Reinforcement Learning from Imperfect Demonstrations

Gao, Yang, et al
International Conference on Machine Learning(2018)

IIE8557-01 동적계획법과 강화학습
경영과학연구실 김변민

Imperfect demonstration

- Demonstration 값은 항상 정확하지 않을 수 있음
- Demonstration 이 실제 모든 action을 수행할 수 있는 것은 아님

가정 : Action 1,2,3 중 시연자는 액션 1 만을 선택할 수 있음

Q(s, a_1) = 10
Q(s, a_2) = 1
Q(s, a_3) = 3

1. Q(s, a_2) 를 업데이트

2. rollout에서의 Max Q값은 Q(s^1, a_1)

3. 시연에 없는 action2 가 Max Q 값으로 인해 과도하게 높아질 수 있음

Demonstration이 imperfect 한 상황을 극복할 수 있는 기법을 개발

Key Idea

- Normalized Actor-critic 방법을 통해 imperfect demonstration을 극복

  -> 시연 데이터에서는 없는(혹은 좋지 않은) action 의 Q 값이 증가하는 것을 방지함

  -> 항상 최적의 시연데이터가 필요한 것이 아니기 때문에 많은 시연데이터에 대해 robust한 학습이 가능함

- Soft value function 을 적용 함으로서 다양한 action을 시도할 확률을 높이고 시연되지 않은 action이 뽑힐 확률을 줄임

Overall Process

For step t ∈ {1,2…}
  if t <= k
    Sample a mini-batch of transitions from D (demonstration buffer) → 시연데이터 샘플링
  else
    Start from s, sample a from π, execute a, observe
    (s',r) and store (s,a,r, s') in M → 일반데이터 샘플링
  end
  Update θ with gradient
  if t mod T =0 then
    θ' <- θ → 타겟 업데이트
  end
end

Maximum entropy reinforcement learning

- 최적의 정책은 미래의 보상을 최대화하는 동시에 행동 분포의 미래 엔트로피도 최대화함

- Action의 분포가 한쪽으로 치우칠 수록 엔트로피는 감소하고 균일할 수록 증가함

$$\pi_{ent} = \underset{\pi}{\operatorname{argmax}} \sum_{t} \gamma^t \underset{s_t, a_t \sim \pi}{\mathbb{E}} [R_t + \alpha H(\pi(\cdot|s_t))]$$

Entropy

Soft value function

- Maximum entropy 강화학습 패러다임으로 가치함수의 정의도 자연스레 변경 됨

- 이는 에이전트가 여러 가능한 행동을 탐색하고 다양한 전략을 학습하는데 도움을 줌

$$Q_\pi(s,a) = R_0 + \mathop{\mathbb{E}}_{(s_t,a_t)\sim\pi} \sum_{t=1}^{\infty} \gamma^t (R_t + \underbrace{\alpha H(\pi(\cdot|s_t))}_{\text{Entropy}}) \tag{3}$$

$$V_\pi(s) = \mathop{\mathbb{E}}_{(s_t,a_t)\sim\pi} \sum_{t=0}^{\infty} \gamma^t (R_t + \underbrace{\alpha H(\pi(\cdot|s_t))}_{\text{Entropy}}) \tag{4}$$

Soft value function

- 가치함수는 Q 값의 가중평균으로 계산 됨 (5)
  -> 큰 Q value 를 가진 action에 더 큰 가중치가 부여됨

- Q value 값이 해당 state의 Q값들의 가중평균보다 클수록 선택될 확률이 높아짐 (6)

