분할 배치 앙상블 DQN

이도혁, 이정우
서울대학교

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

Separated Batch Ensemble DQN

Lee Do Hyeok, Lee Jung Woo
Seoul National Univ.

요 약

본 논문은 기존의 DQN의 과추정 문제를 앙상블 기법을 통해 해결한 앙상블 DQN의 탐색을 분할 배치 학습으로 개선한 분할 배치 앙상블 DQN을 제안한다. 제안하는 기법은 기존 앙상블 DQN 대비 적은 탐색으로도 효율적인 학습을 하며, 최종적으로 더 높은 성능을 냄을 실험을 통해 보인다.

Ⅰ. 서 론

심층 Q-신경망(Deep Q-Network, DQN)[1]의 등장으로, Atari를 비롯한 다양한 환경에서 Q-값(Q-value) 추정을 이용한 다양한 학습이 가능해졌다. 하지만 OpenAI Gym 등 더 복잡하고 어려운 벤치마크 환경이 나오면서, DQN의 고질적인 과추정(overestimation) 문제가 학습을 저하시키는 가장 큰 문제로 대두되었다. 이를 해결하기 위해 여러 개의 Q-값을 학습시켜 평균화하는 앙상블 DQN(Ensemble DQN) 기법[2]이 제시되었고, 좋은 성능을 보였다. 본 논문에서는 모든 목표 Q-값(Target Q-value) 추정이 동일 배치(batch)에서 진행되는 앙상블 기법을 기반으로, Q-신경망 별로 다른 배치를 사용하는 분할 배치 앙상블 DQN(Separated Batch Ensemble DQN)을 제안한다. 이를 통해 행동 공간(action space)이 복잡한 환경에도 좋은 성능을 낼 수 있는 알고리즘을 설계하였다.

Ⅱ. 분할 배치 앙상블 DQN

DQN 알고리즘은 심층 신경망(Deep Neural Network)을 이용한 근사화함수(function approximator)와 재현 버퍼(replay buffer)를 이용해서 Atari 게임에서 좋은 성능을 보여주었다. 목표 Q-값은 재현 버퍼에서 표본 추출(sampling)한 미니 배치(mini-batch)에서 기댓값을 계산하여 추정하고, 이를 이용해 손실 함수를 계산한 후 확률적 경사 하강(Stochastic Gradient Descent)을 통해 학습한다. 이후 에이전트(agent)는 탐색(exploration)을 진행하여 새로운 경험(experience)을 얻고 이를 다시 재현 버퍼에 저장하여, 다음 표본 추출에 활용한다.

DQN은 Atari에서는 준수한 성능을 보였지만, 더 복잡하고 어려운 OpenAI Gym 등의 환경에서는 성능이 떨어지는 현상을 보였다. 가장 주요한 원인으로 Q-값 과추정 문제가 꼽힌다. Q-값이 과추정되면, 해당 상태-행동(state-action)으로 쏠리게 되고 해당 상태에서는 다른 행동을 탐색하지 않고, 과추정된 행동으로만 활용(exploitation)하려는 경향이 생기며 탐색 갇힘(exploration stuck)이 생겨 학습이 저하된다.

이를 해결하기 위해서, 여러 개의 Q-신경망을 두고 이를 평균화하여 한 신경망의 Q-값이 과추정되더라도 나머지 Q-값이 이를 완화시켜주는 앙상블 DQN 이 등장하였다. DQN에서 1개의 Q-신경망을 쓰는 것과 달리, 앙상블 DQN에서는 별도의 신경망 매개변수를 가진 여러 개의 Q-신경망을 사용하여, 목표 Q -값을 추정할 때 여러 Q-신경망을 이용해 추정한 목표 Q-값들의 평균을 취해 최종 목표 Q-값을 추정한다.

본 논문에서는, 앙상블 DQN을 기반으로 탐색을 강화한 분할 배치 앙상블 DQN을 제안한다. 앙상블 DQN은 Q-신경망 학습을 할 때, 전체 재현 버퍼에서 동일한 배치 1개를 표본 추출하여, 모든 Q-신경망이 같은 배치를 사용했다. 분할 배치 앙상블 DQN에서는 각각의 Q-신경망마다 서로 다른 배치를 쓰도록 하였다. 즉, k개의 Q-신경망의 매개변수를 학습할 때, 전체 재현 버퍼에서 똑같이 k개의 미니 배치를 표본 추출하여, 하나의 신경망 매개변수 당 하나의 미니 배치가 할당되어 학습이 진행되도록 하였다.

Algorithm 3 Separated batch Ensemble DQN

Initialize K Q-networks $Q(s,a;\theta^k)$ with random weights θ_0^k for $k \in \{1,\dots,K\}$
Initialize Experience Replay (ER) buffer $\mathcal{B}$
Initialize exploration procedure *Explore*$(\cdot)$
for $i = 1,2,\dots,N$ **do**
  $Q_{i-1}^E(s,a) = \frac{1}{K}\sum_{k=1}^{K} Q(s,a;\theta_{i-1}^k)$
  for $k = 1,2,\dots,K$ **do**
    Sampling mini-batch $\mathcal{B}_k$ from buffer $\mathcal{B}$
    $y_{s,a}^{i,k} = \mathbb{E}_{\mathcal{B}_k}\left[r + \gamma \max_{a'} Q_{i-1}^E(s',a'))|\, s,a\right]$
    $\theta_i^k \approx \text{argmin}_\theta\ \mathbb{E}_{\mathcal{B}_k}\left[(y_{s,a}^{i,k} - Q(s,a;\theta))^2\right]$
  end for
  Explore$(\cdot)$, update $\mathcal{B}$
end for
output $Q_N^E(s,a) = \frac{1}{K}\sum_{k=1}^{K} Q(s,a;\theta_i^k)$

