[POSTER] E-DIAYN: Evolution Strategies for Unsupervised Skill Discovery in Reinforcement Learning

Waris Radji

E-DIAYN: Evolution Strategies for Unsupervised Skill Discovery in Reinforcement Learning

Waris Radji - supervised by Théo Matricon

CS Department - Bordeaux Institute of Technology - ENSEIRB-MATMECA, France

Motivation

- Deep reinforcement learning (RL) has been demonstrated to effectively learn a wide range of reward driven skills, including playing games, controlling robots, and navigating complex environments.
- In nature, intelligent creatures can explore their environments and learn useful skills even without supervision and use those skills to satisfy the new goals quickly and efficiently.
- Environments with sparse rewards have no reward until the agent randomly reaches a goal state. Learning useful skills without supervision may help address challenges in exploration in these environments.

Diversity is All You Need (DIAYN) [1]

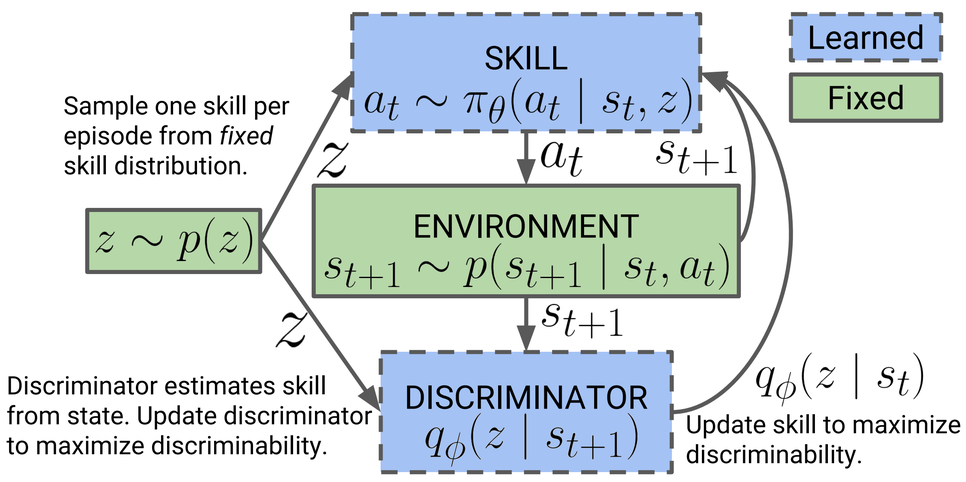

Figure: **DIAYN Algorithm**: We update the discriminator q_ϕ to better predict the skill, and update the skill to visit diverse states that make it more discriminable.

Evolution Strategies (ES) for Reinforcement Learning [2]

An alternative approach to solving RL problems is using black-box optimization algorithms like Evolution Strategies:

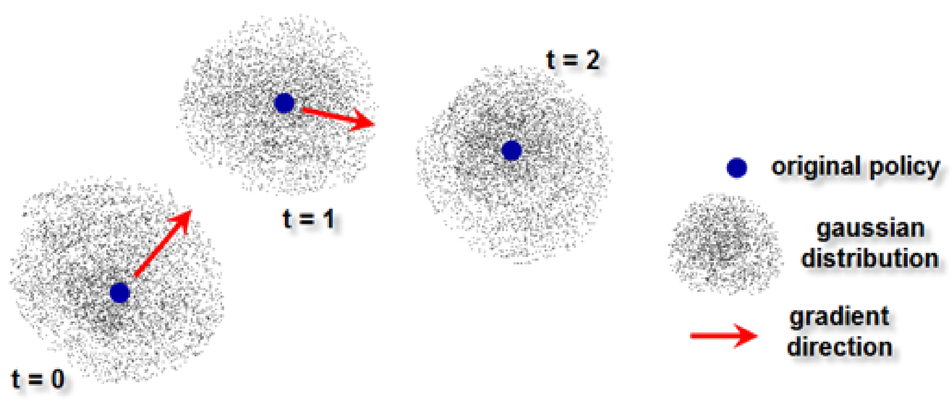

Figure: Behavior of Evolution Strategies.

- No back propagation and policy does not need to be differentiable.
- Better exploration behavior.
- Extremely scalable, due to lack of massive internal data exchange.
- The gradient θ_{t+1} is estimated by averaging *n* perturbations ϵ_i, and weighting them by their scores (given by running a rollout in the environment with $\pi_{\theta+\epsilon_i}$).

Novelty Search (NS) [3]

Inspired by nature's drive towards diversity, NS encourages policies to engage in different behaviors than those previously seen.

