Robotics Reading Group

Session #11
24-07-2020

Giuseppe Paolo

Unsupervised Learning and Exploration of Reachable Outcome Space

G. Paolo, A. Coninx, S. Doncieux, A. Laflaquière

https://arxiv.org/abs/1909.05508

SCIENCES SORBONNE UNIVERSITÉ
cnrs
ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE
SoftBank Robotics

Background: Sparse Rewards

Reinforcement Learning with Hindsight Experience Replay

Or Rivlin Follow

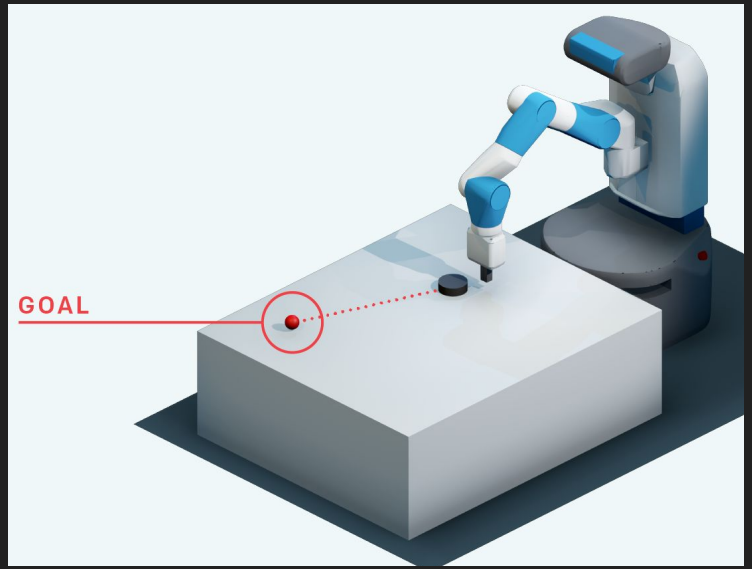
Jan 31, 2019 · 10 min read

AI

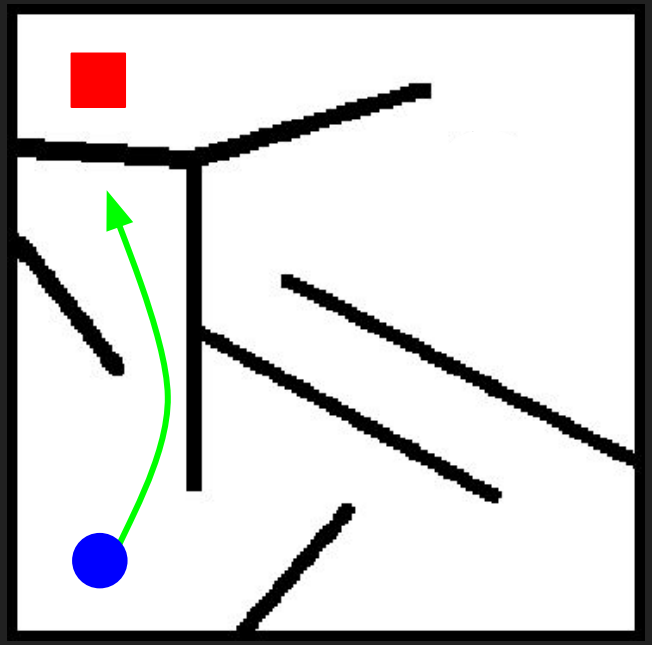
Montezuma's Revenge Solved by Go-Explore, a New Algorithm for Hard-Exploration Problems (Sets Records on Pitfall, Too)

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune

November 26, 2018



Background: Sparse Rewards



Good strategy is to ignore the reward and focus on exploration.

