DMQA Seminar 20231027

Introduction to Exploration in RL

Journey to overcome noisy-TV problem

일반대학원 산업경영공학과
김재훈

고려대학교
KOREA UNIVERSITY

Introduction

- 발표자 소개

- 이름: 김재훈
- 학력
  - ✓ 2020.03 – 현재 | 석박사통합과정 | 고려대학교 산업경영공학과 (지도교수: 김성범)
- 연구분야
  - ✓ Self-supervised learning
  - ✓ Reinforcement learning
- e-mail : jhoon0418@korea.ac.kr

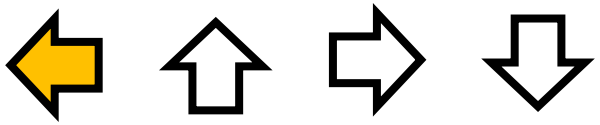
Deep Reinforcement Learning

Objective of RL

❖ 강화학습의 목적

- 강화학습은 **순차적인 의사결정 문제**에서 **누적 보상을 최대화**하기 위해 시행착오를 거쳐서 상황에 따른 행동 정책을 학습
- 강화학습은 에이전트가 속한 상태(state), 선택한 행동(action), 행동에 따른 보상(reward)으로 구성됨

$State$ $Action$ $Reward$



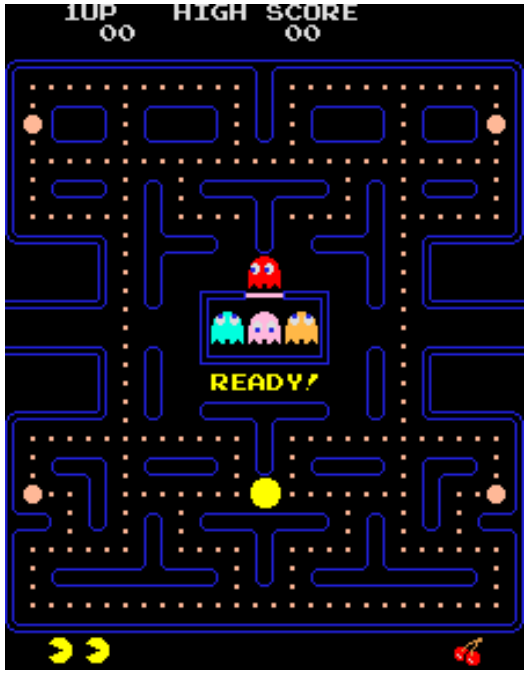
-10 / 0 / 1

Deep Reinforcement Learning

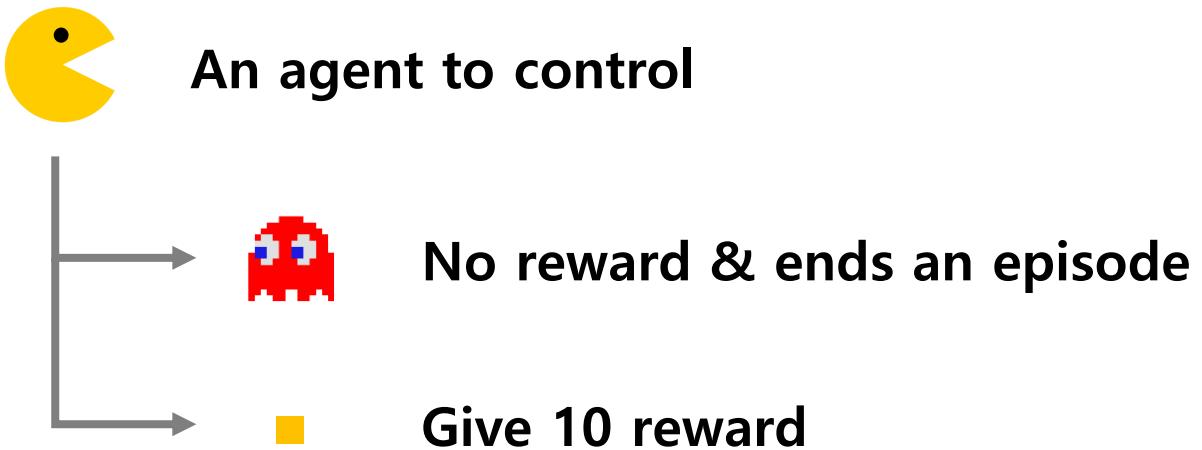
Objective of RL

❖ 예시: PAC-MAN

- 주어진 상태에서 팩맨을 조종하여 유령은 피하되 노란 점을 최대한 많이 수집하는 것이 목표



PAC-MAN

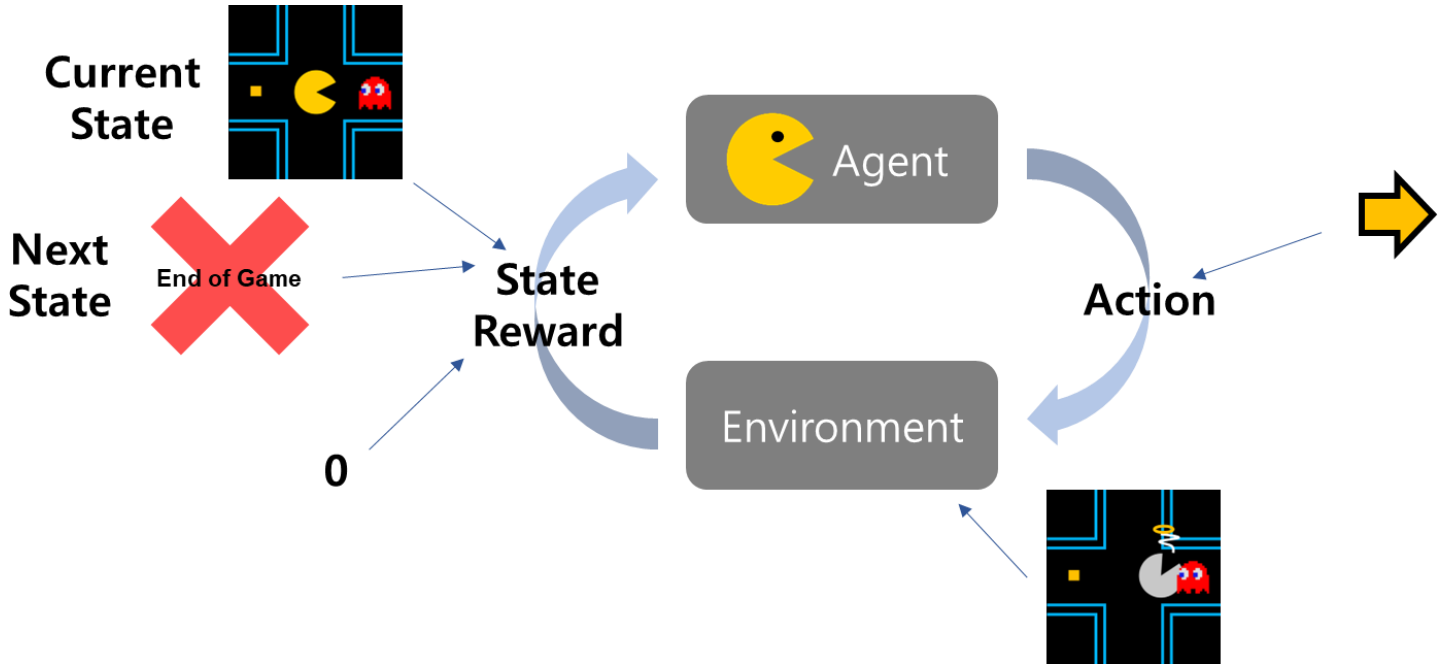


Deep Reinforcement Learning

Objective of RL

❖ 예시: PAC-MAN

