Deep Q-Learning from Demonstrations

2023710158 박경빈

기존의 Deep Reinforcement Learning 은 복잡하고 어려운 문제에서 의미 있고 좋은 결과를 도출해냈습니다. 하지만 이러한 방법은 의미 있는 결과를 얻어내기까지 큰 학습데이터 양을 필요로 한다는 문제를 가지고 있었습니다. 또한 학습 초기에는 성능이 별로 좋지 않았습니다. 이를 해결하기 위해 Deep Q-Learning from Demonstrations (DQfD)가 나오게 되었고 간단하게 설명하면 **Demonstration Data 로 Pretrain** 을 진행하여 학습 초기의 성능을 올리는 방법입니다. 이후 Environment 와 interaction 하면서 학습을 진행합니다

이 논문의 핵심 개념은 **'데모(Demo)**'입니다. 데모는 사람이나 다른 전문가로부터 얻은 경험 데이터입니다. 이러한 데모 데이터를 사용하여 DQN 이 초기 학습을 보다 안정적으로 시작하고, 효율적으로 보상을 최대화하도록 학습을 진행할 수 있습니다.

논문에서 제안하는 Deep Q-Learning from Demonstrations(DQfD) 알고리즘은 두 가지 주요 단계로 구성됩니다. 첫 번째 단계는 전문가의 행동에 대한 데모 데이터를 사용하여 **DQN 을 사전에 학습**시키는 것입니다. 이 단계에서는 데모 데이터와 DQN 의 출력을 최대한 일치시키는 방향으로 네트워크를 조정하여 초기 학습을 돕습니다.

두 번째 단계는 DQN 을 **데모 데이터와 함께 학습시키는 과정**입니다. DQN 은 현재의 경험과 데모 데이터를 기반으로 행동을 선택하고, 이로부터 얻은 데이터로 네트워크를 업데이트합니다. 이를 통해 DQN 은 데모 데이터에서 배울 수 있는 정보를 활용하면서도 새로운 경험을 통해 정책을 계속 개선할 수 있습니다.

위의 과정을 진행하면서 Loss 함수를 계산하게 되는데 논문에서 제안한 Loss 함수는 다음과 같습니다.

$$J(Q) = J_{DQ}(Q) + \lambda_1 J_n(Q) + \lambda_2 J_E(Q) + \lambda_3 J_{L2}(Q)$$

- **One Step Loss :** 1 step Double DQN Loss

$$J_{DQ}(Q) = \left(R(s,a) + \gamma Q(s_{t+1}, a_{t+1}^{\max}; \theta') - Q(s,a;\theta)\right)^2$$

- **N Step Loss :** n step Double DQN Loss

$$J_n(Q) = \mathbb{E}_\pi \left[\left(r_{t+1} + \cdots + \gamma^{n-1} r_{t+n} + \max_a \gamma^n Q(s_{t+n}, a; \theta') - Q(s_t, a; \theta) \right)^2 \right]$$

- **Supervised Loss** : Demo 이후의 학습에서도 Demo 에서의 action 과 같은 action 을 하도록 하는 large margin classification loss 입니다.

$$J_E(Q) = \max_{a \in A} [Q(s, a) + l(a_E, a)] - Q(s, a_E)$$

- **L2 Regularization Loss** : L2 Regularization 을 이용해서 과적합을 방지합니다.

단, self-generated Data 에서는 λ2 = 0 으로 두어 supervised loss 는 반영하지 않습니다.

이러한 loss 함수이외에도 **Prioritized Experience Replay (PER)** 알고리즘을 사용하고 있습니다. PER 이란 경험 **ReplayMemory** 에서 중요한 경험을 더 자주 **sampling** 하여 학습에 활용하는 것입니다. 기존의 경험 리플레이 방법은 모든 경험을 동등한 가중치로 샘플링 하여 사용하였지만 중요한 경험에 더 많은 가중치를 부여하여 학습의 효율성을 높였습니다. 이를 DQfD 알고리즘에 적용하여 학습의 효율성과 안정성을 개선하였습니다.

DQfD 는 Imitation Learning 과 PDD DQN 보다 더 **우수한 성능**을 이룰 수 있었으며 Deep Q-learning form Demonstrations(DQfD)라는 새로운 deep 강화학습 알고리즘을 제안하고, 이 알고리즘이 적은 양의 **인간 데모 데이터**를 활용하여 학습 효율성을 크게 향상시킬 수 있다는 것을 실험적으로 입증하였습니다.