기에 시간 차이 학습으로는 희소 보상 체계에서 정책망 개선 (policy improvement)를 빠르게 진행하기 어렵다. 희소 보상 체계란 강화학습 환경에서 대부분 상태-행동 쌍에서 보상을 받지 못하고, 정해진 목표에 거의 도착해야 보상을 받을 수 있는 시스템을 말하는데, 매 스텝마다 진행되는 정책망 평가

※ 본 연구는 2022년 정부(방위사업청)의 재원으로 국방과학연구소의 지원 (UD190031RD), BK21-plus 및 서울대학교 뉴미디어통신공동연구소 관리로 수행되었습니다.
◆ First Author : Seoul National University, seungyubhan@snu.ac.kr, 학생회원
° Corresponding Author : Seoul National University, junglee@snu.ac.kr, 종신회원
* Seoul National University Department of Electrical and Computer Engineering, {talium, hygnhan, leehs, hjkim}@cml.snu.ac.kr, 학생회원
논문번호 : 202210-24-C-LU, Received October 14, 2022; Revised November 2, 2022; Accepted November 2, 2022

(policy evaluation)이 탐험 과정에서 학습 초기에는 보상 값 합의 변동을 일으키기가 어려운 환경이다. 희소 보상 체계가 아닌 MuJuco와 달리 대표적인 희소 보상 환경으로는 바둑, mini-grid four-rooms와 openAI reacher 환경이 있다. 특히 바둑의 경우 AlphaGo에서 Monte Carlo Tree Search (MCTS)을 기반으로 한 upper confidence bound tree (UCT)와 심층 강화학습을 결합한 방법론으로 희소 보상 체계의 문제를 해결하였다[3]. 하지만 환경에서 주어지는 모든 상태-행동 쌍을 방문 횟수 기반으로 탐험해야 하는 단점이 있다.

본 논문에서는 종단간으로 행동에 대해 희소 샘플링이 가능한 방법론인 BGSS를 제안하여 탐험 공간을 대폭 줄여 기존 방법론 보다 빠른 학습속도를 보일 수 있었다.

Ⅱ. 트리 탐색 기반 강화학습 기법

2.1 문제 정의

종단간 학습을 목표로 하는 희소 보상 환경에서의 이산 행동 공간을 가진 강화학습으로, finite-horizon MDP를 고려한다. Finite-horizon MDP란 (S, A, p, r, γ) 튜플로 이루어져 있고, 상태 공간 $\mathcal{S}$, 행동 공간 $\mathcal{A}$, 보상 함수 $r: S \times A \rightarrow \mathbb{R}$,전이 함수 $p: S \times A \times S \rightarrow$[0,1], discount factor $\gamma \in$[0,1)를 나타내며, 결정 과정의 시간 스텝이 T로 유한한 상황을 말한다.

주어진 환경에서 강화학습 agent의 목표는 총 보상 값의 discounted sum으로 정의되는 Q-함수와 이를 최대화하는 정책망 $\pi(a|s)$를 학습하는 것이다.

$$Q(s_t, a_t) = \mathbb{E}_{\pi}\left[\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})\right] \quad (1)$$

본 논문에서 사용할 행동 정책 (behavior policy)는 소프트 Q-러닝[4]에서의 최대 엔트로피 정책을 사용하기 때문에 학습할 최대 엔트로피 정책 π_{soft}는 다음의 softmax 함수로 정의되며, 이를 위한 Q-함수 $\hat{Q}$는 아래에 정의한다.

$$\pi_{soft}(s) = \exp\left(\frac{1}{\alpha}\,\hat{Q}(s,a) - \sum_{a} \hat{Q}(s,a)\right) \quad (2)$$

α는 온도 hyperparameter이다. 최대 엔트로피 정책을 만들기 위해 Q-함수를 엔트로피를 보상 값에 삽입하여 보상 이외에 각 정책 확률의 엔트로피 값도 커지게 $\hat{Q}$를 정의한다.

$$\hat{Q}(s_t, a_t) = \mathbb{E}_{\pi}\left[\sum_{t'=t}^{T} \gamma^{t'-t}(r(s_{t'}, a_{t'}) + \alpha H(\cdot\,|s_{t'}))\right] \quad (3)$$