- 에이전트를 어떻게 조종하는가에 따라서 달성할 수 있는 최종 점수가 달라짐
- 따라서 에이전트가 좋은 행동 정책을 습득하는 것이 매우 중요함

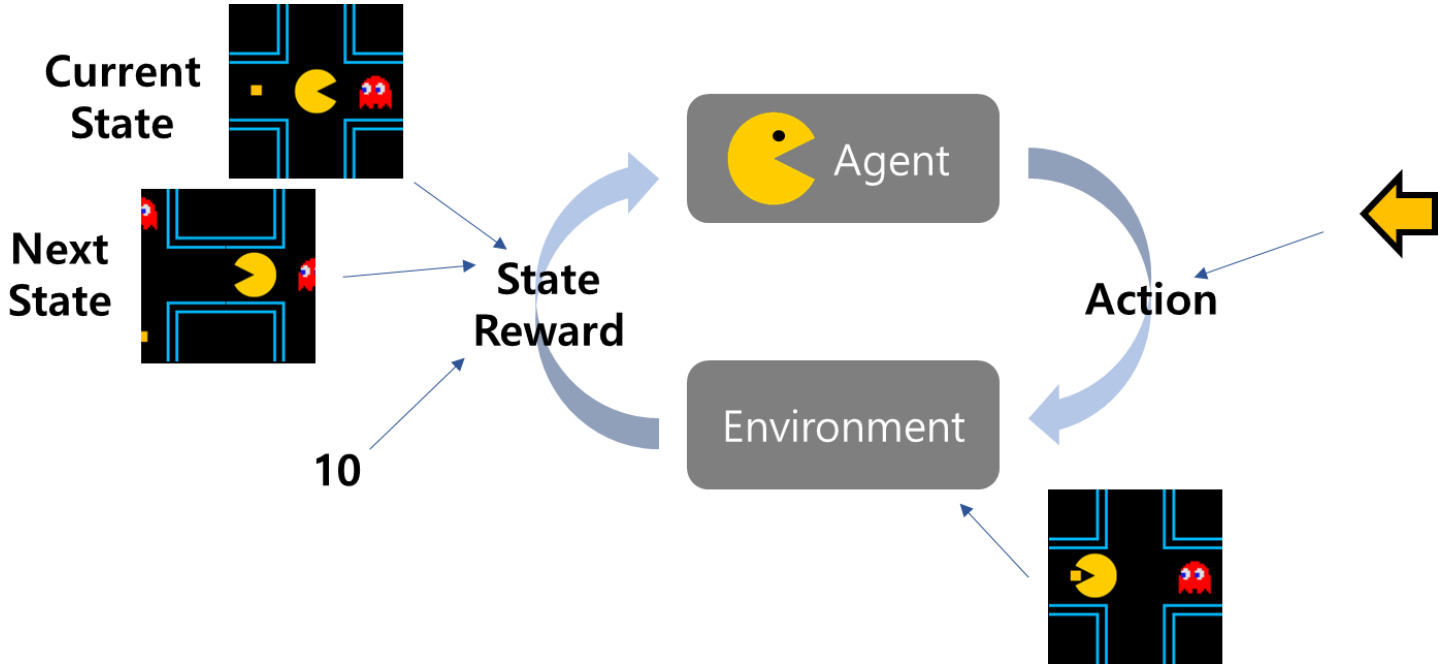


Deep Reinforcement Learning

Objective of RL

❖ 예시: PAC-MAN

- 에이전트를 어떻게 조종하는가에 따라서 달성할 수 있는 최종 점수가 달라짐
- 따라서 에이전트가 좋은 행동 정책을 습득하는 것이 매우 중요함

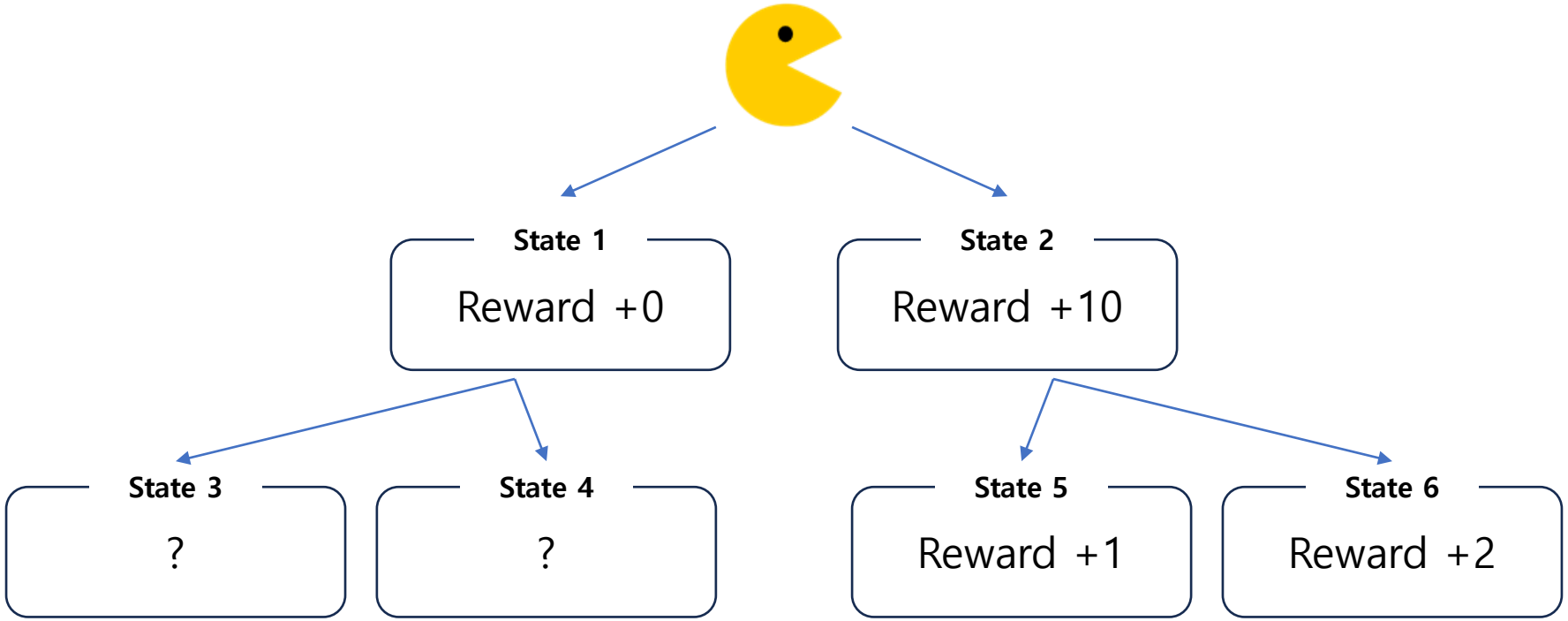


Exploration in RL

Dilemma in RL

❖ Exploration (탐험) vs. Exploitation (착취)

- 학습할 환경을 얼마나 탐험할 것인가는 강화학습의 가장 큰 딜레마 중 하나
- 충분한 탐험을 하지 않는다면 sub-optimal한 정책을 갖기 때문

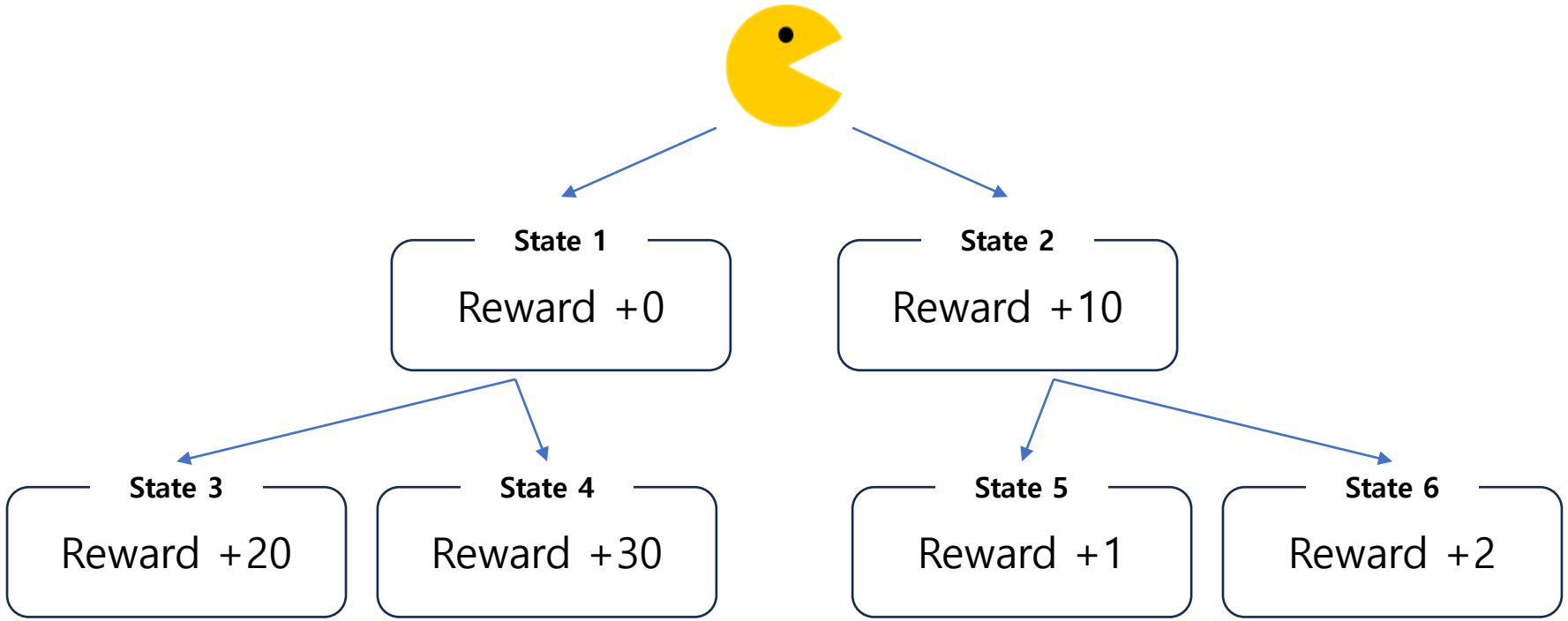


Exploration in RL

Dilemma in RL

❖ Exploration (탐험) vs. Exploitation (착취)

- 학습할 환경을 얼마나 탐험할 것인가는 강화학습의 가장 큰 딜레마 중 하나
- 충분한 탐험을 하지 않는다면 sub-optimal한 정책을 갖기 때문

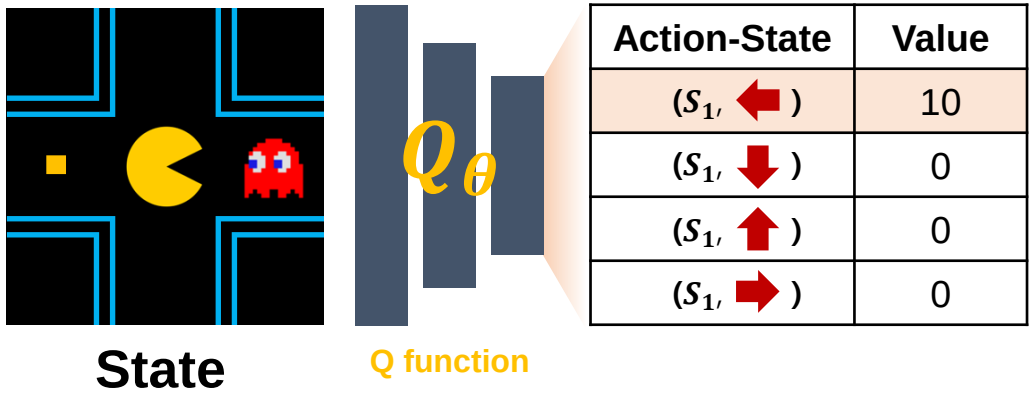


Exploration in RL

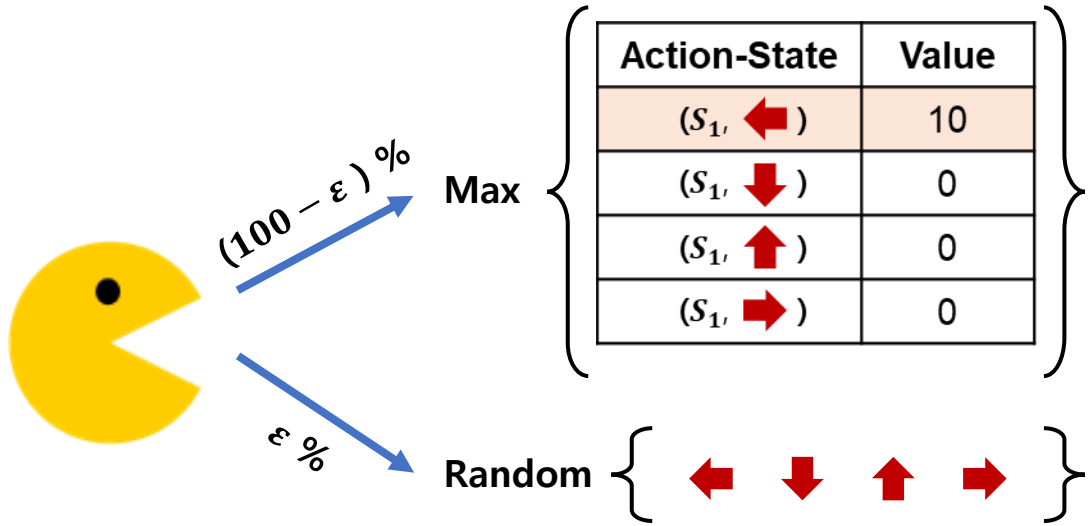
Basic exploration method

❖ 기본적인 exploration 방법론: ε-greedy

- Q-learning 방법론은 action-state value가 가장 큰 값의 행동을 선택함으로써 에이전트의 정책을 생성
- 이 때 일정 확률로 임의의 행동을 선택하도록 하는 탐험 방법론