NS assigns a domain-dependent behavior characterization to a policy $b(\pi)$, for example, in the case of an ant navigation problem, it may be as simple as a two-dimensional vector containing the ant's final location. The novelty of a policy is computed by selecting the k-nearest neighbors from an archive set *A* and computing the average distance between them:

$N(b(\pi_\theta), A) = \frac{1}{|S|} \sum_{j \in S} ||b(\pi_\theta) - b(\pi_j)||_2$ with $S = kNN(b(\pi_\theta, A))$

The E-DIAYN Algorithm

We propose E-DIAYN, an alternative to DIAYN which is built on top of evolutionary strategies instead of SAC (a reference RL algorithm for optimal policy via entropy maximization), in order to better imitate how intelligent creatures evolve in nature and take profit of the scalability of ES.
Instead of trying to maximizing the entropy of the policy, we use NS to find a diverse set of solutions that may not necessarily be the most optimal but can offer new perspectives and ways of solving the problem. The new intrinsic reward is given by :

$$r_t = N(b(\pi_\theta), A) \times (\log q_\phi(z \mid s_{t+1}) - \log p(z))$$

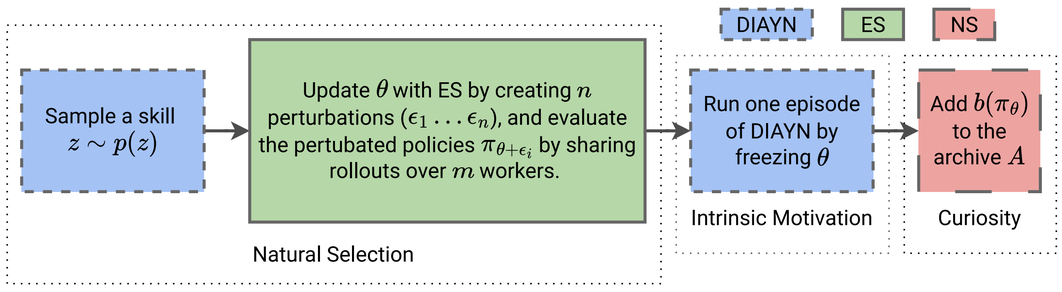

Figure: Episode of the E-DIAYN algorithm

E-DIAYN Early Results

DIAYN and E-DIAYN agents are trained for 20 skills, with the same parameters and seed, on the MuJoCo Ant environment. The trajectories of the ant were tracked on a limited space to compute a state coverage.

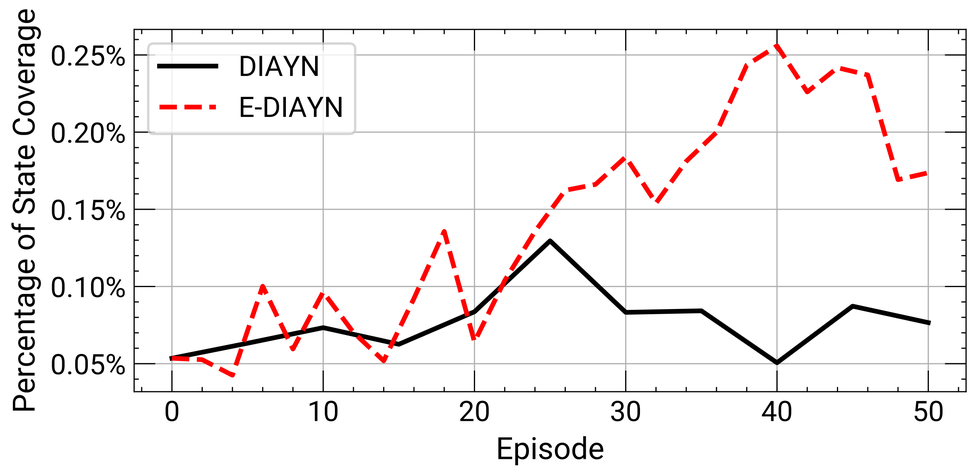

Figure: Comparing percentage of state coverage over episodes.

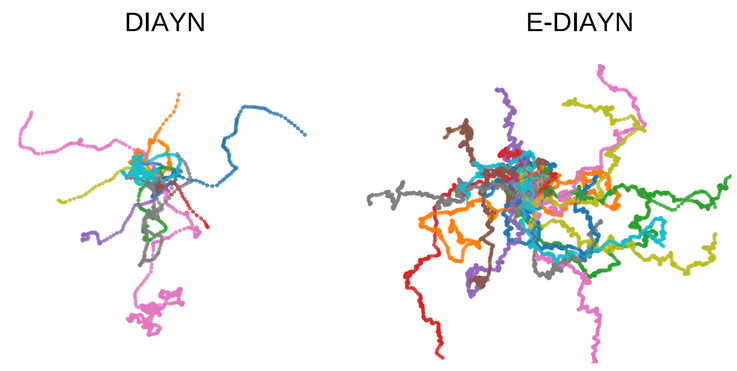

Figure: Trajectories of DIAYN and E-DIAYN Skills after 50 episodes.

Promising results, due to the high amount of rollouts in E-DIAYN (500 shared on 56 cores), compared to DIAYN who performs only one rollout per episode. With a number of rollouts equivalent to the number of cores, E-DIAYN has a similar execution time to DIAYN.

References

[1] Eysenbach et al. (ICLR 2018) Diversity is All You Need: Learning Skills without a Reward Function;
[2] Salimans et al. (2017) Evolution Strategies as a Scalable Alternative to Reinforcement Learning;
[3] Conti et al. (NeurIPS 2018) Improving Exploration in Evolution Strategies for Deep Reinforcement Learning via a Population of Novelty-Seeking Agents;