$$V^*(s) = \alpha \log \sum_a \exp(Q^*(s,a)/\alpha) \qquad (5)$$

$$\pi^*(a|s) = \exp((Q^*(s,a) - V^*(s))/\alpha) \qquad (6)$$

Normalized actor critic

- 시연에 없던 action은 Max Q로 인해 Q값이 급격하게 증가할 수 있음

- V_Q(s) 항을 통해 시연되지 않은 action 의 Q 값이 급격히 증가하는 것을 방지함

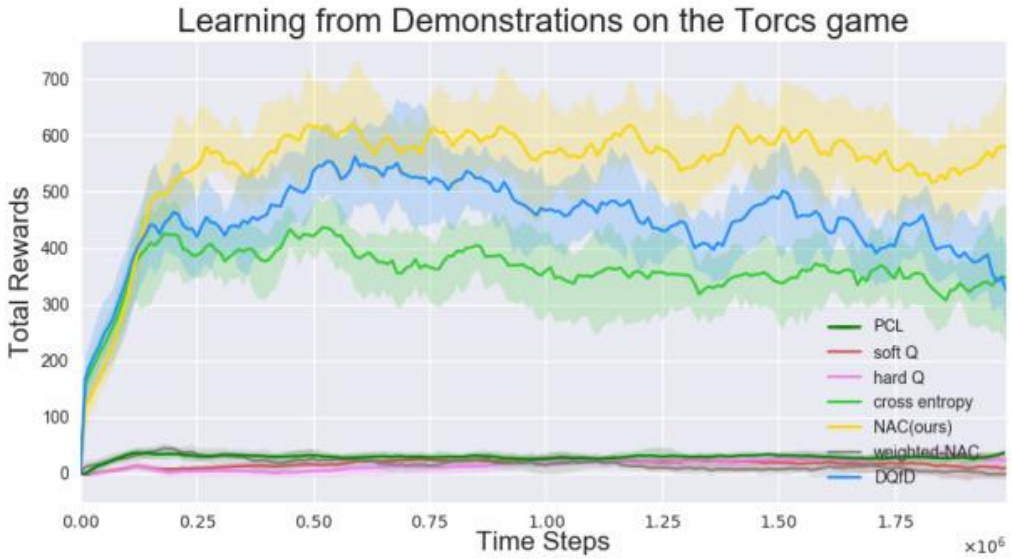
$$\nabla_\theta J_{PG} = \mathop{\mathbb{E}}_{s,a\sim\pi_Q}\left[(\nabla_\theta Q(s,a) - \nabla_\theta V_Q(s))(Q(s,a) - \hat{Q})\right]$$

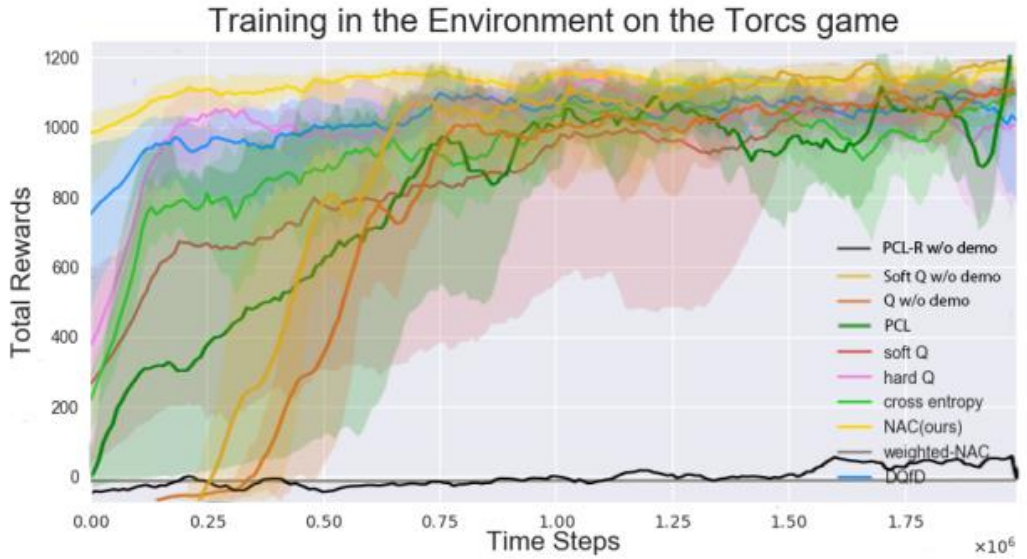
$$V_Q(s) = \alpha \log \sum_a \exp(Q(s,a)/\alpha) \qquad (11)$$

$$\pi_Q(a|s) = \exp((Q(s,a) - V_Q(s))/\alpha) \qquad (12)$$

Torcs game (racing game)

