SUSC Summer 2023

강화학습 논문 리뷰

동아대 AI학과 김지선

01

논문의 수식 해석 예시

실제 논문 수식으로 살펴보는 강화학습 기본

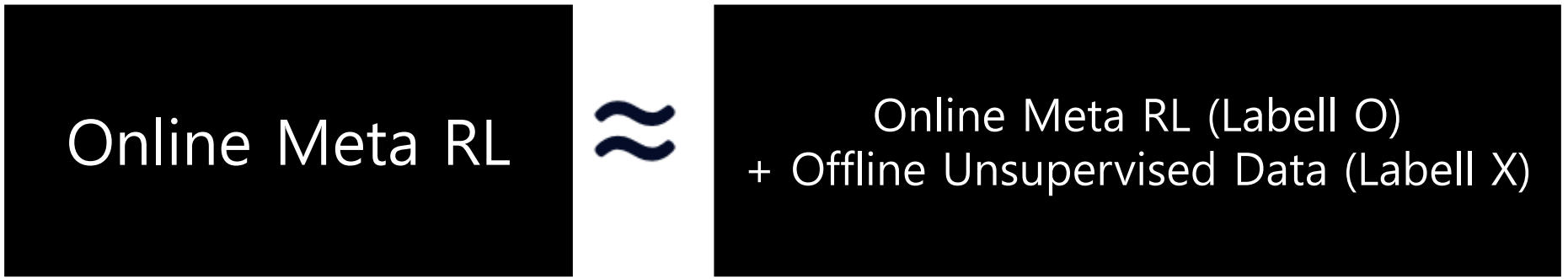
강화학습 논문 소개

- 지금까지 배운 내용
  - 용어: Agent, Action, State, Environment
  - 개념: Actor-Critic, REINFORCE 알고리즘(몬테카를로), 벨만 방정식, A2C

- 왜 강화학습 논문을 ?
  - 기초 이론을 배우고 논문(활용)을 보니까 웬만한 내용이 다 이해되고,
  - 기존에 공부했던 내용도 더 실질적으로 와 닿았음.

오늘 소개할 강화학습 논문

- 학회지: NeurIPS 2021, 12월에 발표된 논문
- 제목: Offline Meta-Reinforcement Learning with Online Self Supervision
- 요약
  - 기본 Online Meta-RL은 성능은 좋지만 고비용의 문제가 있다.
  - Offline meta-RL에 추가로 labelling되지 않은, Online Unsupervised data를 추가해서 학습하면 그냥 Online Meta-RL만큼 성능이 좋아짐.



용어 다시 Remind

- Actor-Critic: 행동하고, 한 행동에 대해 평가하는 강화학습의 대표적인 학습법
- On-Offline 학습
  - Online 학습
  - Offline 학습

- Labelling
  - ML에서 데이터의 특성과, 정답 개념 Remind
  - Feature: 데이터의 특성을 나타내는 속성 (Target 칼럼을 제외한 모든 칼럼)
  - Labell: Target, Class로도 불리며, 데이터의 정답을 나타냄.

- Meta Reinforcement Learning
  - Meta-Learning: Meta 정보를 이용한 학습
  - Meta-Reinforcement Learning: Meta-Learning을 사용한 학습

