# 잠재 공간에서 Belief State를 이용해 미래 상태를 예측하는 모델 기반 강화학습에 관한 연구

권세이, 최계원
성균관대학교
say0423@g.skku.edu, kaewonchoi@skku.edu

# A Study on Model-Based Reinforcement Learning Imagination Using Belief State in Latent Space

Se I Kwon, Kae Won Choi
Sungkyunkwan Univ.

요 약

모델 기반 강화학습은 환경을 모사한 모델을 사용하여 최적 정책을 학습하는 것이다. Dreamer[1]는 모델 기반 강화학습과 월드 모델을 사용하고 잠재 공간에서 미래 상태를 예측하는 것으로 학습이 잘 된다는 것을 보여줬다. 본 논문에서는 현재 상태에 대한 믿음 상태 belief state만으로 잠재 공간에서 미래 상태를 예측하는 모델 기반 강화학습을 제안한다. Dreamer[1]는 RSSM을 사용하여 믿음 상태 없이 모델의 현재 상태와 미래 상태를 잠재 공간에서 예측하였다. 우리 모델은 상태를 나타내는 state와 현재 상태에 대한 믿음을 나타내는 belief state를 구별하고, 현재 상태와 미래 상태를 belief state에 기반하여 추론한다. Imagination test를 통해 belief state만으로 사람이 생각하는 것과 유사하게 월드모델도 미래 상태를 예측하고 있음을 보여준다.

## Ⅰ. 서 론

최근 강화학습이 자율주행, 로봇, 통신 패킷의 네트워킹 등 다양한 분야에 적용되면서 관심을 받고 있다. 강화학습은 행동심리학에서 영감을 받은 기계 학습의 한 영역이다. 강화학습은 기존의 지도학습과 달리 라벨링 된 데이터 없이 학습할 수 있다는 것이 특징이다. 학습과 의사결정의 주체자를 에이전트라고 하고, 에이전트 이외의 모든 것을 환경이라고 한다. 에이전트는 시행착오를 거듭하면서 역학관계를 모르는 환경과 상호 작용하고 보상을 받는다. 에이전트는 훈련을 통해서 기대되는 보상을 최적화하는 행동 정책을 학습한다.

강화학습은 모델 프리 강화학습과 모델 기반 강화학습으로 나눌 수 있다. 모델 프리 강화학습은 에이전트가 환경과 직접 상호작용하여 최적 정책을 학습하는 것이다. 모델 기반 강화학습은 환경을 모사한 모델을 사용하여 최적 정책을 학습하는 것이다. 모델 기반 강화학습은 환경과 직접 상호작용하는 모델 프리 강화학습에 비해 모델로 학습을 진행하기 때문에 실행 속도가 빠르며 높은 sample 효율성을 가진다. 잘 학습된 모델은 업무가 바뀌더라도 금방 학습하여 동작할 수 있다. 모델은 환경을 모사한 것이기 때문에 환경이 어떻게 변할지 예측할 수 있다. Imagination은 모델을 사용하여 미래 상태를 시뮬레이션하는 것이다. 모델 기반 강화학습에서 이러한 Imagination을 통해 정책을 만들고 향상시키는 것을 Planning이라고 한다. 모델 프리 강화학습은 학습하는 동안 환경에 직접 상호작용하면서 정책을 향상시키는 것을 Learning이라고 한다.

Dreamer[1]는 모델 기반 강화학습과 월드 모델을 사용하여 잠재 공간에서의 Imagination 미래 상태를 예측하는 것으로 학습이 잘 된다는 것을 보여줬다. 사람의 뇌는 경험들을 모으고 머릿속에서 추상적인 모델을 만들어 학습한다. 이러한 뇌의 동작 방식과 매우 유사하게 학습을 하는 딥러닝 방법을 월드 모델이라고 한다. Dreamer[1]는 월드 모델로 RSSM(Recurrent State Space Model)을 사용한다. 상태를 확률적인 모델과 결정적인 모델을 결합하여 정의한다. 확률적인 모델만 사용할 경우 여러 단계를 예측하기가 어렵고, 결정적인 모델만 사용할 경우 최적화가 어렵기 때문이다. RSSM을 적용하여 학습이 잘 되는 월드 모델을 구현하였다. Dreamer[1]는 잠재 공간에서 Imagination을 통해 미래 상태를 예측하면서 학습을 진행한다. Imagination의 길이가 길어질수록 예측하기 어려워했다. 하지만 Dremaer[1]에서는 RSSM과 인코더를 통과한 이미지 관측 상태를 결합하여 사용하는지에 대한 설명은 나와있지 않다.

