SOGANG UNIVERSITY

Gymnasium –Humanoid v5

박재형 120250384

Github link: https://github.com/wowharry1/humanoid-rl-project

프로젝트 주제 및 목표

프로젝트 주제

Gymnasium Humanoid-v5 환경에서 3D 휴머노이드 로봇이 자율적으로 보행하도록 강화학습(RL) 기반 제어 정책을 학습

프로젝트 목표

- 알고리즘 비교
  - PPO, SAC, TD3 알고리즘을 동일한 조건에서 학습
  - 수렴 속도, 최종 성능, 안정성을 비교
- 보상 함수(Reward shaping) 실험
  - Torque penalty를 추가하여
    - 에너지 효율적인 보행
    - 보다 안정적이고 자연스러운 움직임에 미치는 영향 분석
- 재현 가능한 실험 파이프라인 구축
  - 공통 코드 구조(train / eval) 설계
  - 여러 seed에서 실험하여 성능 분산(variance)까지 평가

환경 및 데이터셋 설명 (Humanoid-v5)

환경: Gymnasium — Humanoid-v5

- 3D 휴머노이드 로봇의 보행 제어 시뮬레이션 환경
- 관절 17개 / 연속 액션 공간 (torque control)
- 보상 항목에 포함되는 요소
  - 전진 속도
  - 넘어지지 않을수록 높은 보상
  - 에너지/토크 사용량 패널티

데이터셋

- 실제 데이터셋 사용되지 않음
- 로봇이 시뮬레이션 안에서 움직이는 과정에서 상태, 행동, 보상 데이터가 계속 생성되는 형태로 학습됨.

State, Action, Reward 설계 설명

State

Humanoid-v5 환경에서 하나의 상태 벡터는 다음을 포함:

- 몸 중심 위치 및 속도
- 관절 각도 / 각속도
- 지면 접촉 정보
- 관절 제한 및 충돌 정보

Action

- 17개 관절에 대해 지속적인 torque 값을 출력
- 범위: 각 관절별 [-1, 1] (정규화된 torque 명령). 곧 행동 명령

Reward

기본 환경 보상:

- 전진 이동 속도 보상
- 넘어지면 즉시 큰 페널티
- 에너지 과다 사용 / 불안정한 진동 패널티 포함

추가 보상:

- 추가 요소: Torque Penalty (에너지 최소화)

State, Action, Reward 설계 설명

알고리즘	특징	기대 효과
PPO (Proximal Policy Optimization)	On-policy / 안정적 학습 / 가장 널리 사용	기준선 알고리즘 역할
SAC (Soft Actor-Critic)	Off-policy / 엔트로피 기반 탐색 / 높은 샘플 효율	높은 최고 성능 기대
TD3 (Twin Delayed DDPG)	Off-policy / Q-function overestimation 방지	정밀 제어 성능 기대

공통 Hyperparameters

항목	값
Total timesteps	300,000 per algorithm
Discount factor (γ)	0.99
Optimizer	Adam
Policy	MLP (2–3 fully-connected layers)
Seeds	3 (0 / 10 / 20)
Reward shaping	torque_penalty

강화학습 알고리즘 및 Hyperparameters

알고리즘별 주요 Hyperparameters

알고리즘	핵심 Hyperparameters
PPO	n_steps, batch_size, clip_range, gae_lambda, learning_rate
SAC	buffer_size, alpha (temperature), tau, train_freq, learning_rate
TD3	policy_delay, target_noise, noise_clip, buffer_size, learning_rate

