Transfer Learning in Deep Reinforcement Learning : A Survey

Presented by Byun Min Kim

Introduction

Transfer Learning 이란 ?

- Source domain 을 풀기 위해 학습된 모델만 있는 상황에서, 유사하지만 다른 domain 으로 상이한 task 를 풀기 위해 기존 모델이 학습한 지식을 전이하는 방법론

Transfer Learning 이 필요한 이유

- 강화학습에서는 환경과 상호작용하는 방식으로 샘플을 생성함
- 부분관찰만이 가능한 환경 혹은 상태와 행동 공간의 복잡성 할 때 충분한 양의 샘플이 필요하지만 이를 얻는 것이 어려울 수 있음
- Transfer learning 을 이용하면 실제로 필요한 샘플의 양을 줄일 수 있음
- 결국 Transfer learning을 이용하면 학습기간을 줄일 수 있음

Transfer Learning 이 어려운 이유

- 강화학습을 더 복잡하게 만들 수 있으며 MDP 요소들의 여러가지 형태에 맞게 전문가의 지식을 전달해야 하므로 적절한 전달 방식을 찾는 것이 어려울 수 있음

Transfer learning 을 적용할 때 필요한 질문

1. 어떤 지식이 전이되는지 ?

- 전문가의 시연, 전문가의 정책, 전문가의 가치함수, 전문가 행동의 확률분포, 잠재함수

2. 어떤 RL 프레임워크가 transfer learning 에 적합한가 ?

- Transfer learning 의 종류에 따라 적용가능한 RL 프레임워크가 다름
- Learning from demonstration 기법에서는 모든 프레임워크에 적용이 가능하지만 Policy transfer 기법은 DQN과 같은 구조에 부적합할 수 있음

3. Target domain 과 Source domain 에는 어떠한 차이가 있는가 ?

- Source domain(전문가 환경)과 Target domain (agent의 환경)은 같을 수 도 있고 다른 경우도 존재함
- 다를 경우 상태 공간과 보상의 분포가 다를 수도 있음

Transfer learning 을 적용할 때 필요한 질문

4. 타겟 도메인에서 활용 가능한 정보는 무엇인가?

- Source domain에서의 지식은 대부분 활용 가능하지만 Target domain에서의 sampling은 금지 되거나 보상이 지연될 수 있음
- 예시로 자율주행 에이전트 학습을 위해 실제 환경의 데이터를 이용하는 경우 원하는 만큼의 sampling 이 불가능함

5. 샘플 효율성이 어느 정도로 높아야 하는 전이학습을 사용할 것인가?