본 논문에서는 POMDP(Partially Observable Markov Decision Process)에 기반한 Latent Space Model을 제안한다. 우리 모델은 잠재 공간에서 현재 상태에 대한 믿음 상태인 belief state를 사용하여 미래 상태를 예측하면서 학습을 진행한다. 현실에서는 항상 부분적인 정보만 제공되기 때문에 POMDP를 기반으로 모델을 구성하였다. POMDP를 가정하면 한 시점의 관측이 실제 관측과 다를 수 있기에 belief state를 사용해야 한다. 우리는 Dreamer[1]와 달리 belief state를 포함한 모델을 구성하여 학습을 진행하고 Imagination test를 통해 Dreamer[1]의 결과와 비교하였다.

## Ⅱ. 본론

### 가. 월드 모델 구조

본 논문에서 제안한 POMDP에 기반하여 Belief state를 포함한 월드 모델의 구조는 그림 1과 같이 구성된다. $s_t$는 t 시점에서 상태, $a_t$는 t 시점에서 행동, $o_t$는 t 시점에서 관측, $r_t$는 t 시점에서 보상, $h_t$는 t 시점에서 RNN state, $b_t$는 t 시점에서 belief state를 나타낸다.

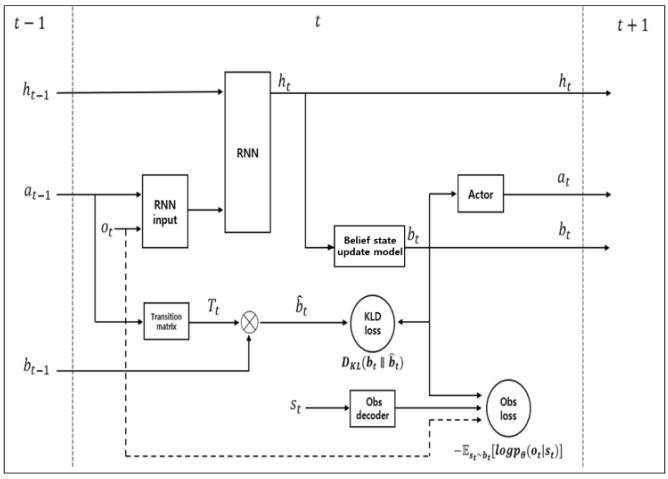


그림 1. 월드 모델 구조

Observation decoder는 입력으로 $b_t$의 분포에서 샘플한 $s_t$가 들어오면 $o_t$의 분포를 출력한다. Transition matrix는 $a_{t-1}$이 입력으로 들어오면 다음 시점의 믿음 상태로 변환할 수 있는 행렬 $T_t$를 출력한다. 그 행렬과 $b_{t-1}$를 곱하면 다음 시점의 믿음상태 $\hat{b}_t$가 나온다. Belief state update model은 입력으로 $b_{t-1}$, $a_{t-1}$, $o_t$이 들어가서 다음 시점의 믿음 상태 $b_t$의 분포를 출력하는 모델이다. 먼저 RNN의 입력으로 $a_{t-1}$, 인코더를 거쳐서 축약된 $o_t$과 $h_{t-1}$를 넣으면 $h_t$가 나온다. $h_t$를 Belief state update model에 넣으면 다음 시점의 믿음 상태 $b_t$의 분포를 얻을 수 있다. Reward model은 입력으로 $b_t$의 분포에서 샘플한 $s_t$를 넣으면 $r_t$의 분포를 출력한다. Actor model은 입력으로 $b_t$를 넣으면 $a_t$의 분포를 출력한다. Value model은 입력으로 $s_t$와 $b_t$를 넣으면 value를 출력한다.

나. 학습 과정

모델의 학습과정은 크게 다섯 부분으로 나뉜다. 첫 번째, 학습을 시작하면 랜덤 행동으로 환경과 상호작용한 결과를 버퍼에 저장한다. 두 번째, 모델을 학습시키기 위해서 저장된 샘플들을 버퍼에서 불러온다. 세 번째, 월드 모델 loss를 backpropagate한다. 네 번째, Imagination을 통해 구한 state와 reward로 actor-critic loss를 backpropagate한다. 다섯 번째, actor model을 사용하여 환경과 상호작용하고 결과를 버퍼에 저장한다. 2~5까지를 하나의 에피소드로 정하고 이를 반복하여 모델을 학습시킨다.