Estimating action-state value with Q function



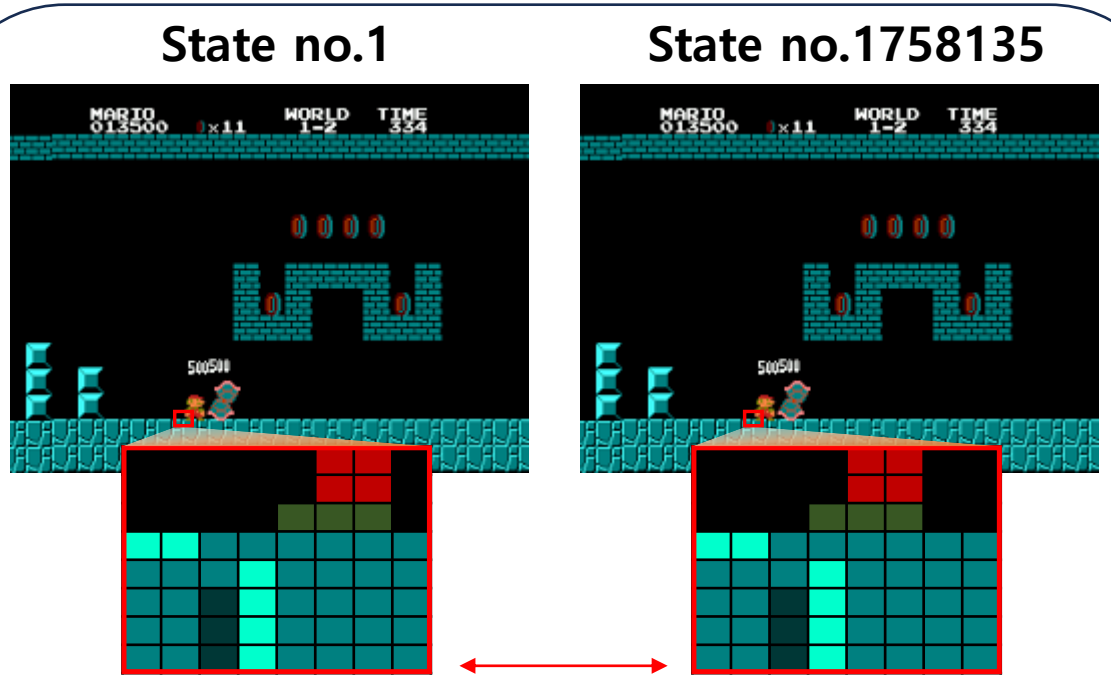
ε-greedy

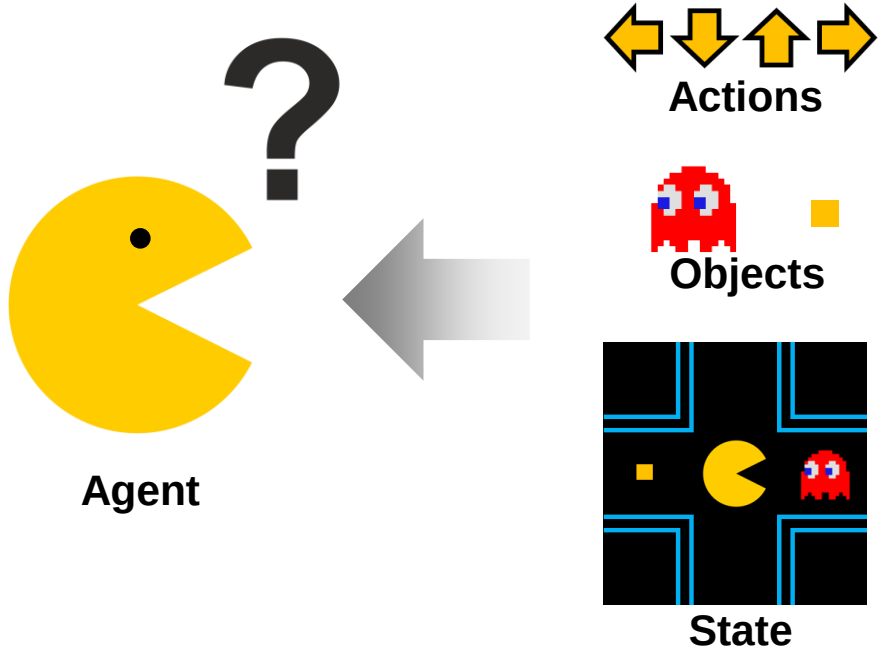
Exploration in RL

Needs of advanced exploration method

❖ 고도화된 exploration 방법론의 필요성

- 규모가 크고 복잡한 환경에서 모든 상태를 탐험하는 것은 불가능에 가까움
- 따라서 시간 효율적인 탐험을 위한 다양한 방법론이 연구되고 있음



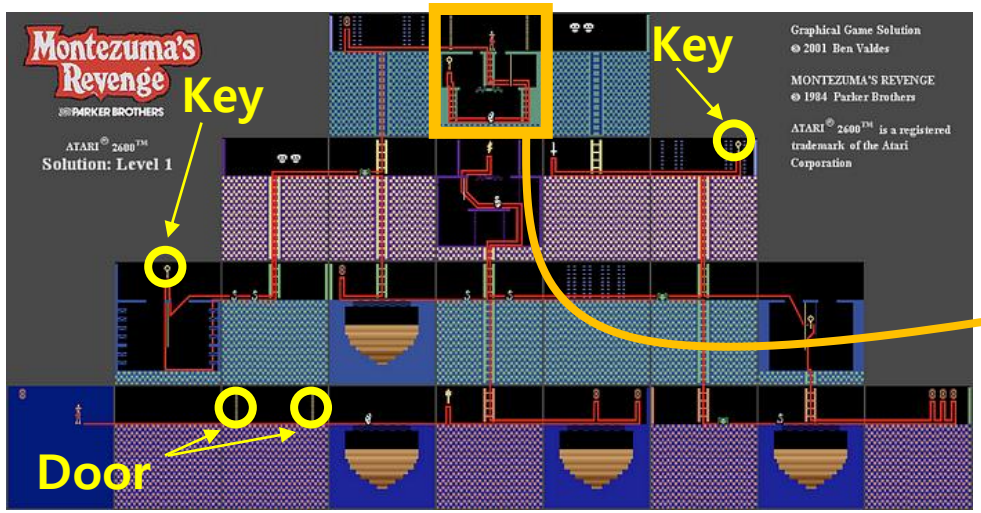


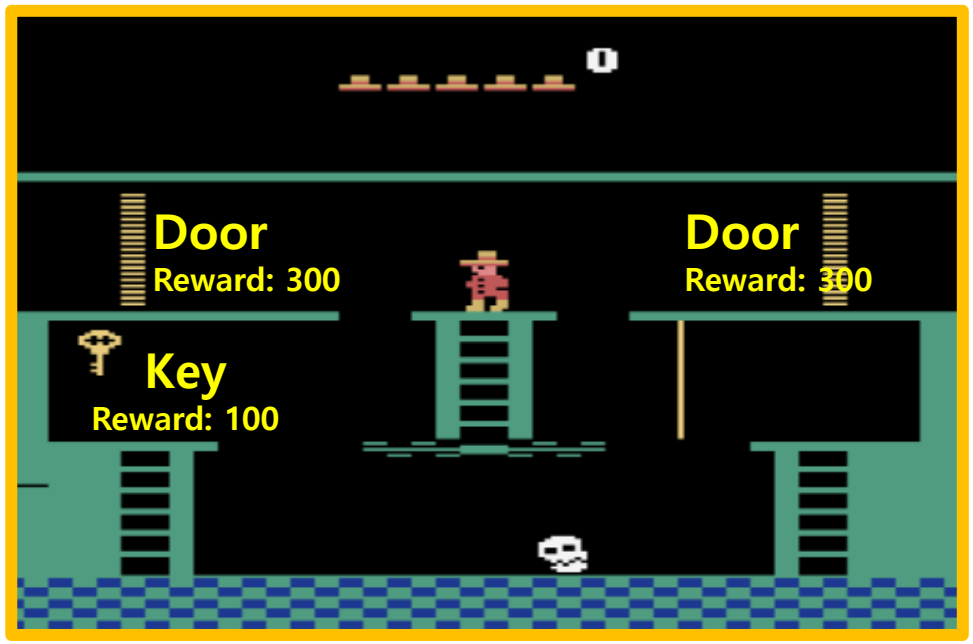
Exploration in RL

Key exploration problems

❖ Hard-exploration problem

- 보상이 매우 희소하여 에이전트가 적절한 피드백을 받기 어려운 문제 상황을 의미하며 대표적으로는 Montezuma's Revenge 게임이 있음
- 희소한 보상으로 인해 에이전트가 최적의 정책을 학습하기 위한 난이도가 높으며 시간이 매우 오래 걸림





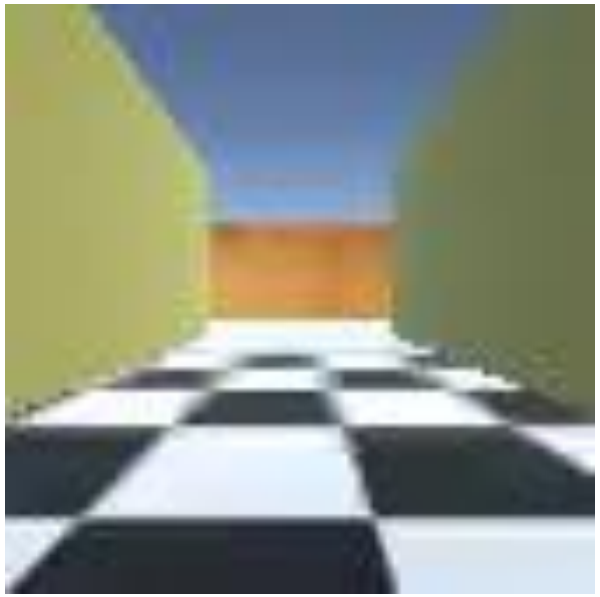
Montezuma's Revenge

Exploration in RL

Key exploration problems

❖ Noisy-TV problem

- 특정 상태에서 강한 random noise가 있을 때 에이전트가 해당 noise에만 집중하는 문제
- Curiosity-driven 방법론 일부가 해당 문제에 특히 취약한 모습을 보임



w/ Noisy-TV

w/o Noisy-TV

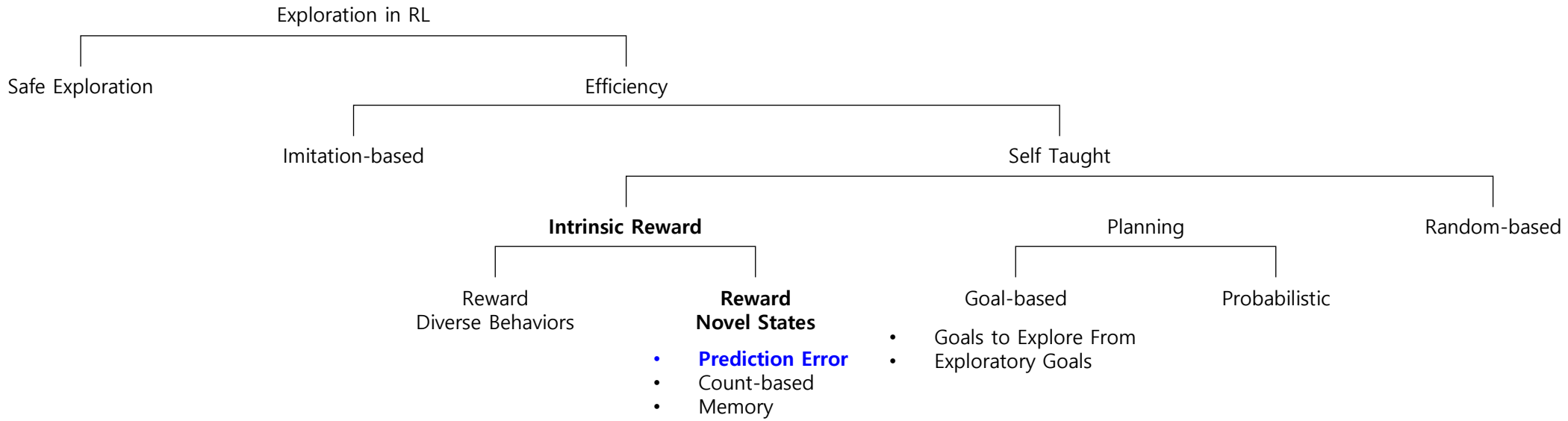
Burda, Yuri, et al. "Large-Scale Study of Curiosity-Driven Learning." *International Conference on Learning Representations*. 2018.