- 샘플링 비용을 기준으로 다음과 같이 분류 가능함

Zero-shot transfer

∧

Few-shot transfer

∧

Sample efficient transfer

새로운 평가 지표

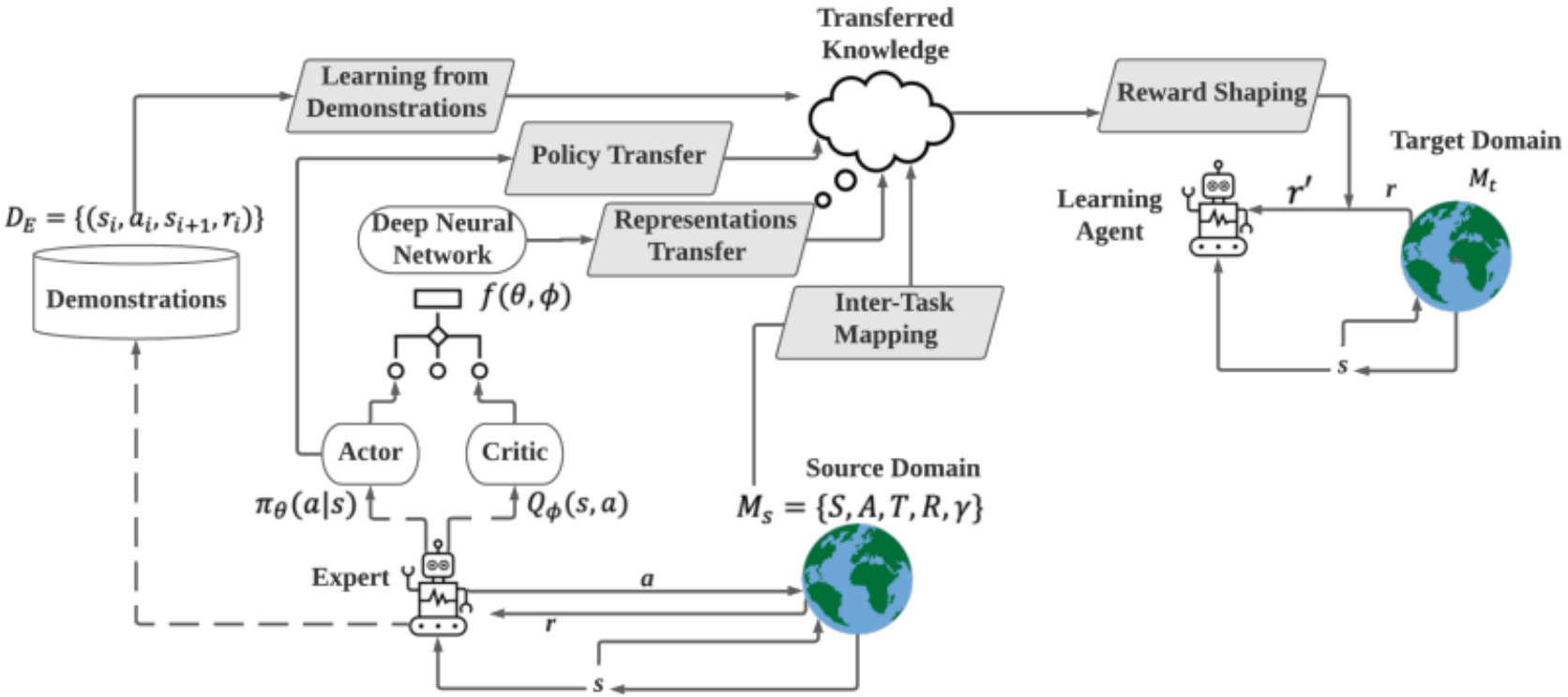
Transfer 할 지식의 양

- 특정 성능의 임계값을 달성하기 위해 전이 학습에 필요한 지식의 양
- 예시 : 전문가가 시연한 샘플 개수, Source domain의 개수

Transfer 할 지식의 품질

- 효과적인 전이 학습을 가능하게 하기 위해 필요한 지식의 품질
- 최적이 아닌 지식이 주어졌음에도 효과적인 전이학습이 가능하다면 효율적인 것으로 볼 수 있음

Overall Method



1. Reward shaping
2. Learning from demonstration
3. Policy transfer
4. Inter-Task mapping

Reward shaping

Reward shaping 이란 ?

- 외부 지식을 활용하여 타겟 도메인의 보상을 재구성하는 기법

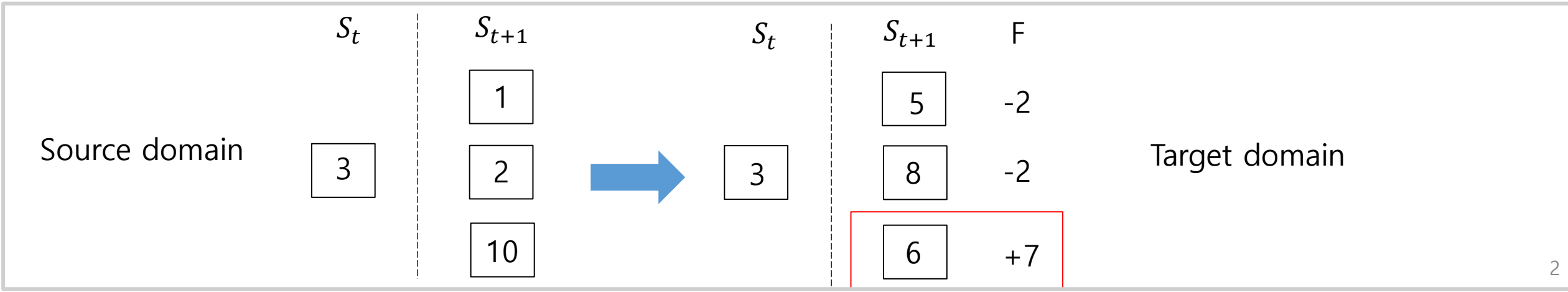
Reward 형태

$\mathcal{R}' : \mathcal{R}' = \mathcal{R} + \mathcal{F}$ → 전문가 지식을 활용하여 계산된 F 가 추가됨

$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathcal{R}) \rightarrow \mathcal{M}' = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathcal{R}')$.