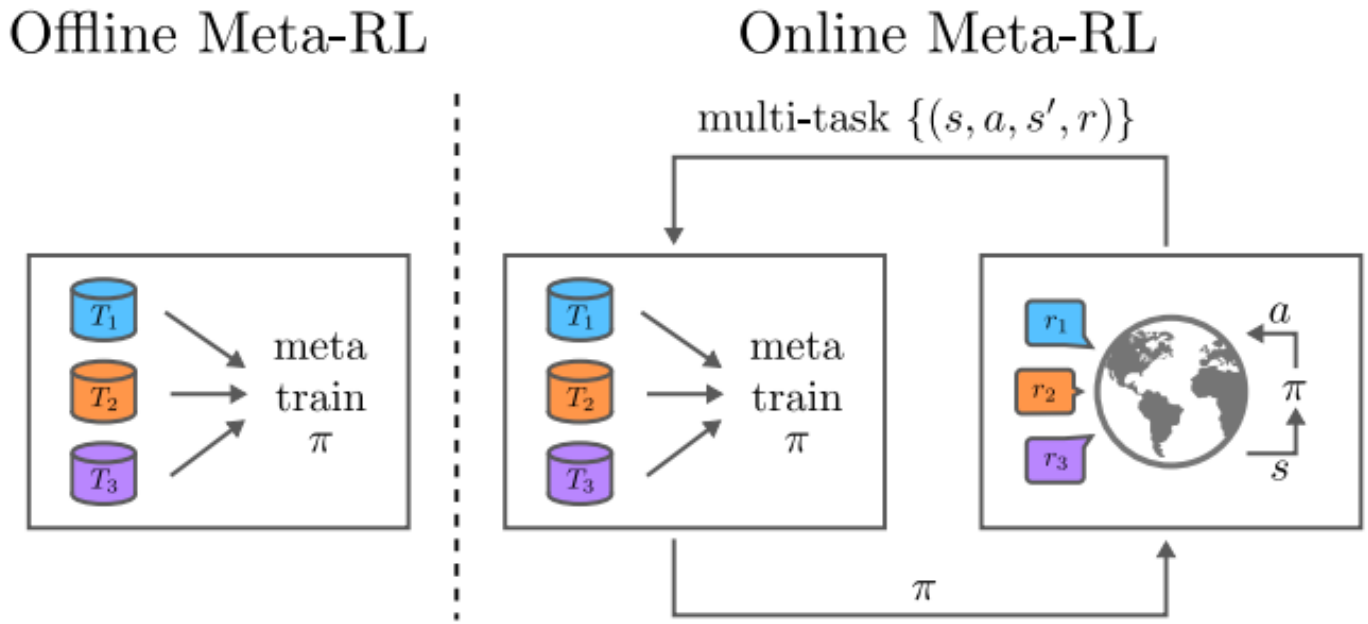
추가로 등장하는 용어

- KL divergence: 두 확률분포 사이의 차이를 계산하는데 사용하는 함수 $D_{KL}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}$
- Semi-Supervsied Learning (준지도학습)
  - 지도학습에 준하는 학습, 레이블이 달려 있는 데이터와 레이블이 달려 있지 않은 데이터를 모델 학습에 동시에 사용하는 방법, 레이블이 없는 데이터로부터 추가 정보를 얻어 일반화 성능 향상 가능
- Self-Supervsied Learning (자기 지도학습) (in BERT, Google) (음성신호에 주로 사용)
  - 레이블이 지정되지 않은 샘플 데이터에서 학습하며, 자기 스스로 학습 데이터에 대해 레이블을 생성하거나, 예측하는 방법.
- Supervision: 모델 학습에 사용되는 지도 신호 레이블로, 지도 신호를 통해 원하는 출력을 예측 가능.
- Benchmark: 측정의 기준이 되는 대상을 설정하고, 비교 분석을 통해 따라 배움.

참조: Self-supervised learning (자기지도 학습) 이란?

논문 요약 (Abstract)

- 기본 Online Meta-RL은 Online Interaction(최근 데이터)을 기반으로 학습해서 성능이 좋지만 고비용의 문제점이 있음,
- Labelling된 Offline meta-RL에 추가로 Labelling 되지 않은 Online Unsupervised data를 학습시키면, 그냥 Online Meta-RL만큼 성능이 좋아짐



논문에서 제기한 문제점

- 정책(Policy)이 고정된 Offline dataset에서 학습되어서, Online dataset (탐색적 데이터)에 적응할 때 행동 예측이 불가능해서

  => 예측하는 분포의 차이가 발생한다.

  (잘 예측하지 못한다, 일반화 성능이 떨어진다.)