Exploration in RL

Hierarchy of exploration methods

❖ Exploration 방법론 개요

- 해당 개요는 Ladosz, Pawel, et al. "Exploration in deep reinforcement learning: A survey." Information Fusion 85 (2022):를 인용하였음
- Exploration 방법론은 종류가 매우 세분화 되어 있으며 그 중에서 prediction error 기반의 방법론 위주로 소개

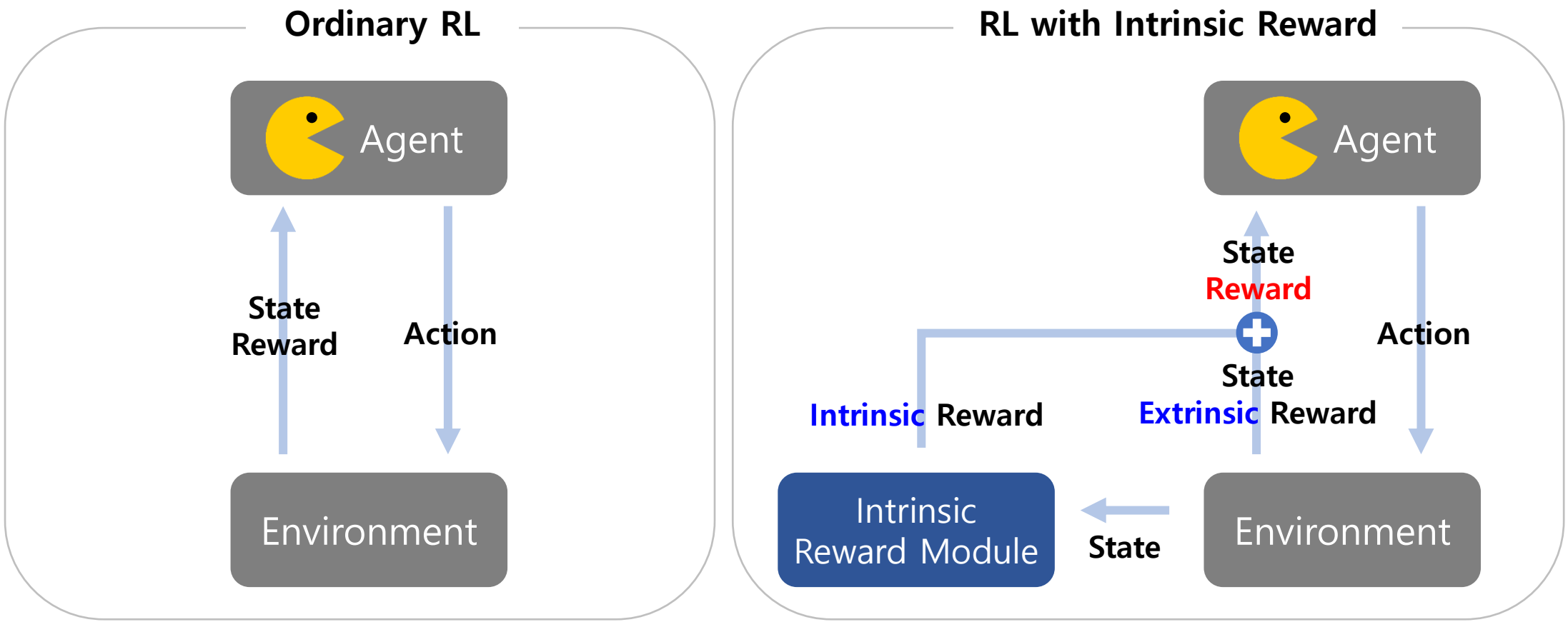


Exploration in RL

Exploration with intrinsic reward

❖ Intrinsic reward

- 환경에서 제공하는 보상(extrinsic reward)과 모델에서 제공되는 보상(intrinsic reward)을 더하여 정책을 업데이트

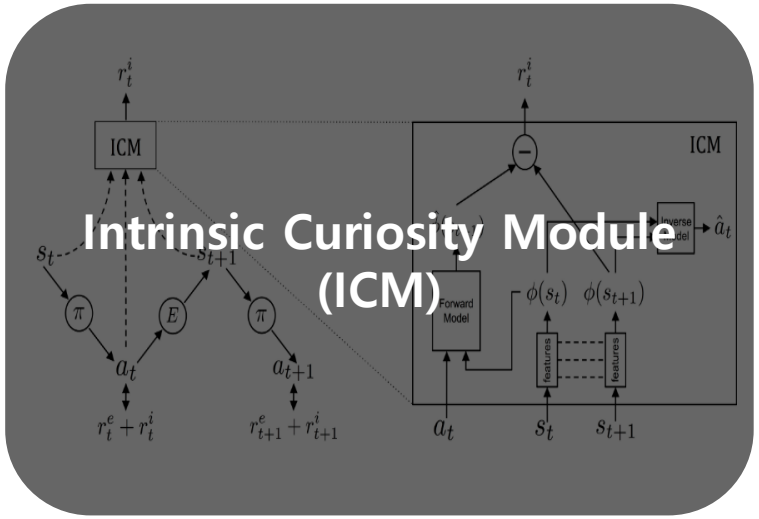


Exploration in RL

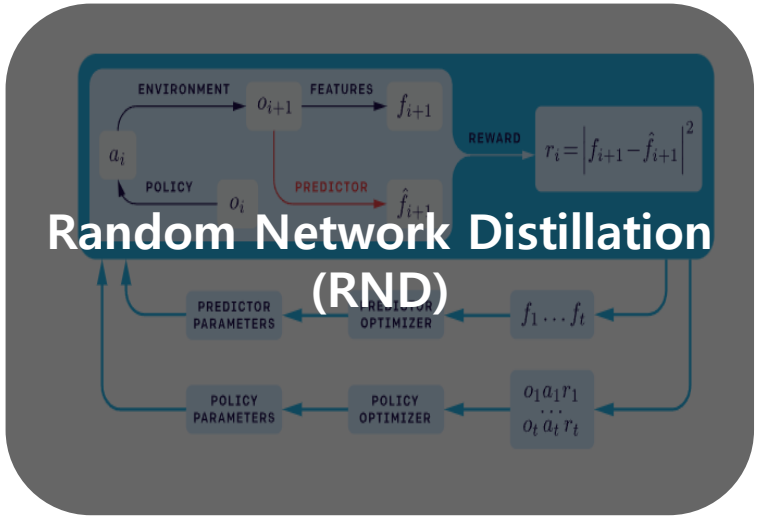
Exploration with intrinsic reward

❖ Curiosity as an intrinsic reward

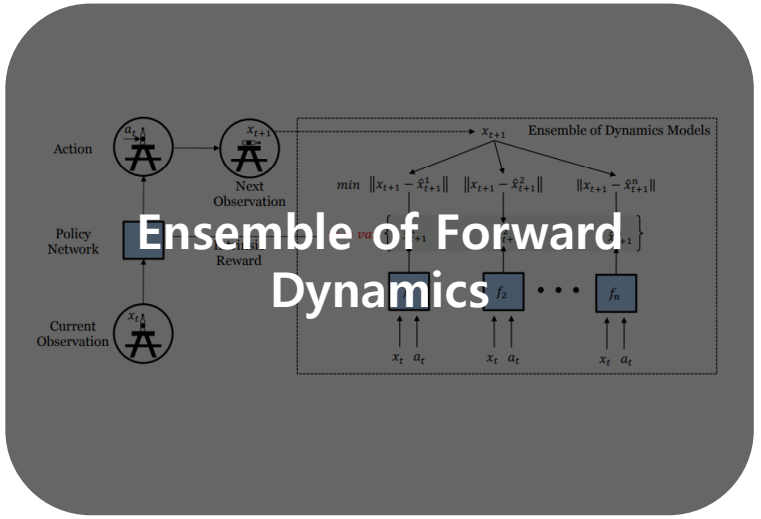
- Curiosity는 상태에 대해 불완전한 정보를 가지고 있을 수록 값이 커지는 지표로써 이 값이 큰 상태 위주로 탐험하도록 모델을 유인
- 예측 모델의 예측 오차 혹은 출력 값의 분산을 기반으로 curiosity를 계산하며 intrinsic reward로 활용



Intrinsic Curiosity Module (ICM)



Random Network Distillation (RND)



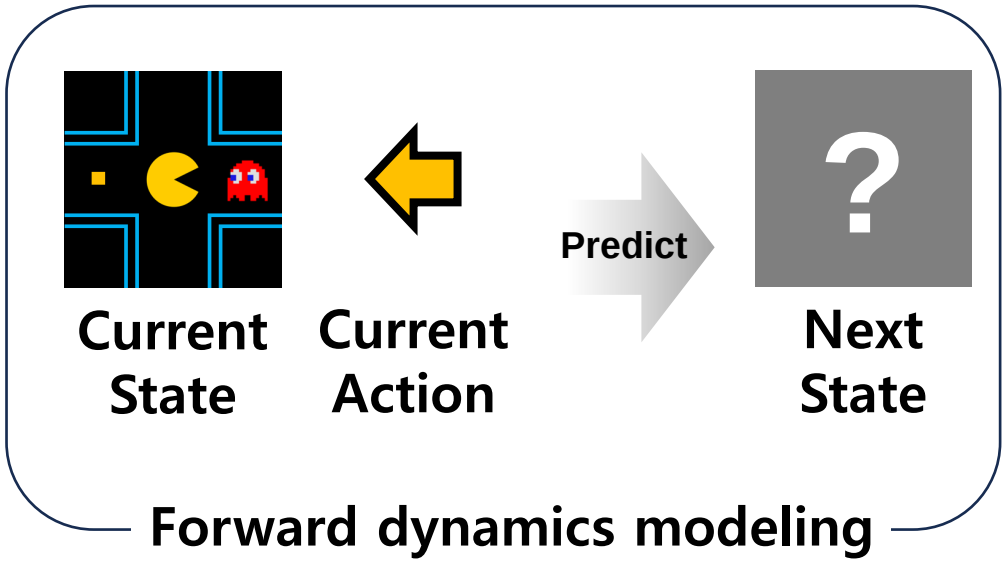
Ensemble of Forward Dynamics

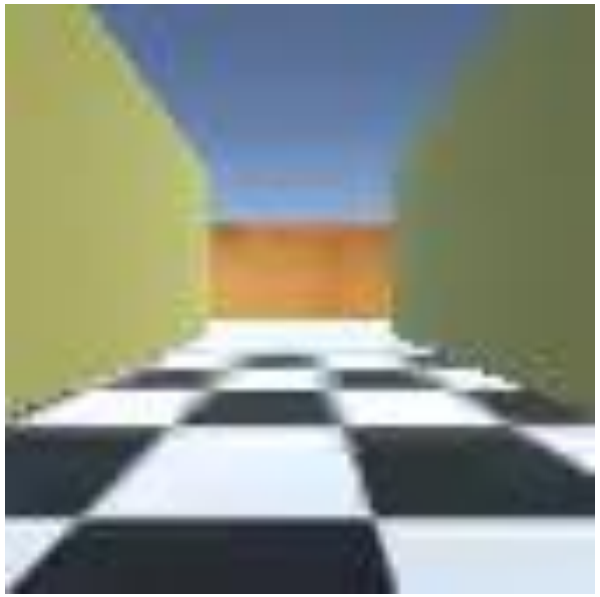
Exploration in RL

Exploration with intrinsic reward

❖ Problems of previous research

- 기존 방법론[1]은 forward dynamics modeling 기반으로 진행되었음
- Forward dynamics modeling은 현재 상태와 행동으로 다음 상태를 예측하며 발생한 오차를 intrinsic reward로 사용
- 예측 대상이 환경이므로 환경의 랜덤성이 강한 경우(ex. Noisy TV)에는 특정 상태에 고착되거나 무의미한 탐험을 수행하는 경우가 발생함





