PPO 알고리즘

(Proximal Policy Optimization) 중심부 정책 최적화

- $\boxed{\text{TRPO}}$ Trust Region Policy Optimization (신뢰 영역 정책 최적화)을 연산에 용이하도록 알고리즘 수정한것
- Actor 와 Critic 을 이용해서 최적화

※ Reinforcement Learning 에서의 목표함수 Expected Reward

$$\eta(\pi) = \hat{E}_t \left[\Sigma \log \pi_\theta (a_t | s_t) \hat{A}_t \right]$$

→ t의 시점에서 추정되는 Advantabe

↳ Stochastic Policy [확률 정책]

↳ 파라미터 θ에 대해서 Gradient를 사용하여 최대화하는 방향으로 업데이트 한다.