동일한 조건을 유지하기 위해

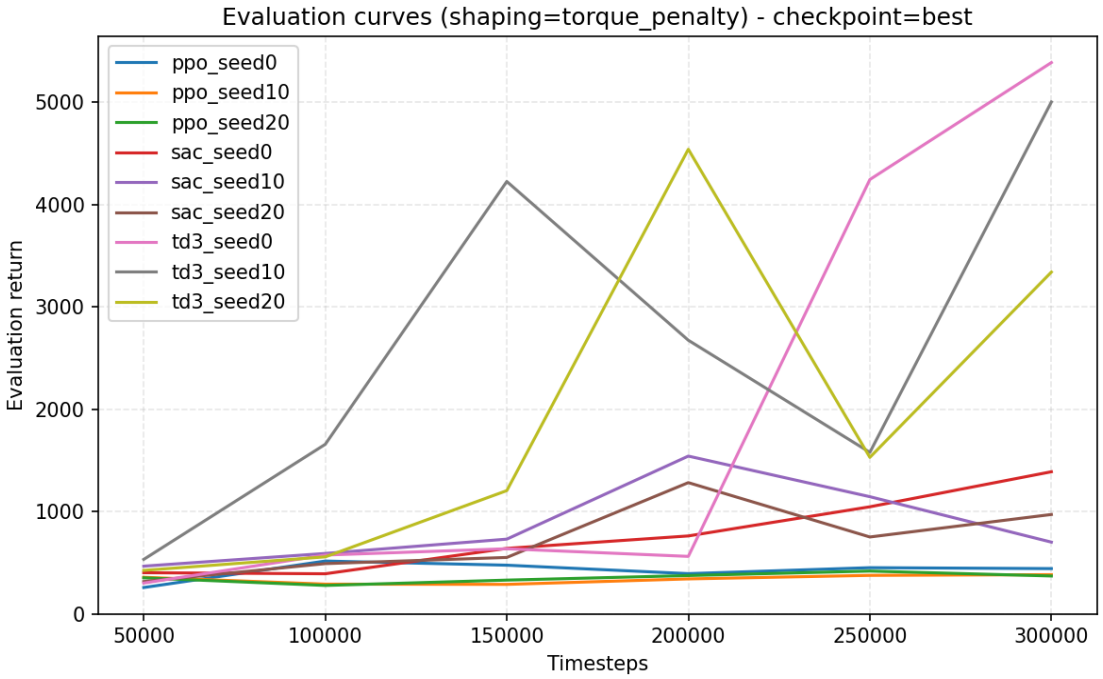
- 학습 환경, 보상, seed, total timesteps, policy architecture는 통일
- 각 알고리즘 고유 파라미터만 합리적인 기본값으로 유지

실험 셋업

실험 환경

항목	내용
강화학습 환경	Gymnasium — Humanoid-v5
라이브러리	Stable-Baselines3 (SB3)
시뮬레이터	MuJoCo
관측 공간	376-dim continuous state
행동 공간	17-dim continuous torque control
보상	기본 보상 + 보상 패널티(Shaped reward)
종료 조건	로봇 넘어짐 or 시간 제한

실험 결과



그래프 해석

- TD3는 성능 상승 폭이 크고 빠르지만, SAC과 PPO는 매우 완만한 향상을 보임
- TD3는 seed 간 편차가 매우 큼 → seed10은 최고 성능을 보였지만 seed0 · seed20은 불안정
- SAC은 비교적 일관된 상승 곡선을 보이며 안정적으로 학습
- PPO는 세 seed 모두 낮은 편차, 낮은 성능 → 가장 안정적이지만 가장 낮은 return

결과

TD3는 성능이 좋지만 불안정. SAC는 안정적으로 학습됨. PPO는 거의 학습을 못 함.

결론

- 연속 제어가 복잡한 Humanoid 환경에서는 단순 On-policy 업데이트(PPO)보다 Off-policy + entropy/target smoothing 기반의 SAC/TD3가 훨씬 효과적
- TD3의 제일 좋은 성능은 가장 높았지만 seed에 매우 민감하여 실험 재현성이 떨어지는 trade-off 존재
- SAC은 성능/안정성/재현성 모두 균형적 → 종합적으로 가장 우수
- PPO는 안정성을 유지했지만 정책 업데이트 크기가 제한적이라 고난도 Humanoid 문제에서는 학습력이 떨어짐

개선 방안

- Rewarding Shaping을 했을 때와 안 했을 때의 차이점을 봐, reward shaping 효과가 있었는지에 대해서 확인