diversity is all your need

陈前辛小雨

tju

2025 年 6 月 6 日

Motivation

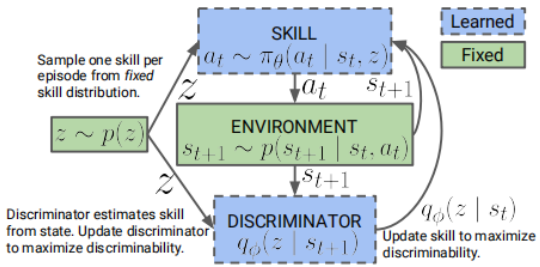
Current DRL can effectively learn skills driven by rewards, while intelligent creatures can explore the environment and learn skills in an unsupervised manner. As a result, they can use these skills to quickly and effectively meet new, specific goals.

Related Work

- Used to solve exploration problems in sparse reward environments
- For long-horizon tasks, unsupervised learned skills can serve as primitives in hierarchical reinforcement learning to reduce episode length (high-level policies select skills, and low-level skills execute actions)
- Used in scenarios where interacting with the environment is free but evaluating rewards requires human feedback, reducing the amount of supervision needed for learning tasks
- Used to determine which tasks an agent should learn in an unfamiliar environment (unsupervised emergence of diverse skills)

Contribution

- Proposes the DIAYN method, which can learn useful skills without a reward function
- Presents a concise optimization objective that induces the unsupervised emergence of diverse skills, such as walking and jumping; capable of solving benchmark tasks without receiving real task rewards
- Proposes methods for applying learned skills to hierarchical reinforcement learning and imitation learning
- Demonstrates how to quickly apply discovered skills to solve new tasks



Algorithm 1: DIAYN

while *not converged* **do**
  Sample skill $z \sim p(z)$ and initial state $s_0 \sim p_0(s)$
  for $t \leftarrow 1$ **to** *steps_per_episode* **do**
    Sample action $a_t \sim \pi_\theta(a_t \mid s_t, z)$ from skill.
    Step environment: $s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$.
    Compute $q_\phi(z \mid s_{t+1})$ with discriminator.
    Set skill reward $r_t = \log q_\phi(z \mid s_{t+1}) - \log p(z)$
    Update policy (θ) to maximize r_t with SAC.
    Update discriminator (ϕ) with SGD.

Core Ideas and Approach

Core Ideas

- Skills guide the states visited by the agent, with different skills visiting different states, thus making skills distinguishable $s \sim \pi(z)$.
- Skills need to be as random as possible to encourage exploration, while being as diverse as possible, and also need to avoid randomness affecting distinguishability (stay away from the state spaces of other skills).
- Use state rather than action to distinguish skills, because actions that do not change the environment are unobservable.

Approach

- Use maximum entropy policies and maximize an information-theoretic distinguishability objective to learn skills.
- Maximize the mutual information between skills and states to enable skills to control the agent to visit corresponding states.
- Maximize the entropy of the mixed policy (the mixed policy is the prior distribution of all skills and latent variables).
- Minimize the mutual information between actions and skills given states, so that skills are distinguished by states rather than actions.

$$
\begin{aligned}
\mathcal{F}(\theta) &\triangleq I(S; Z) + \mathcal{H}[A \mid S] - I(A; Z \mid S) \\
&= (\mathcal{H}[Z] - \mathcal{H}[Z \mid S]) + \mathcal{H}[A \mid S] - (\mathcal{H}[A \mid S] - \mathcal{H}[A \mid S, Z]) \\
&= \mathcal{H}[Z] - \mathcal{H}[Z \mid S] + \mathcal{H}[A \mid S, Z]
\end{aligned}
$$