2.2 희소 샘플링 기반 행동 선택 법칙

매시간 스텝마다 탐험을 위한 행동 정책 기반으로 얻어진 (s_t, a_t, r_t, s_{t+1}) 쌍을 리플레이 메모리에 저장하였다, 샘플하여 한 스텝 시간 차이만큼의 업데이트를 하는 시간 차이 강화학습과 달리, 본 논문에서 제안할 행동 선택 법칙은 탐험 과정에 사용되는 행동 선택과정을 트리 탐색 기반 강화학습이다.

기존의 트리 탐색 기반 강화학습[3]과 달리 한 시간 스텝에서 행동 선택을 점진적으로 효율적인 의사 결정을 하기 위해 희소 샘플링 기반 방법론인 BGSS를 다음 장에서 정의한다.

π트리 탐색 기반 행동 선택은 노드 확장 (node expansion), 롤아웃 (rollout)과정으로 나뉘는데, 주어진 D 시간 스텝만큼은 초기 상태에서 가능한 행동 공간에 있는 행동을 주어진 희소 샘플링 정책으로 탐험하여 가능한 가지 노드 조합을 방문한 후, D 시간스텝 이후에는 노드 확장을 더 이상 하지 않는다. 이후 를 활용하여 롤아웃을 하는데, 마지막 시간 T까지 제어하여 얻어지는 $\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$를 계산하여 노드의 Q 값을 업데이트한다.

즉, T 길이만큼의 에피소드를 시뮬레이션 상에서 탐험하는 동안 D 시간만큼은 트리의 가지 노드를 늘리는 방식으로 행동 선택을 하며, D 이후에는 롤아웃으로 D 이후의 하나의 궤적만 얻는 방식을 사용한다. 이러한 과정을 총 M 번 반복하며 각 노드의 Q값은 터미널 상태에서의 보상 값의 discounted sum 기준으로 업데이트 된다. 최종적으로, M번의 시뮬레이션 에피소드 이후 실제환경에서의 다음 스텝으로 넘어가기 위한 행동 a_t를 시행한다.

Ⅲ. 종단간 학습가능한 정책 기반 희소 샘플링 기법

3.1 제안 기법

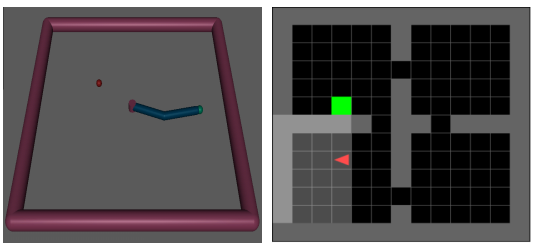
BGSS는 에너지 및 온도 기반 소프트 Q 함수를 사전에 학습된 고정된 값이 아닌 종단간으로 실시간 업

데이트를 해가며 주어진 행동 공간에서 행동 선택을 할 때 희소 샘플링을 할 수 있게 해준다. 희소 샘플링은 행동 공간에서 가능한 모든 행동들을 방문해보지는 않고 희소 샘플링 확률 $P(a_t|s_t)$에 따라 높은 보상을 얻을 수 있는 행동들만 선택적으로 방문하게 된다. 이때 기존의 사전 정보를 통해 얻어진 확률로 진행하는 희소 샘플링 방법론들과 달리 소프트 Q-러닝 기반으로 종단간으로 학습 가능하여 실시간으로 업데이트 되는 희소 샘플링 방법을 제안한다.

희소 샘플링 확률을 위에서 정의한 π_{soft}를 직접 트리 탐색에 결합하여 사용하면 기존의 UCT 및 사전 정의된 희소 샘플링 확률을 사용하는 것보다, MCTS를 통한 행동 선택을 점진적으로 업데이트 된 정책에서 선택할 수 있을 뿐 아니라, 시간 차이 학습법에서 한 스텝으로 업데이트 된 값의 오류보다 본래의 벨만 업데이트 과정에 더욱 가까운 정확한 업데이트를 할 수 있다. 또한 제안된 BGSS 방법론은 마지막 상태까지의 정확한 Q 값을 평가할 수 있으며, 롤아웃 과정에서 사용되는 기존의 정책에 비해 종단간으로 실시간 업데이트 되는 정책으로 평가하기 때문에 더 정확한 마지막 가지 노드의 값을 얻을 수 있다.