- 고정된 Offline labeled data로 학습하고, 추가적인 Online unlabeled data를 학습시키면, 정책 (Policy)의 적응 능력이 매우 향상되고, 분포 변화 문제를 해결 가능

  => Semi-Supervised meta Actor-Critic (SMAC)

논문에서 제기한 문제점

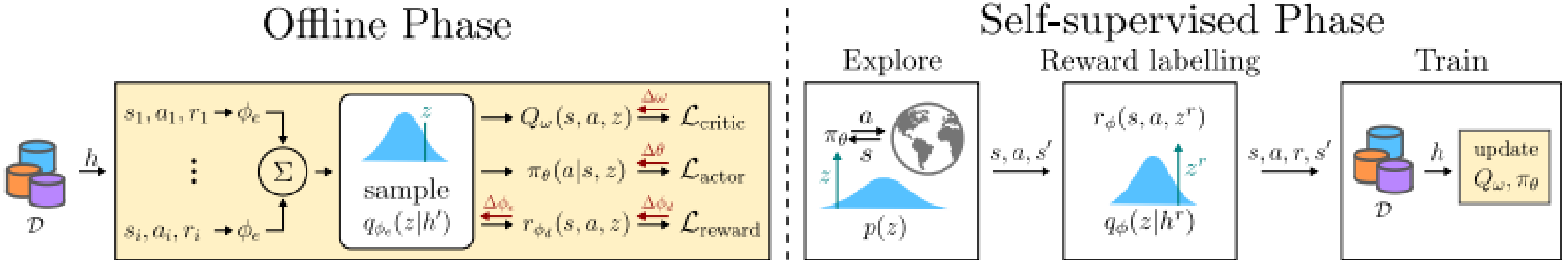
- Offline dataset에서 학습한 행동 정책 $\pi_\beta(\mathbf{a}, \mid \mathbf{s}, \mathbf{z})$
- Online dataset에서 학습한 탐색적 행동 정책 $\pi_\theta(\mathbf{a}, \mid \mathbf{s}, \mathbf{z})$

- 이 둘 사이의 차이가 예측 분포 변화 문제를 유발한다.

$$p(\mathbf{z} \mid \mathbf{h}_{\text{offline}}) \text{ and } p(\mathbf{z} \mid \mathbf{h}_{\text{online}})$$

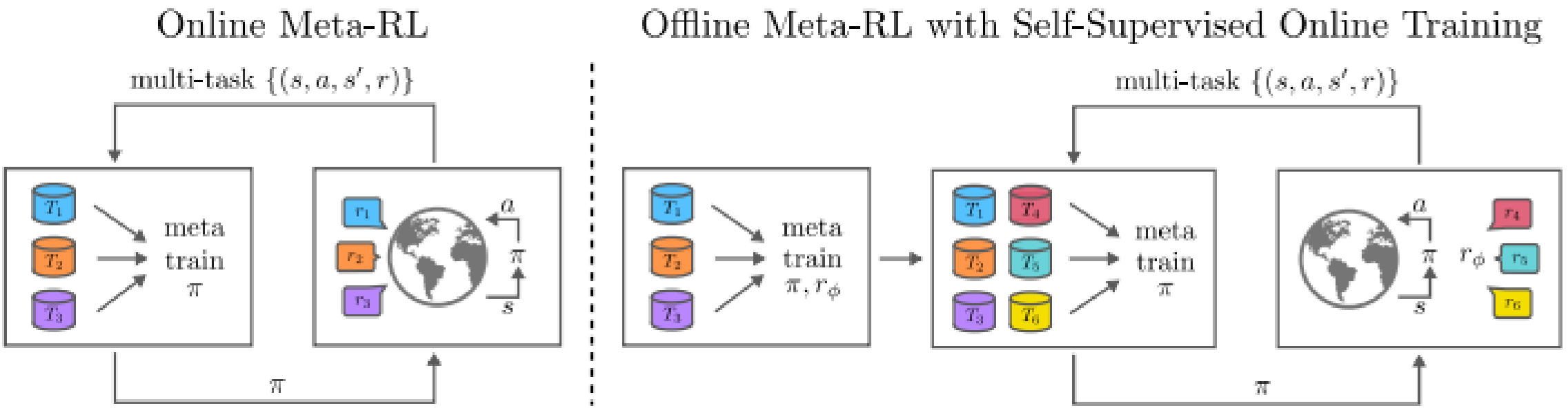
논문에서 제시한 방법론 (SMAC)