w/ Noisy-TV

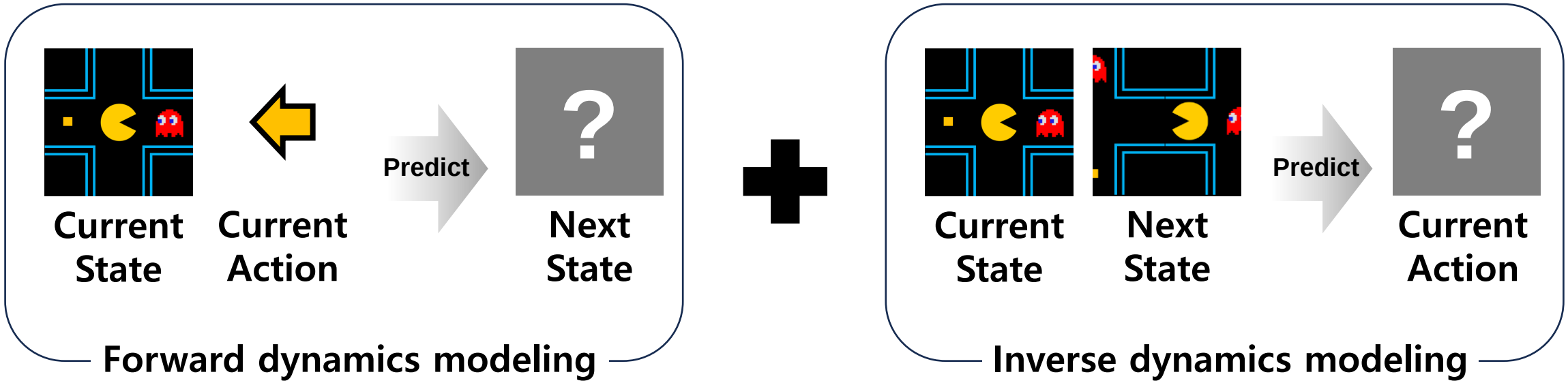
[1] Stadie, Bradly C., Sergey Levine, and Pieter Abbeel. "Incentivizing exploration in reinforcement learning with deep predictive models." *arXiv preprint arXiv:1507.00814* (2015).
[2] Burda, Yuri, et al. "Large-Scale Study of Curiosity-Driven Learning." *International Conference on Learning Representations*. 2018.

Exploration in RL

Exploration with intrinsic reward

❖ **Curiosity-driven Exploration by Self-supervised Prediction (ICML 2017 / 2,322회 인용)**

- Intrinsic reward module에 두 가지 environmental dynamics modeling을 사용
- Inverse dynamics modeling은 현재 상태와 다음 상태로 현재 행동을 예측하는 작업을 수행
- Inverse dynamics modeling을 사용하여 환경의 랜덤성이 강한 경우 발생하는 문제를 완화

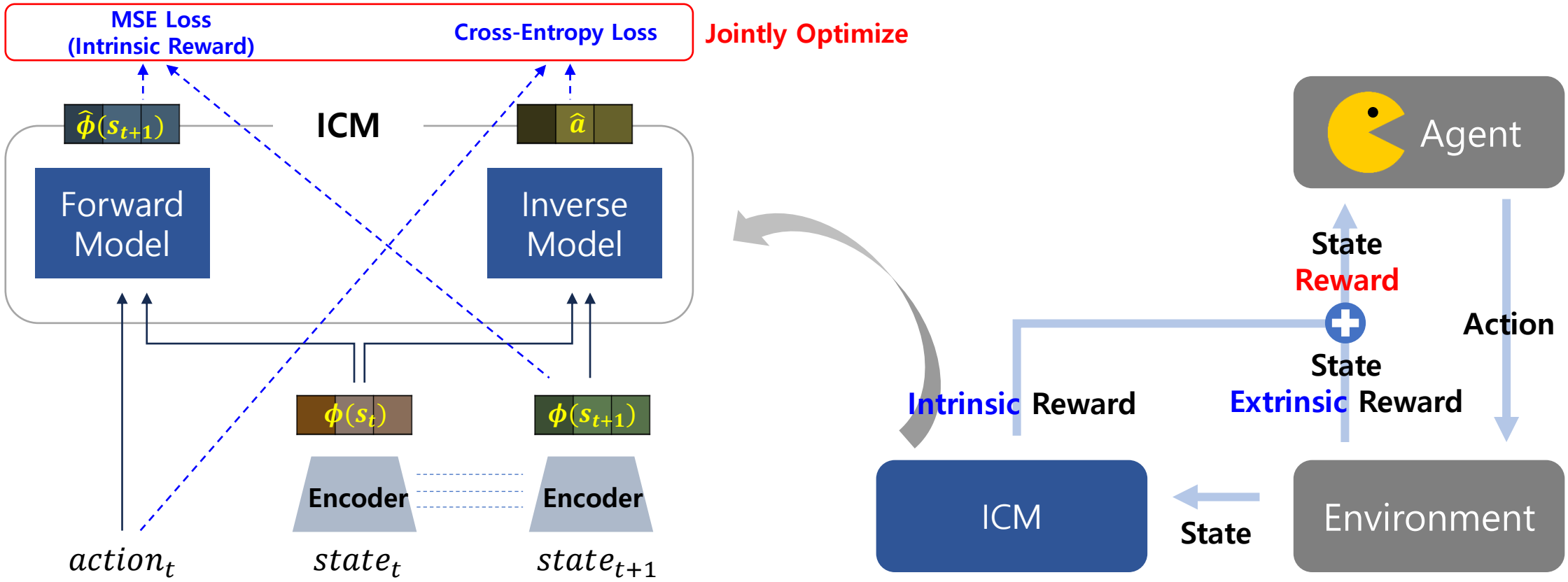


Exploration in RL

Curiosity-driven Exploration by Self-supervised Prediction

❖ Overview of Intrinsic Curiosity Module (ICM)

- 각 dynamics에서 나온 오차와 환경에서 주어지는 보상을 기반으로 에이전트를 업데이트



Exploration in RL

Curiosity-driven Exploration by Self-supervised Prediction

❖ Role of Inverse Dynamics in ICM

- Forward dynamics는 환경을 예측해야 하기 때문에 에이전트와 무관한 환경의 변화에도 민감하게 반응할 수 있음
- 반면 inverse dynamics는 에이전트의 움직임을 예측하는 문제이기 때문에 에이전트와 무관한 환경 요소는 무시하도록 학습
- 따라서 intrinsic reward를 계산할 때 정책 학습에 있어 불필요한 환경 요소들을 배제할 수 있음

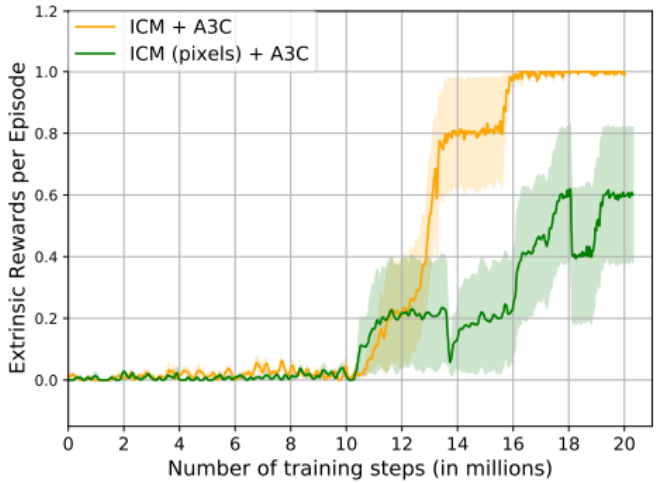


Figure 6. Evaluating the robustness of ICM to the presence of uncontrollable distractors in the environment. We created such a distractor by replacing 40% of the visual observation of the agent by white noise (see Figure 3b). The results show that while ICM succeeds most of the times, the pixel prediction model struggles.

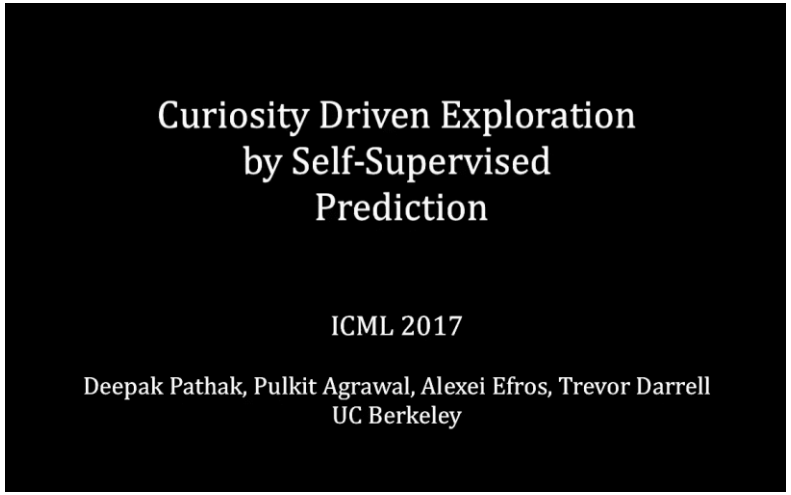
Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Exploration in RL

Exploration with intrinsic reward

❖ Large-Scale Study of Curiosity-Driven Learning (ICLR 2019 / 734회 인용)

- ICM을 보다 다양한 실험에 적용해본 연구 (환경, 아키텍처 등...)
- 순수하게 intrinsic reward만을 사용해서 에이전트를 학습하고 평가
- 세 가지 관점에서 어떤 방식으로 학습한 특징 공간이 탐험을 수행하는데 효과적인지 실험
  - 관점 1. Compactness: 주어진 상태에서 얼마나 쉽게 불필요한 요소를 제외하고 낮춘 차원의 벡터로 표현할 수 있는가
  - 관점 2. Sufficiency: 행동을 결정할 때 필요한 중요한 정보를 얼마나 잘 보존하고 있는가
  - 관점 3. Stability: Intrinsic reward가 실제로 에이전트의 curiosity를 얼마나 올바르게 표현해줄 수 있는가



Intrinsic reward만으로 학습된 에이전트 예시 (Pathak et al., 2018)