월드 모델 loss는 observation loss, reward loss, kld loss를 더해서 만들어진다. Observation loss는 실제 observation과 현재 belief state 분포에서 샘플한 state를 observation decoder에 넣어서 만든 observation이 얼마나 다른지 나타낸다. Reward loss는 실제 reward와 현재 belief state 분포에서 샘플한 state를 reward model에 넣어서 만든 reward가 얼마나 다른지 나타낸다. Kld loss는 Transition matrix로 만든 $\hat{b}_t$와 Belief state update model에서 나온 $b_t$가 같아지도록 하는 loss이다. Actor-critic loss를 구하기 위해서는 먼저 Imagination을 해야한다. Imagination을 통해서 예측한 state, belief state를 reward model, value model에 넣어서 imag_reward, imag_value를 구한다. 이렇게 예측한 것들로 actor-critic loss를 backpropagate할 수 있다.

다. Imagination 과정

Actor와 valu를 학습하기 위해서는 $b_t$를 이용한 Imagination이 필요하다. Imagination미래 상태를 예측하는 time step의 개수를 horizon이라 하고, 0에서부터 시점 H까지 예측하는 Imagination 과정을 설명한다. 먼저 랜덤하게 $b_0$를 결정한다. $b_0$에 기반하여 $s_0$를 샘플링한다. 시점 t가 0에서부터 H-1이 될 때까지 reward model에 $s_t$를 넣어서 $r_t$를 샘플링하고, actor model에 $b_t$를 넣어서 $a_t$를 샘플링하고, transition model에 $a_t$, $s_t$를 넣어서 $s_{t+1}$을 샘플링하고, $s_{t+1}$을 observation decoder에 넣어서 $o_{t+1}$을 샘플링하고, belief model에 $o_{t+1}$, $a_t$, $h_t$를 넣으면 $b_{t+1}$이 나온다. 이러한 과정을 반복해서 수행하면 Imagination을 통해 모델이 미래 상태를 예측할 수 있다.

라. Imagination test 결과

그림 2는 본 논문에서 제안한 모델의 Imagination test의 결과이고 그림 3은 Dreamer[1]의 Imagination test 결과이다. 환경은 아타리게임의 미니어처 버전이고 horizon은 둘다 10으로 설정하였다. Belief state만을 사용하여 Imagination한 결과가 rssm을 사용하여 Imagination한 결과만큼 모델이 미래 상태를 예측하고 있음을 확인할 수 있다.

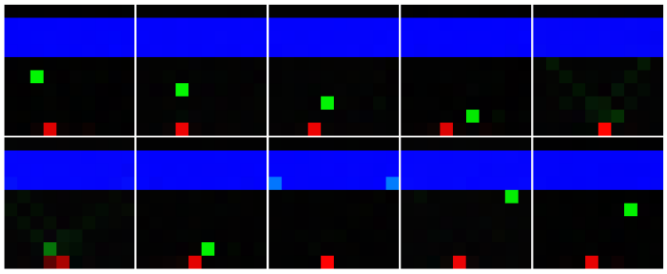

그림 2. Our model Imagination test

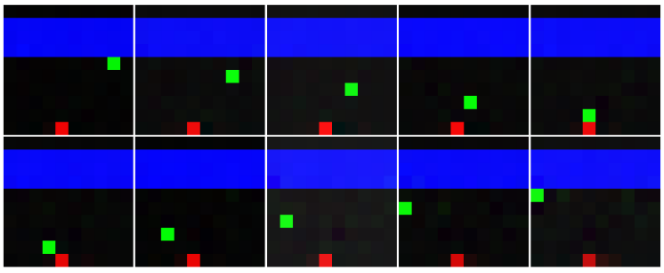

그림 3. Dreamer Imagination test

## Ⅲ. 결론

본 논문에서는 잠재 공간에서 월드 모델을 이용해 Imagination을 수행한다는 점은 Dreamer[1]와 유사하지만, POMDP에 기반하여 현재 상태에 대한 belief state를 추가하였다. 우리 모델은 잠재 공간에서 Imagination을 Belief state에서 추론한 state로 미래 상태를 예측하였다. Imagination test를 확인해본 결과 미래 상태를 예측하는 정확도가 높은 것을 확인할 수 있었다. Belief state만으로 사람이 생각하는 것과 유사하게 월드모델도 미래 상태를 예측하고 있음을 보여준다.

## ACKNOWLEDGEMENT

이 논문은 4단계 BK21 사업의 지원을 받아 수행된 연구임.

## 참 고 문 헌

[1] Hafner, Danijar, et al. "Dream to control: Learning behaviors by latent imagination." arXiv preprint arXiv:1912.01603 (2019).