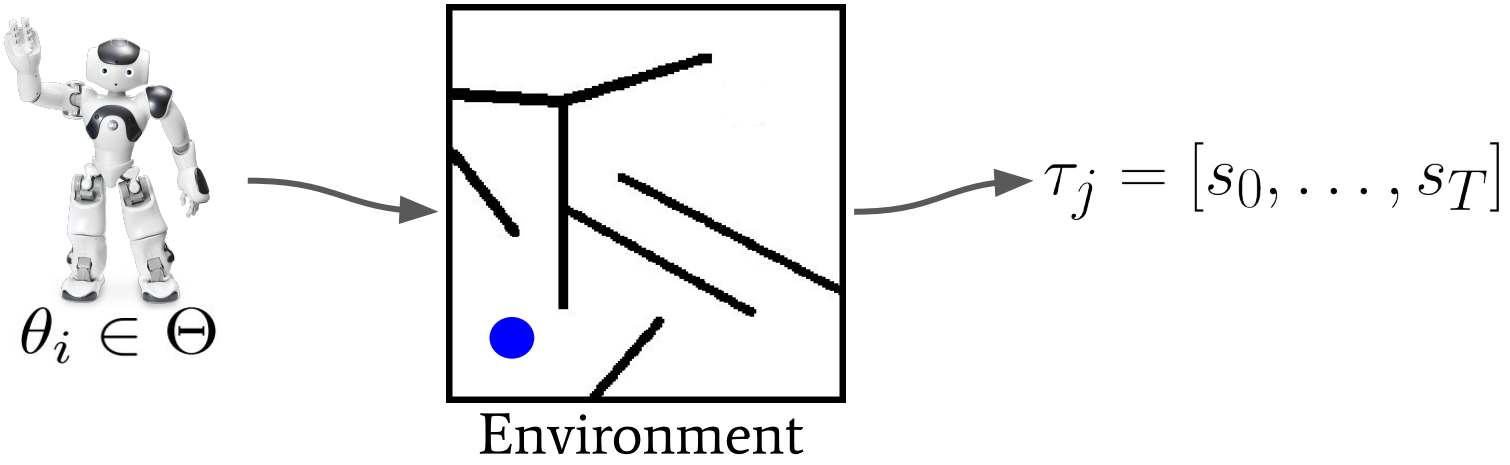
Divergent policy search methods focus on exploring the space of possible policies.

- Population based algorithms
- The search is driven through a measure of novelty[1], surprise[2] or diversity[3].

[1] Lehman, Joel, and Kenneth O. Stanley. "Exploiting open-endedness to solve problems through the search for novelty." ALIFE. 2008.
[2] Gravina, Daniele, Antonios Liapis, and Georgios Yannakakis. "Surprise search: Beyond objectives and novelty." Proceedings of the Genetic and Evolutionary Computation Conference 2016. 2016.
[3] Mouret, J.-B. and Doncieux, S. (2012). Encouraging Behavioral Diversity in Evolutionary Robotics: an Empirical Study. Evolutionary Computation. Vol 20 No 1 Pages 91-133.

Background: Novelty Search

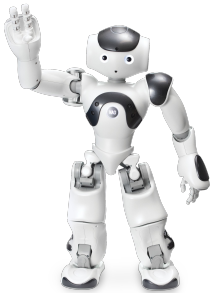
Novelty search[1] performs the search in a hand designed low-dimensional outcome space.



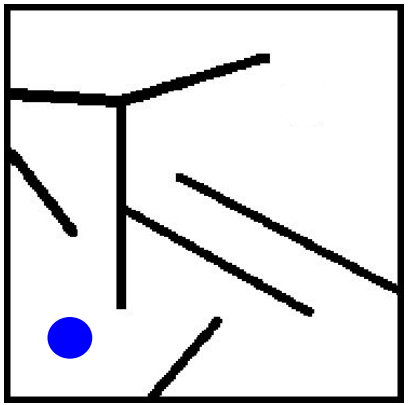
[1] Lehman, Joel, and Kenneth O. Stanley. "Exploiting open-endedness to solve problems through the search for novelty." ALIFE. 2008.

Background: Novelty Search

Novelty search[1] performs the search in a hand designed low-dimensional outcome space.