Ex) PBRS(Potential based Reward Shaping)

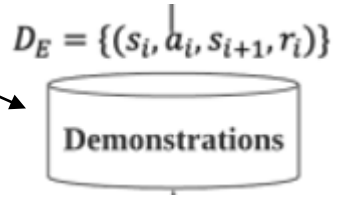
$F(s, a, s') = \gamma\Phi(s') - \Phi(s)$ 포텐셜 함수Φ(·) : 전문가로부터 얻은 상태나 행동의 가치를 평가하는 함수



Learning from demonstration

Learning from demonstration이란?

- 효율적인 학습을 위해 전문가의 시연을 활용하여 강화학습을 지원하는 기법



Offline vs Online

- 전문가의 시연을 언제 지식전이에 사용하는지에 따라 Offline과 Online 으로 나뉨
- Offline 방식에서는 모든 전문가 시연 데이터가 실제 학습전에 수집 됨
- Online 방식에서는 모델이 학습되는 동안 실시간으로 시연 데이터가 수집되면서 활용 됨

Example of Learning from demonstration

DPID(direct policy iteration with demonstration algorithm)

1. 전문가의 샘플 데이터 D_E와 agent의 샘플 데이터 D_π 를 얻음
2. Q value 업데이트를 위해 D_E 와 D_π 가 각각 일정한 확률로 이용됨 (MC 방식으로 업데이트)

DQFD(Deep Q learning from demonstration)

1. 전문가 정책으로 얻은 샘플 데이터 D_E와 agent 정책으로 얻은 샘플 데이터 D_π 를 저장
2. 각각의 버퍼에 D_E 와 D_π 를 저장
3. 일정한 확률로 각 버퍼의 샘플들을 뽑은 후 Q-learning 방식에 기반하여 파라미터 업데이트

LFDS(Learning from demonstrations with Reward shaping)

1. 판별자가 전문가의 샘플 데이터와 agent 의 샘플 데이터를 입력으로 받음
2. 판별자가 전문가의 샘플 데이터인지, agent의 샘플 데이터인지 구분함
3. 판별자가 구분을 할 수 있다면 reward에 Penalty 를 부여하고 반대면 추가 보상을 부여함

Learning from demonstration 한계

부정확한 전문가 시연 데이터

- 전문가의 시연 데이터가 좋은 데이터가 아니라면 성능이 더 나빠질 수 있음
- 한 가지 해결방법은 전문가 시연 데이터를 초기 학습 단계만을 향상시키기 위해 사용하는 것임

부족한 수의 전문가 시연 데이터

- 제한된 수의 전문가 시연 데이터를 사용하여 모델을 학습시킬 경우, 모델이 편향적으로 학습될 수 있음
- 한 가지 해결방법은 시간이 지남에 따라 전문가 시연 데이터의 사용을 줄이는 것임

Policy transfer

Policy transfer 란 ?

- 하나 이상의 Source domain 에서 사전 훈련 된 정책을 활용한 transfer learning 기법

EX) Policy reuse

1. Target domain 환경에서 사전 훈련된 Source domain 의 정책을 통해 에피소드 진행
2. Target domain agent의 정책 혹은 Q_value 업데이트
3. 해당 Source domain의 가중치가 평가됨
4. 해당 가중치를 이용해 Softmax 방식으로 하나의 Source domain을 선택함
5. 다시 1번 반복

한계점

- 매 에피소드마다 Source domain의 가중치를 계속 계산해야 해서 학습속도가 느릴 수 있음

Inter-Task Mapping

Inter-Task Mapping 이란?

- Source domain과 Target domain 이 다른 형태일 때 매핑 함수를 사용하여 지식 전달을 지원하는 Transfer Learning 기법

EX) A deep convolutional encoder – decoder architecture for image segmentation,2017

- 인코더-디코더 구조를 활용
- Encoder Input : Source Domain의 상태
- Encoder Output : 압축 된 형태의 벡터
- Decoder Output : Target Domain에 사용될 새로운 상태
- 매핑 된 Source Domain에서의 정책을 따라할 수 있도록 reward shaping 진행

Unresolve Question

- Source domain 에서 좋은 action이 다른 보상함수를 갖는 Target domain 에서는 안 좋게 작용할 수 있음

  이처럼 도메인 간 보상함수의 극단적인 차이가 있을 때 좋은 성능을 내는 Transfer learning 이 가능한가 ?

- 전문가 시연 데이터가 부족하거나 부정확할 경우 좋은 성능을 내는 Transfer learning 이 가능한가?

- 상태공간과 행동공간이 모두 다를 때 좋은 성능을 내는 Transfer learning 이 가능한가?

Thank you