3.2 실험 결과

제안된 BGSS의 경우 효율성을 증대하기 위해 시간 차이 학습법으로 먼저 20000 에피소드 동안 학습하였다. 실험에 사용된 두 시뮬레이션 환경은 그림 1에서 나타냈고, 두 환경 모두 그림에서 표시된 빨간색 점, 초록색 점에 도달하여야 희소보상을 얻을 수 있는 환경이다. Action space는 [0,1]의 연속 행동 공간을 다음 5개의 이산 행동 [0, 0.01, 0.1, 0.5, 1]으로 축소한 reacher 환경에서 사용하였고, [256,256] fully connected layer를 사용하여 0.001 learning rate로 학습하였다. 그림 2의 경우 각 방법론의 논문에서 제시된 방법으로 학습 후 두 환경에서 테스트한 결과이다.



(a) Reacher (b) Four rooms

그림 1. 실험에 사용된 다중 목표 환경
Fig. 1. Multi-goal environments

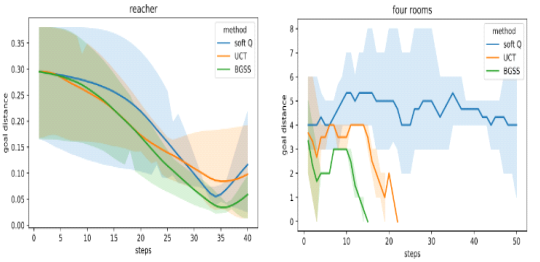


그림 2. 목표와의 거리 성능 비교
Fig. 2. Distances between goal and position

표 1. 목표 지향 강화학습 환경에서의 실험 결과
Table. 1. Results on multi-goal RL environments

Method	openAI reacher-v2		Minigrid Four-rooms	
	return	D_{min} (m)	return	D_{min} (칸)
UCT	-10.41	0.052	769.2	0
Soft Q(TD)	-10.09	0.0409	271.9	2.67
BGSS	-9.077	0.0168	825.0	0

5번의 테스트 결과를 평균과 표준편차, 그리고 최소값을 그림2와 표 1에서 나타냈다. BGSS는 희소 보상 환경에서 높은 성능을 보여주었지만 그림 2와 표 1에서 확인할 수 있듯이 UCT+MCTS와 시간 차이 학습 기반 소프트 Q러닝의 경우 목표 지점에 도착하여 얻는 희소 보상을 얻는 데에 실패하였다.

Ⅳ. 결 론

제안된 방법론은 기존의 심층 강화학습에서 사용되는 탐험 과정을 좀 더 정확하고 효율적으로 진행할 수 있는 방법이라 할 수 있다. 종단간으로 업데이트 가능한 희소 샘플링 정책으로 트리 탐색과정으로 탐험과정을 진행한 것이 핵심이다. 실험결과에서도 희소 보상을 얻을 수 있는 확률을 높여 효율적인 탐험을 할 수 있었다. 향후 후속연구로 트리 탐색 기반을 탐험 과정에서만 쓰는 새로운 탐험기법을 제안하여 제어 과정에서는 계산 복잡도를 줄이는 연구를 진행해 볼 수 있을 것이다.

References

[1] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in *Proc. ICLR 2016*, San Juan, Puerto Rico, May 2016.

(https://doi.org/10.48550/arXiv.1509.02971)

[2] J. Moon and B. Shim, "Deep reinforcement learning-based vehicle-to-vehicle resource allocation," *J. KICS*, vol. 47, no. 10, 2022. (https://doi.org/10.7840/kics.2022.47.10.1565)

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484-489, 2016. (https://doi.org/10.1038/nature16961)

[4] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *ICML 2017*, pp. 1352-1361, Sydney, Australia, Aug. 2017. (https://doi.org/10.48550/arXiv.1702.08165)