	VAE	IDF	RF	Pixels
Compactness	Yes	Yes	Maybe	Maybe
Sufficiency	Yes	Maybe	Maybe	Yes
Stability	No	No	Yes	Yes

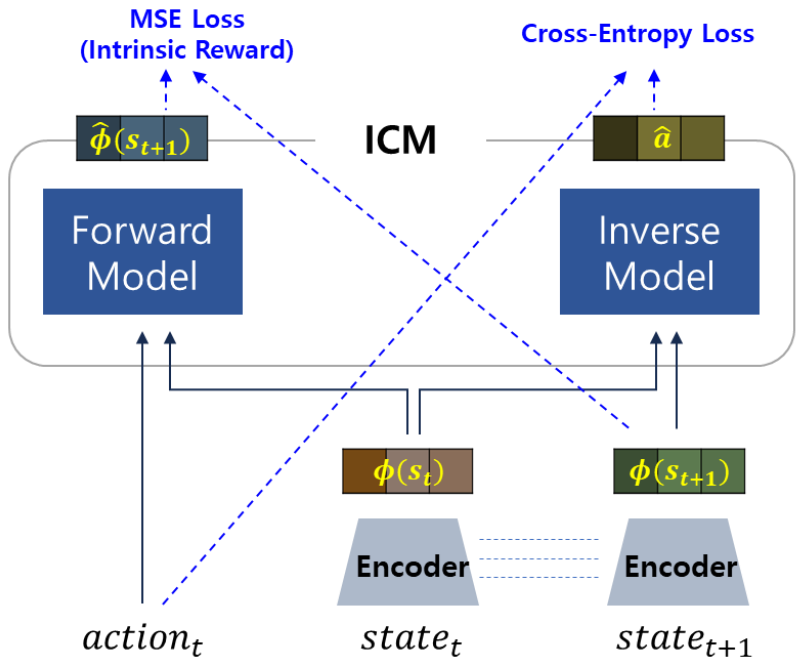
특징 공간을 학습하는 다양한 방식 및 관점 별 평가 (Burda et al., 2019)

Exploration in RL

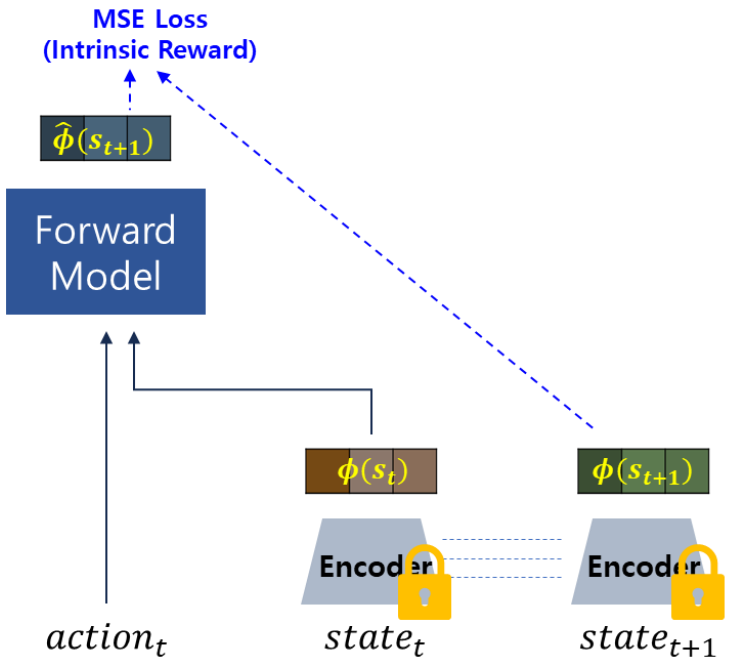
Exploration with intrinsic reward

❖ Large-Scale Study of Curiosity-Driven Learning (ICLR 2019 / 734회 인용)

- Inverse dynamics feature (IDF): ICM과 동일한 구조
- Random feature (RF): 인코더를 초기 파라미터 상태로 고정한 뒤 나오는 특징 벡터를 기반으로 탐험을 수행



Inverse dynamics feature



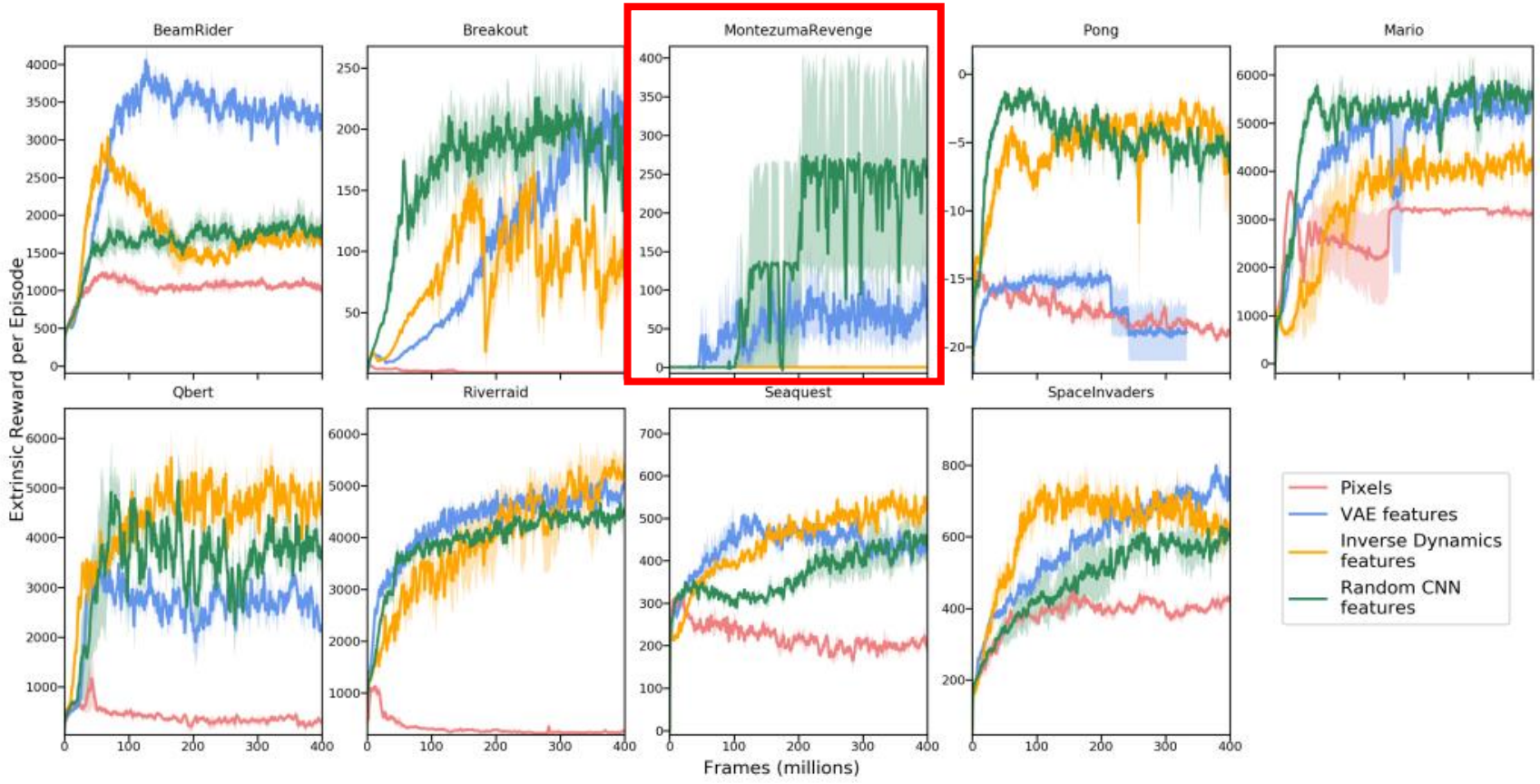
Random feature

Exploration in RL

Large-Scale Study of Curiosity-Driven Learning

❖ Potential of learning random feature space

- 실험 전반에서 뿐만 아니라 대표적인 Hard exploration 게임인 Montezuma's Revenge에서도 뛰어난 성능을 보임

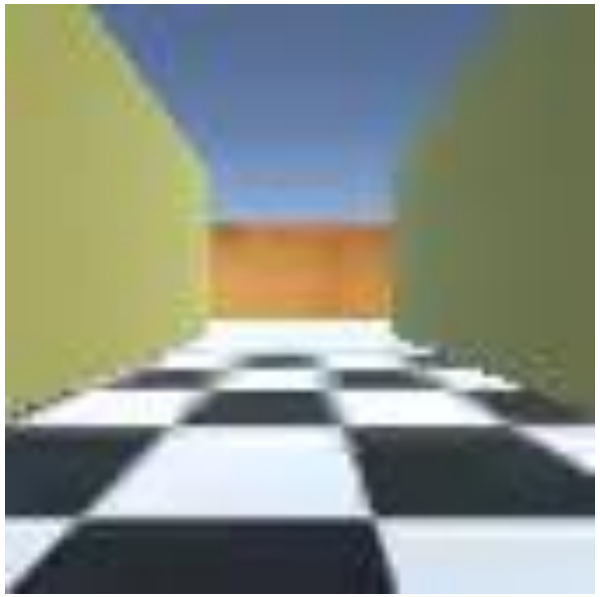


Exploration in RL

Large-Scale Study of Curiosity-Driven Learning

❖ Noisy-TV problem in Unity Maze

- 앞선 논문(Pathak et al., 2018)에서 제시한 noisy-TV 문제는 해결이 되었으나 Burda et al., 2019에서 제시한 형태에서는 여전히 문제가 발생
- 기존 문제의 경우 처음부터 유사한 패턴의 white noise가 존재하는 반면, 해당 연구에서는 탐험 중에 발견하며 다채로운 패턴을 가짐
- 전이 확률이 랜덤성을 가지는 환경에서는 prediction error 기반의 모델이 더 강한 랜덤성을 가진 상태를 탐색하게 됨



w/ Noisy-TV

w/o Noisy-TV

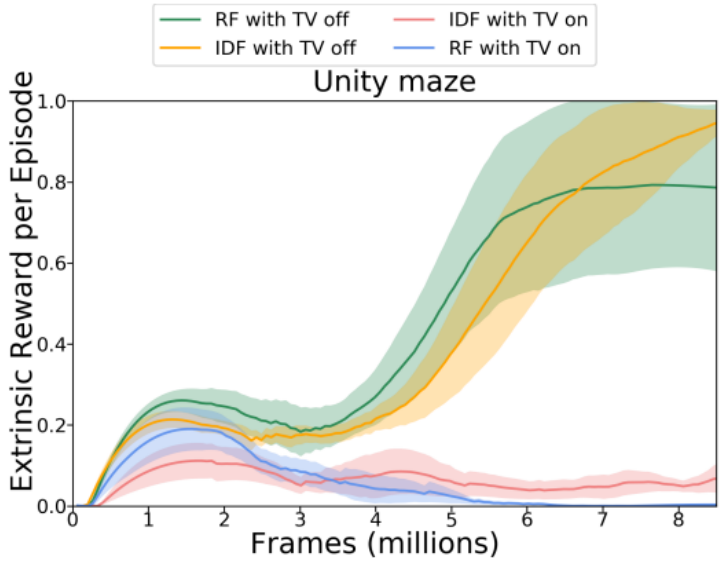


Figure 6: We add a noisy TV to the unity environment in Section 3.3. We compare IDF and RF with and without the TV.