- SMAC – Semi-Supervised Actor-Critic
- 앞서 정책의 차이가 예측 분포의 차이를 이끈다는 것을 설명했는데, 이 문제를
- 아래 구조와 같이 Offline 데이터에서는 Encoder와 Decoder만 사용하고, Offline에서 Policy을 학습시켜서 문제를 해결함 !!
  - Encoder (입력), Decoder (출력)



논문에서 제시한 방법론 (SMAC)

- SMAC – Semi-Supervised Actor-Critic



벨만 에러 방정식 활용 예시

- 이 논문에서도 벨만 에러 방정식이라고 아래와 같은 보상의 차이를 구함.
- Q(s,a,z) 해석
  - 현재 상태 s에서 a라는 행동을 하고, 다음 상태 z를 고려할 때 받을 수 있는 기대 보상값
  - Q는 행동 가치 함수, a는 행동(action), s는 상태(state), z는 다음상태
- Q(s′,a′,z′) 해석
  - 다음 상태 s′에서 다음 행동 a′를 하고, 다음 상태 z를 고려할 때 받을 수 있는 기대 보상값
  - Q는 행동 가치 함수, a는 행동(action), s는 상태(state), z는 다음상태,r은 보상(reward)

$$\mathcal{L}_{\text{critic}}(w) = \mathbb{E}_{(\mathbf{s},\mathbf{a},r,\mathbf{s}')\sim\mathcal{D}_i, z\sim q_{\phi_e}(\mathbf{z}|\mathbf{h}), \mathbf{a}'\sim\pi_\theta(\mathbf{a}'|\mathbf{s}',\mathbf{z})}\Big[\\ (Q_w(\mathbf{s},\mathbf{a},\mathbf{z}) - (r + \gamma Q_{\bar{w}}(\mathbf{s}',\mathbf{a}',\mathbf{z})))^2\Big],$$

벨만 에러 방정식 활용 예시

- $(Q_w(\mathbf{s}, \mathbf{a}, \mathbf{z}) - (r + \gamma Q_{\bar{w}}(\mathbf{s}', \mathbf{a}', \mathbf{z})))^2$ 해석
- 활용 예시- 논문에서도 벨만 에러 방정식이라고 아래와 같은 보상의 차이를 구함.
- 현재 행동 가치에서 다음 행동 가치의 차이를 빼서 현재 시점에서 다음 행동을 했을 때 받는 보상의 차이를 계산해서 그 값의 기댓값 (E)을 평가하는(critic) Loss 함수로 함. (Lcritic)

$$\mathcal{L}_{\text{critic}}(w) = \mathbb{E}_{(\mathbf{s},\mathbf{a},r,\mathbf{s}')\sim\mathcal{D}_i, z\sim q_{\phi_e}(\mathbf{z}|\mathbf{h}), \mathbf{a}'\sim\pi_\theta(\mathbf{a}'|\mathbf{s}',\mathbf{z})}\Big[(Q_w(\mathbf{s}, \mathbf{a}, \mathbf{z}) - (r + \gamma Q_{\bar{w}}(\mathbf{s}', \mathbf{a}', \mathbf{z})))^2\Big],$$

벨만 기대 방정식

- 벨만 기대 방정식은 현재 State와 다음 State의 가치함수 사이의 관계를 식으로 나타낸 것.

