1. 어떻게 MDP를 적용할 수 있는가?
2. 

자동차 운전의 예시를 기억하십니까?

PPT :이 사진 자체로는 마코브 상태가 아니지만, 그룹을 묶어 평가하면 마코브 상태에 가까워졌었음. 비슷한 아이디어가 사용됨.

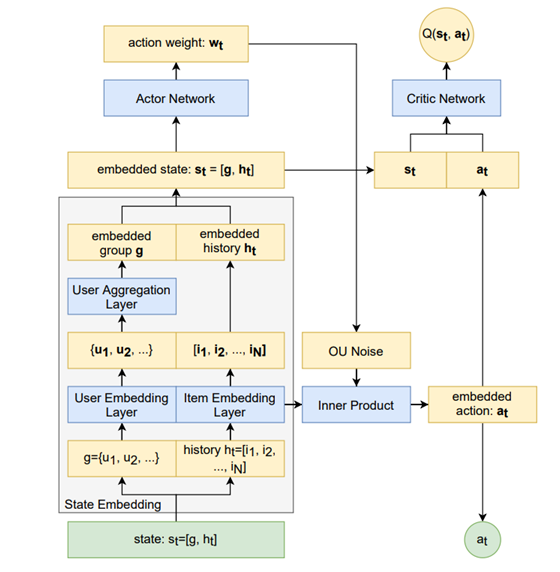
먼저 유저를 그룹으로 묶고, 그 그룹이 본 영화를 상태로 정의합니다.

액션은 에이전트가 그룹에게 추천할 영화인데, 이론적으로는 모든 영화가 추천목록에 들어갑니다.하지만 에이전트가 그룹에게 추천할 영화를 뽑아주도록 학습하는 것이 이 모델의 목표입니다.

보상의 경우 그룹이 그 영화를 선택하는 1 그외에는 0이 됩니다.

전이확률 p(s(t+1)/st,at)이며 마코브 property를 만족합니다 .상태 s에는 그 그룹이 영화를 평가한 history가 담겨있고, 그것을 다 묶어서 하나의 현재로 정의했습니다. 그러므로 현재 상태와 action을 더해 추가 상태를 만드는 것은 이상하지 않습니다.

2. DDQN를 사용하고 Actor를 policy-gradient로, critic network를 벨만방정식, q-learning을 통해 학습한다. DDQN을 사용함으로서, Deterministic policy를 가정하기 전의 수식에서는 at+1을 골랐던 순간의 policy로 Q에 대한 Expection을 원래 구해야하기 때문에 off-policy가 아니지만, Determinsitic policy를 가정한다면 update 할 당시의 policy로 at+1를 구할 수 있기 때문에 off-policy가 됩니다. 평점 시스템에서는 내가 평가한 평점이 평점 시스템에 영향을 주지 않고 추천 영화를 반환하여 좋은 영화를 반환할 수 있게 되겠습니다.



아키텍쳐입니다.

Embedding을 거치면서, 데이터를 actor-network에 맞게 embedding하고->자세하게 말하면 아래:

policy를 탐험함에 따라 environment에서 생성되며, replay buffer에 (현재 상태, 현재 취한 행동, 보상, 다음 상태)의 튜플 형식으로 저장되며 이를 trainsition이라고 부른다. 매 timestep마다 actor와 critic은 buffer로부터 minibatch형태로 sample들을 받아 업데이트 합니다.

데이터는

actor network는 policy gradient 방식으로 학습되는데, actor는 ou 노이즈를 활용하여, i(j)값과 내적하여 그중 가장 값이 큰 것을 action으로 취해져, action이 조건을 만족하면 group에 추천되도록 학습될 것입니다.

critic network는 Q-learning에서와 같이 벨만 방정식을 사용하여 학습(=파라미터 업데이트)됨, 해당 논문에서는 Deep Q-learning을 사용할 것입니다.