$\theta_i \in \Theta$



Environment

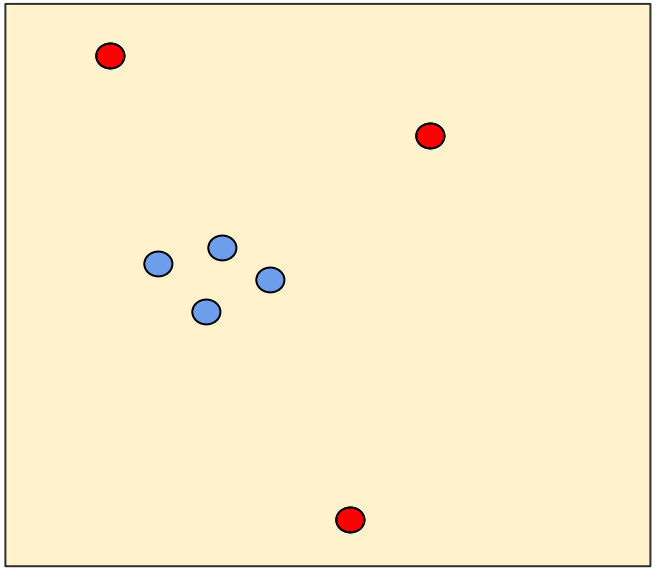
$\tau_j = [s_0, \ldots, s_T]$

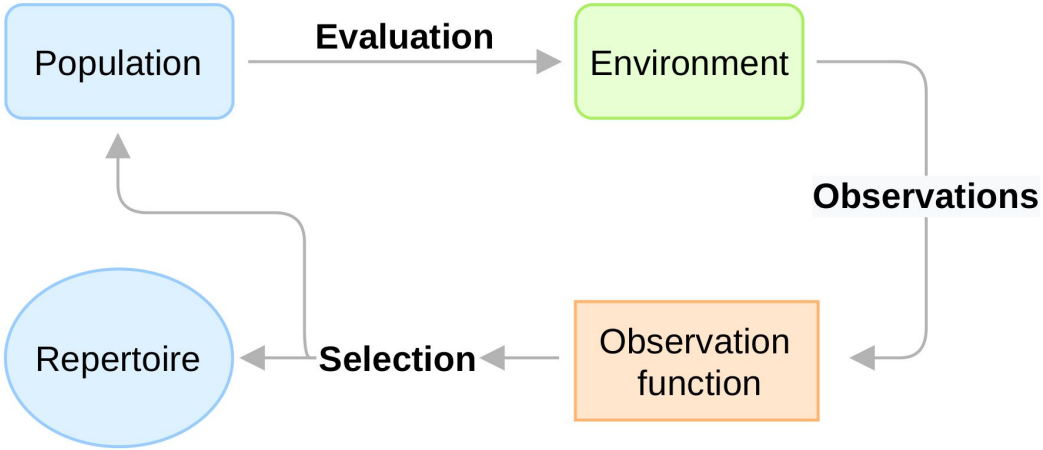
$$O_B(\tau_j) = b_j$$

$$\phi_B(\theta_j) = b_j$$

[1] Lehman, Joel, and Kenneth O. Stanley. "Exploiting open-endedness to solve problems through the search for novelty." ALIFE. 2008.

Background: Novelty Search



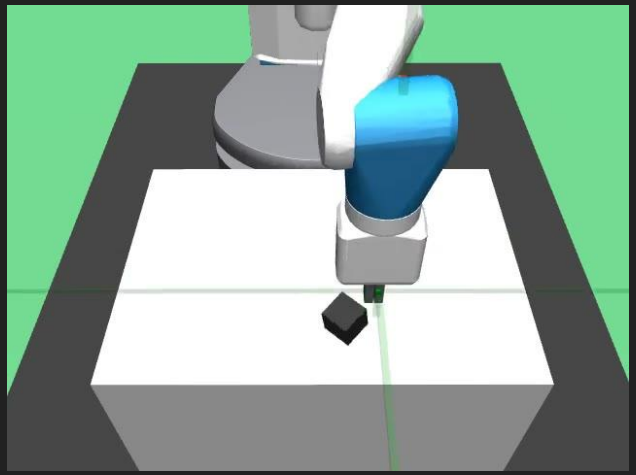


$$\eta_j = \frac{1}{k} \sum_{i=1}^{k} dist(\phi_B(\theta_j), \phi_B(\theta_i))$$

Designing the outcome space

Design of the outcome space and observer function can be problematic.

- Huge amount of prior knowledge
- Designer induced bias
- Important features not always obvious



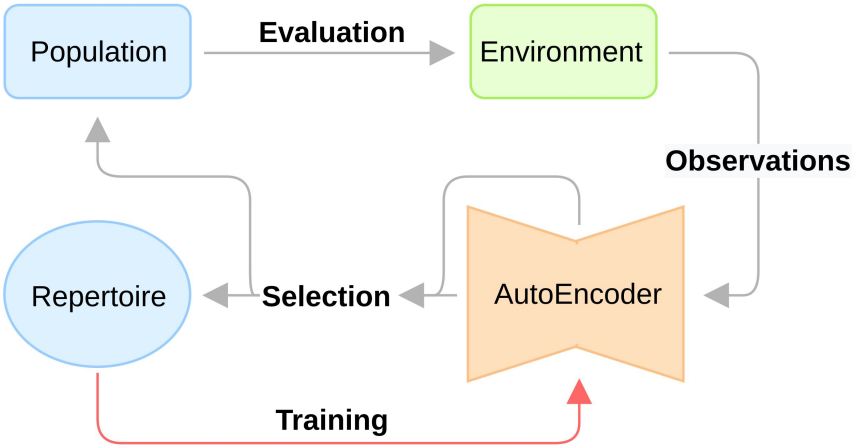
What can be done?