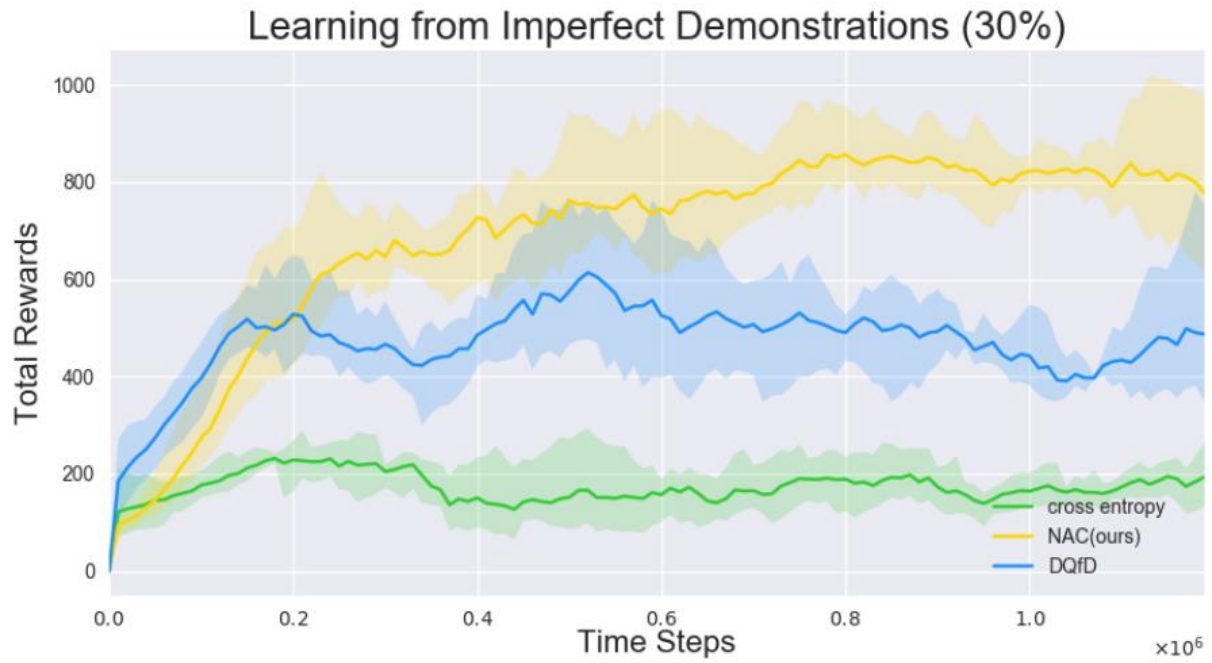
- 충돌을 피하면서 빠르게 주행해야 하는 레이싱 게임
- Demonstration을 이용한 학습과 그 이후 일반 학습에서 모두 제안한 방법이 가장 좋은 성능을 냄





Torcs game (racing game) with imperfect demonstration

- Optimal action을 70 % 만 맞추는 demonstration을 이용한 실험을 진행
- Imperfect demonstration 상황에서 특히 제안한 방법이 가장 우수한성능을 보임



Conclusions

- 최적이 아닌 시연을 극복할 수 있는 알고리즘을 최초로 제안

- 해당 알고리즘을 이용했을 때 특히 최적이 아닌 시연데이터에서 명확한 성능차이를 보임

Q & A