$$V^{\pi}(x_t) = \int_{u_t} Q^{\pi}(x_t, u_t)\pi(u_t|x_t)du_t$$

$$= \mathbb{E}_{u_t \sim \pi(u_t|x_t)}[Q^{\pi}(x_t, u_t)]$$

<이론>

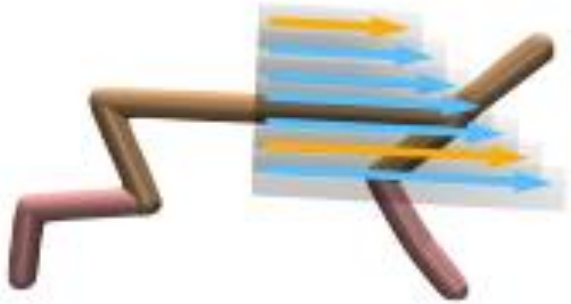
$$\mathcal{L}_{\text{actor}}(\theta) = -\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}' \sim \mathcal{D}, \mathbf{z} \sim q_{\phi_e}(\mathbf{z}|\mathbf{h})}\Big[\log \pi_\theta(\mathbf{a} \mid \mathbf{s}) \times \exp\left(\frac{Q(\mathbf{s},\mathbf{a},\mathbf{z}) - V(\mathbf{s}',\mathbf{z})}{\lambda}\right)\Big]. \quad (2)$$

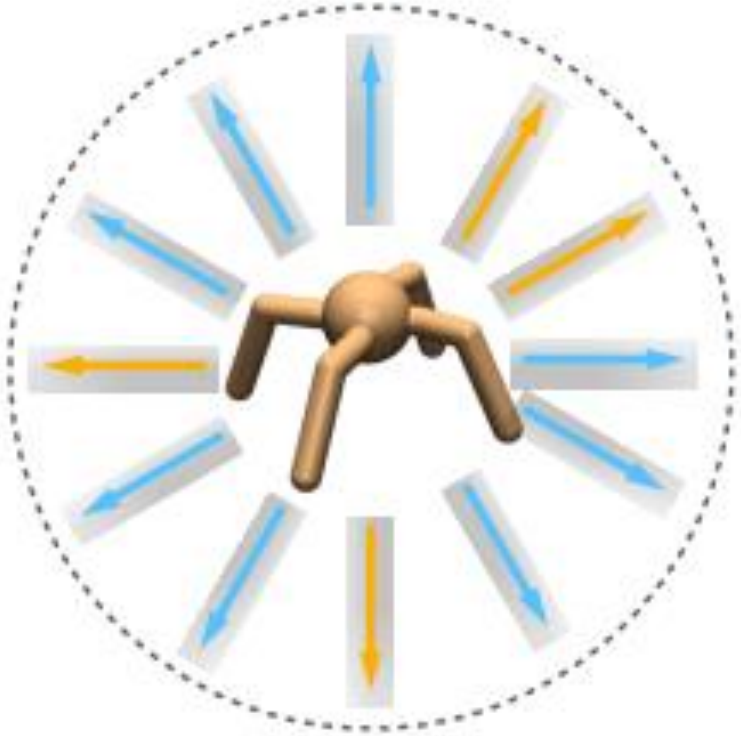
We estimate the value function $V(\mathbf{s},\mathbf{z}) = \mathbb{E}_{\mathbf{a} \sim \pi_\theta(\mathbf{a}|\mathbf{s},\mathbf{z})} Q(\mathbf{s},\mathbf{a},\mathbf{z})$ with a single sample, and λ is the resulting Lagrange multiplier for the optimization problem. See Nair et al. (2020) for a full derivation.

<활용>

논문 실험

- 본 실험에서는 치타 속도, 개미 방향, 휴머노이드, Walker Param, Hopper Param, 문제를 이전 Online meta-RL task에 기반해서 평가한다.





STUDY PRESENTATION

THANK YOU

발표를 마치겠습니다

궁금한 점은 바로 질문해 주세요!