Burda, Yuri, et al. "Large-Scale Study of Curiosity-Driven Learning." *International Conference on Learning Representations*. 2018.

Exploration in RL

Exploration with intrinsic reward

❖ Self-Supervised Exploration via Disagreement (ICML 2019 / 307회 인용)

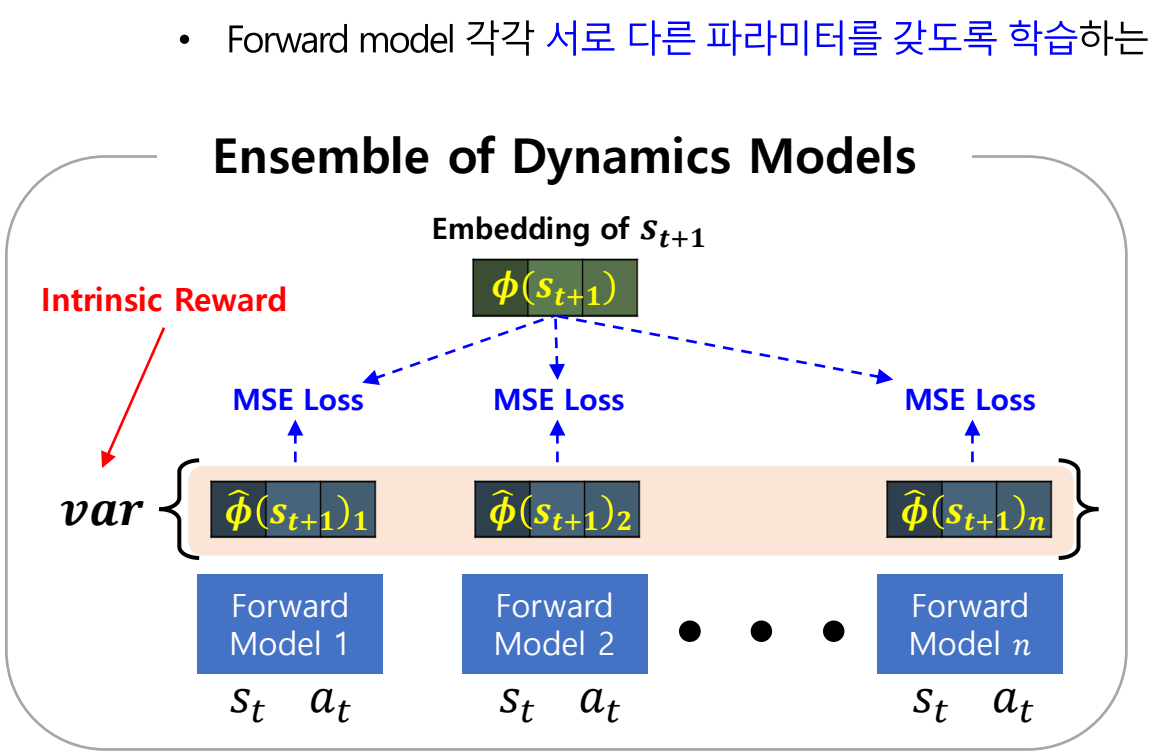
- Stochastic dynamics를 가진 환경에서도 잘 작동할 수 있는 탐험 방법론이 연구 됨
- 해당 연구에선 forward 모델의 앙상블로부터 나온 intrinsic reward만으로 정책을 학습 (extrinsic reward 사용 안 함)
- 기존 연구와 달리 예측 오차가 아닌 예측 값들의 분산을 기반으로 intrinsic reward를 계산 → 새로운 noisy-TV 문제를 해결

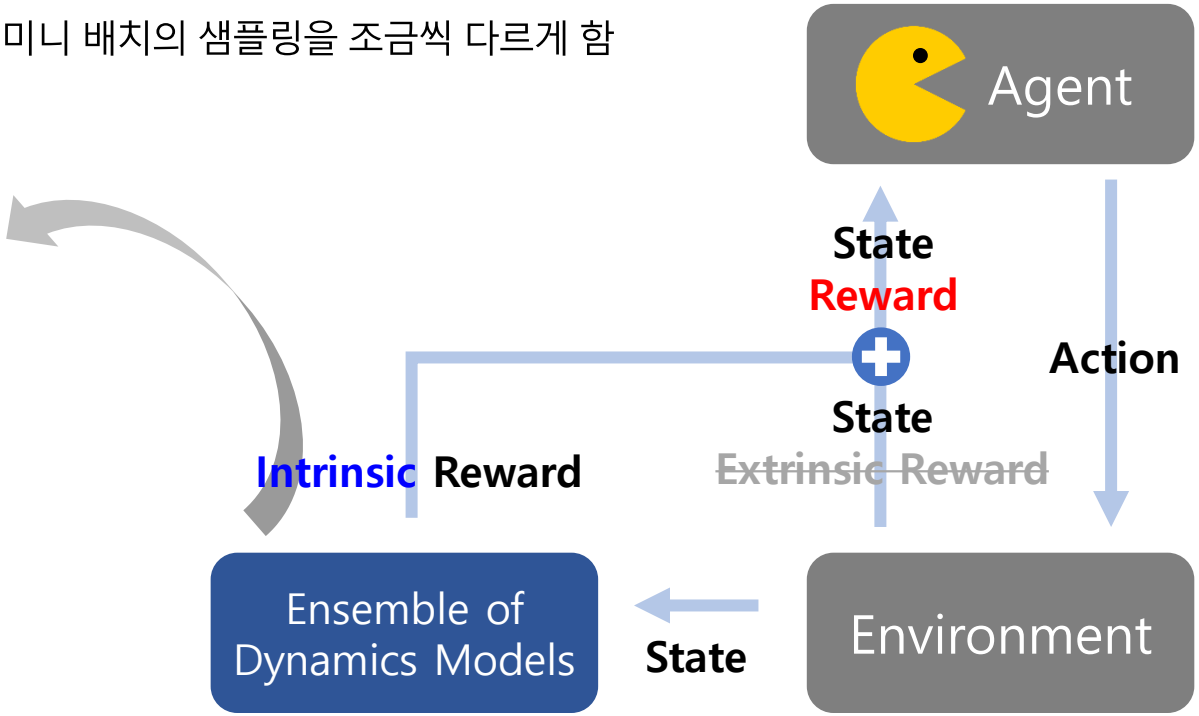
Exploration in RL

Self-Supervised Exploration via Disagreement

❖ Overview of Ensemble of Dynamics Models

- 다수의 forward 모델에서 나온 출력의 분산을 disagreement라고 정의하여 intrinsic reward로 사용
- Intrinsic reward를 최대화하는 방향으로 정책을 업데이트 하여 더 많은 탐험이 이루어지도록 함
- Forward models는 MSE loss를 통해서 오차를 줄이도록 학습
- Forward model 각각 서로 다른 파라미터를 갖도록 학습하는 미니 배치의 샘플링을 조금씩 다르게 함



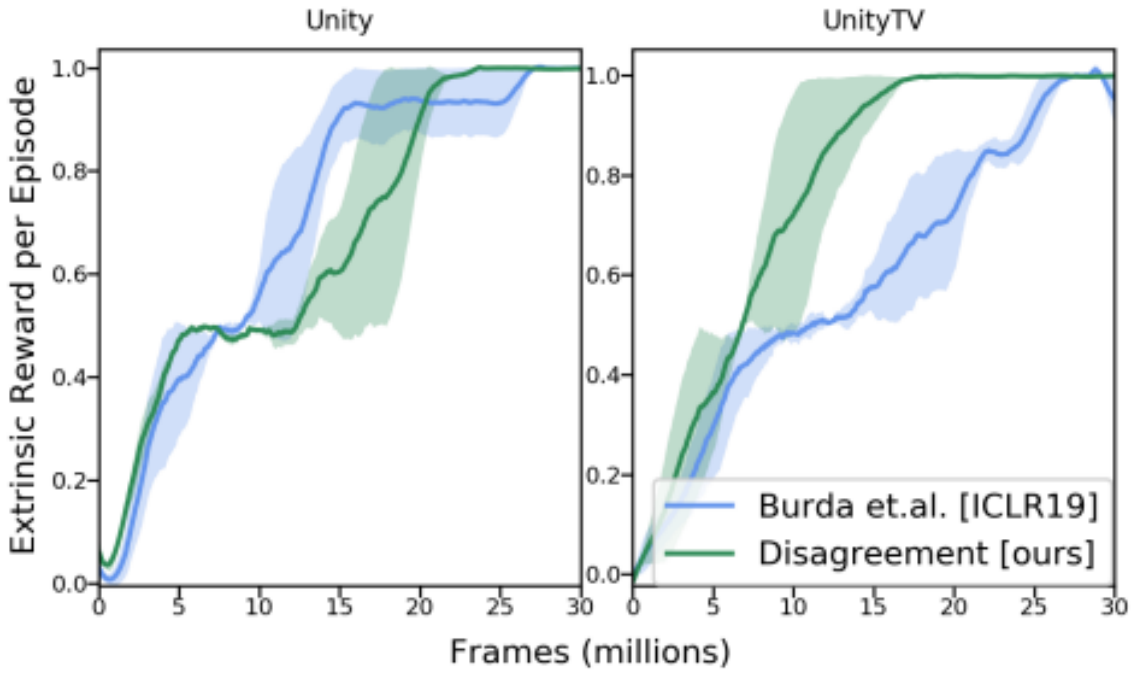


Exploration in RL

Self-Supervised Exploration via Disagreement

❖ Robust to stochastic dynamics

- [1]에서 제안한 random feature space를 학습한 forward dynamics 모델과 성능 비교
- Ensemble of forward models는 출력 값의 분산을 사용하기 때문에 실제 next state 값과는 상관 없이 자주 본 상태에 대해서는 유사한 값을 출력
- 따라서 이러한 특성으로 Noisy-TV 문제에서 에이전트가 고착되는 문제를 해결하였으며 기존보다 빠른 수렴 속도를 보임



[1] Burda, Yuri, et al. "Large-Scale Study of Curiosity-Driven Learning." *International Conference on Learning Representations*. 2018.