그림 1. 제안하는 알고리즘 의사코드

Ⅲ. 실험 결과 및 결론

이번 프로젝트에서는 OpenAI Gym 환경 중 Cartpole-v1과 Acrobot-v1을 사용하였다. Cartpole-v1은 수레 위에 연결된 막대기를 쓰러트리지 않는 것을 목적으로 하는 환경으로, 전통적으로 DQN 기반 알고리즘에서 성능이 잘 나오지 않는 것으로 잘 알려진 환경이기 때문에 선택하였다. Acrobot-v1 환경은 두개의 관절로 연결된 팔들의 토크를 조절하여, 기준선 위로 세우는 것을 목적으로 하는 환경이다. Cartpole-v1보다 더 복잡한 상태 공간(state space)을 가지고 있기 때문에 선택했다.

두가지 환경에 DQN, 앙상블 DQN, 분할 배치 앙상블 DQN 세가지 알고리즘을 각각 적용하여 실험하였다. 앙상블 DQN과 분할 배치 앙상블 DQN 모두 10개의 Q-신경망을 사용하였고, 배치 크기는 64로 실험하였다. 분할 배치 앙상블 DQN은 Q-신경망과 같은 개수로 10개의 배치를 표본 추출하였다. 그림 2,3에 두 환경 모두에 대해 10개 난수 시드(random seed)에 대한 500 에피소드(episode)까지의 평균 이득(mean return)을 나타냈다.

Cartpole-v1과 Acrobot-v1 모두 분할 배치 앙상블 DQN이 가장 좋은 성능을 보였다. (그림 2,3) 학습 초기에는 재현 버퍼에 들어있는 표본 수가 적어서, 모든 Q-신경망의 배치가 거의 동일해진다. 따라서 기존 앙상블 DQN와 거의 비슷한 학습 수준을 보인다. 하지만 학습이 진행되며 재현 버퍼의 크기가 늘어나면서, 각각의 배치가 다양해지고, 탐색이 강화되며 더 좋은 성능을 보여주었다. 또한 Cartpole-v1보다는 Acrobot-v1에서 다른 알고리즘과의 학습 차이가 더 큰 양상을 보였다. Cartpole-v1에서는 150 에피소드가 넘어가면서 성능이 올라오기 시작하였지만, Acrobot-v1에서는 학습 초반부터 다른 알고리즘보다 더 높은 평균 이득을 보여주었다. 상태 공간이 크기에, 더 많은 탐색을 필요로 하는 Acrobot-v1 환경에서 분할 배치 앙상블 DQN이 더 월등한 성능을 낸 것으로 분석된다.

본 논문에서는 앙상블 DQN을 기반으로, 탐색을 더 강화하는 분할 배치 앙상블 DQN을 설계하여 실험을 진행하였다. 탐색을 강화하고 과추정을 줄여주는 효과를 주기 위해, Q-신경망 각각에 하나의 배치를 할당하는 방식으로 설계하였다. 실험 결과 Cartpole-v1 환경에서 성능 향상을 볼 수 있었고, 특히 상태 공간이 복잡한 Acrobot-v1에서 월등한 성능을 보여주었다.

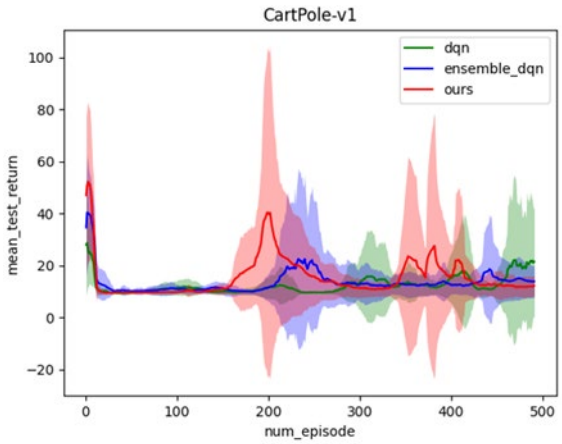

그림 2. CartPole-v1 의 실험 결과

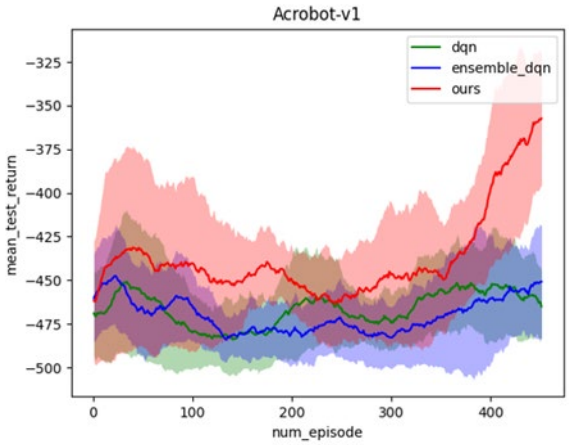

Figure 3. Acrobot-v1 의 실험 결과

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] Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou, Ioannis, Wierstra, Daan, and Riedmiller, Martin. Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.

[2] Anschel, Oron, Nir Baram, and Nahum Shimkin. "Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning." International conference on machine learning. PMLR, 2017.