Exploration in RL

Exploration with intrinsic reward

❖ Exploration by Random Network Distillation (ICLR 2019 / 1,119회 인용)

- Curiosity(prediction error)가 높은 원인을 세 가지 제시함

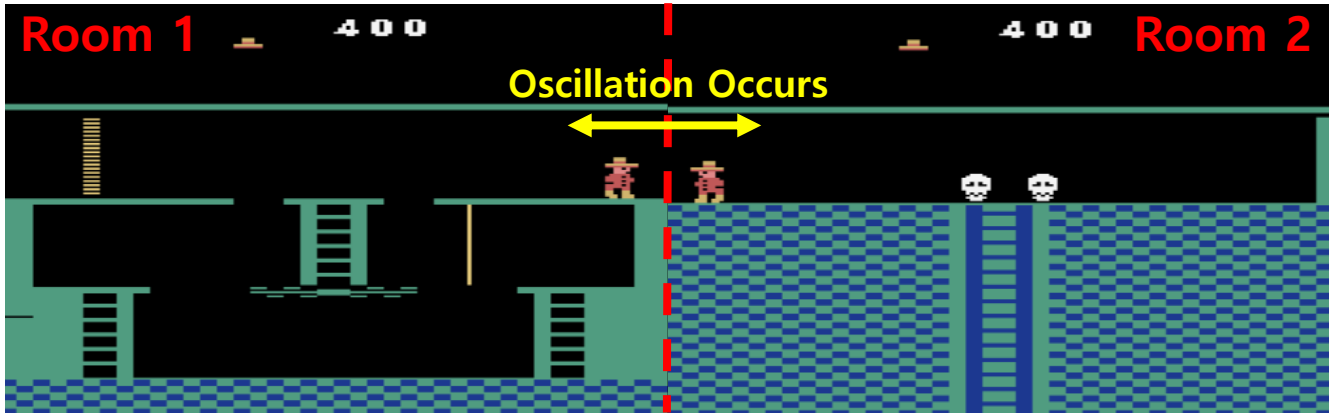
Need → 1. Amount of training data: 아직 충분히 학습되지 않은 데이터(상태)에 해당하기 때문 → **Curiosity**

Avoid → 2. Stochasticity: 예측할 대상(next state)의 변동성이 높기 때문

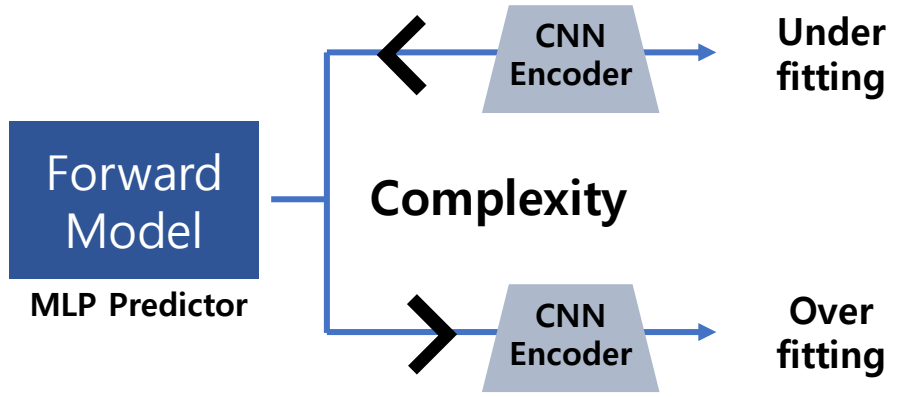
Avoid → 3. Model misspecification/Learning dynamics: 예측 모델에 쓰일 모델 크기가 커도 작아도 문제가 발생함

} **Noisy-TV 문제를 발생시킴**

- 원인 2와 3은 탐험 효과를 저해하기 때문에 이를 피하기 위해서 새로운 아키텍처를 제안함



Stochasticity 예시



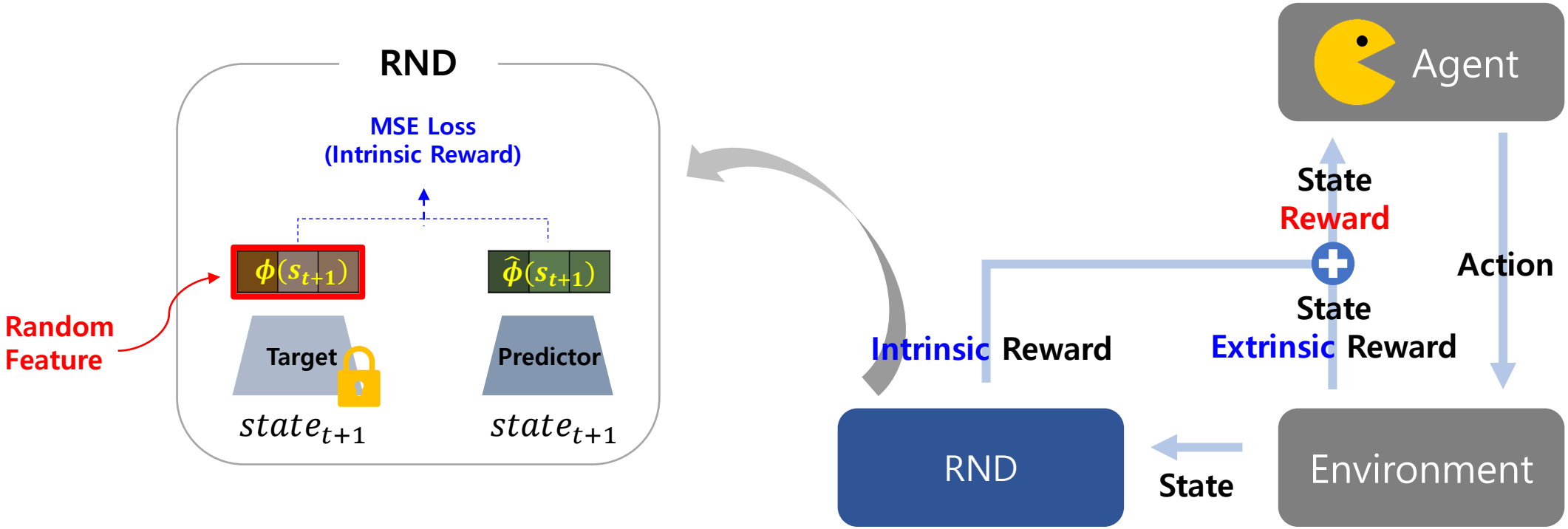
Model misspecification

Exploration in RL

Exploration by Random Network Distillation

❖ Overview of Random Network Distillation (RND)

- Target network와 predictor는 서로 동일한 아키텍처로 구성 (원인 3 해결)
- Target network는 초기 파라미터 그대로 고정하여 random feature를 출력하도록 함 (deterministic target function; 원인 2 해결)
- 유사한 상태를 경험할 수록 random feature를 더 잘 예측할 수 있게 되므로 curiosity로 활용 가능

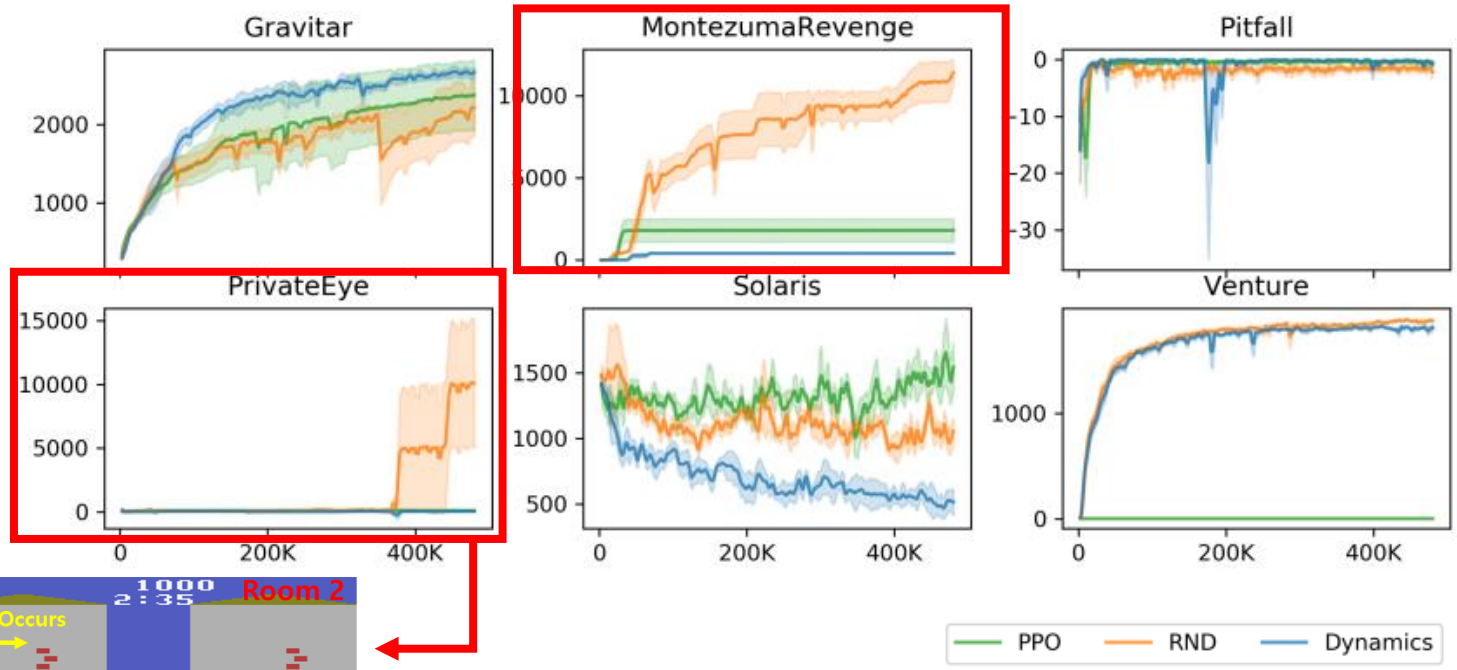


Exploration in RL

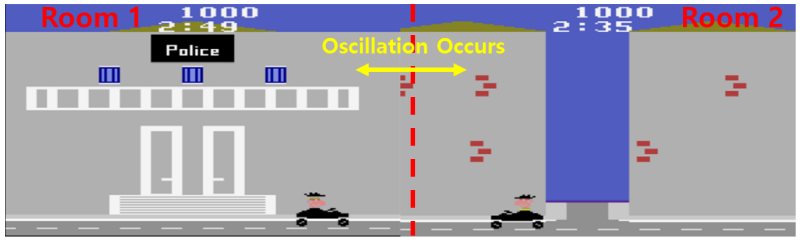
Exploration by Random Network Distillation

❖ Exploration by Random Network Distillation (ICLR 2019 / 1,119회 인용)

- Atari2600에서 대표적인 hard exploration 게임 6 종으로 탐험 성능을 평가함
- 특히 Stochasticity가 심하고 게임 내에서 얻은 아이템을 다른 상태에서 사용하는 것이 필요한 게임에서 매우 우월한 성능을 보임



PrivateEye



Summary

❖ **환경의 정보를 불완전하게 알고 있는 상황에서 최적의 정책을 찾기 위해선 탐험이 반드시 필요**

❖ **기존의 예측 기반의 탐험 방법론(forward dynamics modeling)은 매우 확률적인 환경(ex. Noisy TV)에서 탐험이 어려움**

❖ **이를 극복하기 위하여 다음 상태를 예측하는 오차 대신 다른 방법론이 연구 되었음**

- Ensemble of Forward Dynamics: 서로 다른 다수의 forward dynamics model 출력 간의 분산을 사용
- Random Network Distillation: 특정 시점의 상태에서 random feature를 추출한 뒤 동일한 시점의 상태를 기반으로 예측

Citation

- Deep Reinforcement Learning

Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Ladosz, Pawel, et al. "Exploration in deep reinforcement learning: A survey." *Information Fusion* 85 (2022): 1-22.

Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *International conference on machine learning*. PMLR, 2017.

Burda, Yuri, et al. "Large-Scale Study of Curiosity-Driven Learning." *International Conference on Learning Representations*. 2018.

Pathak, Deepak, Dhiraj Gandhi, and Abhinav Gupta. "Self-supervised exploration via disagreement." *International conference on machine learning*. PMLR, 2019.

Burda, Yuri, et al. "Exploration by random network distillation." *International Conference on Learning Representations*. 2018.