TAXONS

Autoencoder learns low-dimensional features from last observation of trajectory

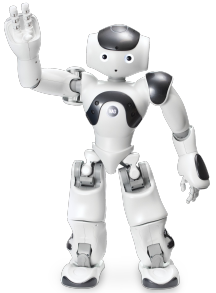
$$\mathcal{E} : \mathcal{O} \rightarrow F$$
$$D : F \rightarrow \mathcal{O}$$

- Feature space is the outcome space
- Encoder is the observer function

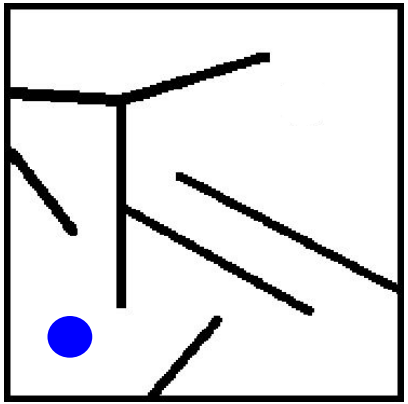


Task Agnostic eXploration of Outcome space through Novelty and Surprise

TAXONS

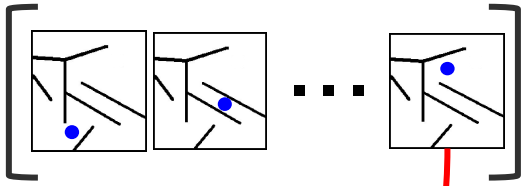


$\theta_i \in \Theta$



Environment

$\tau_j =$



Outcome descriptor: b_j

Reconstruction error

AutoEncoder

TAXONS: Novelty and Surprise

Policy selection done through:

Novelty

$$n(\theta_i) = \frac{1}{k}\sum_{j=1}^{k} dist(f(\theta_i), f(\theta_j)) \quad \text{with} \quad f(\theta_i) = \mathcal{E}(o_T^{(\theta_i)})$$

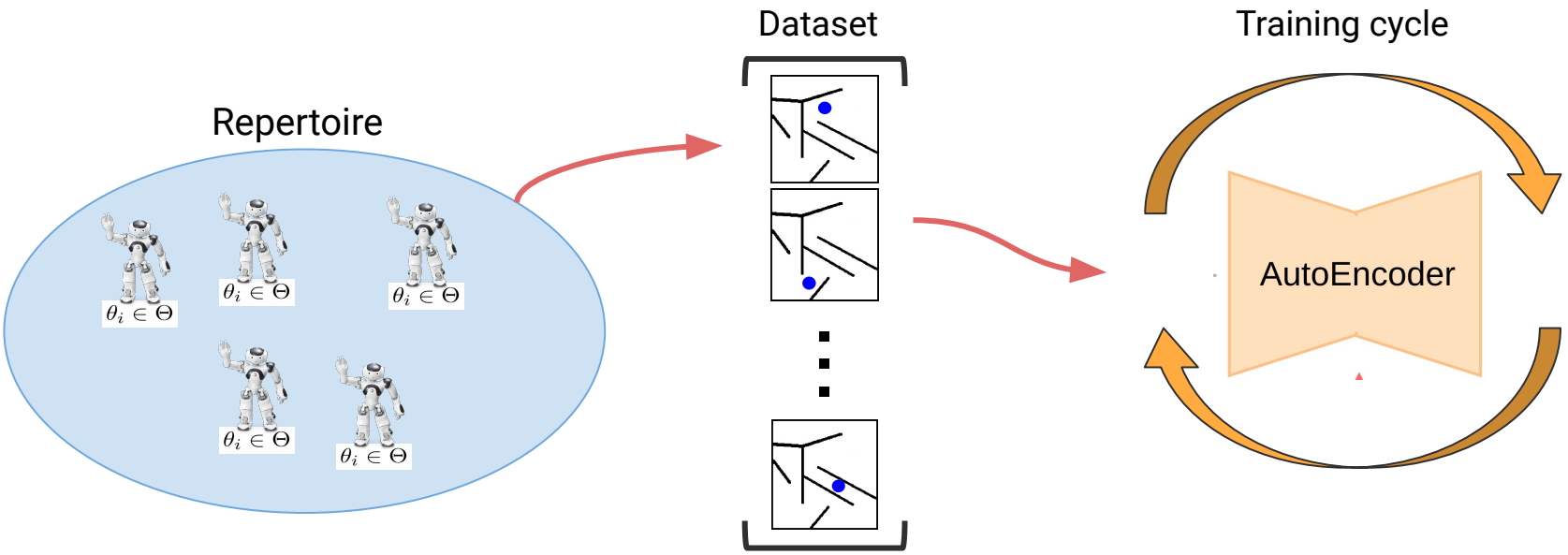
Surprise[1]

$$s(\theta_i) = ||o_T^{(\theta_i)} - \mathcal{D}(\mathcal{E}(o_T^{\theta_i}))||_2^2$$

Random choice between the two.[2]

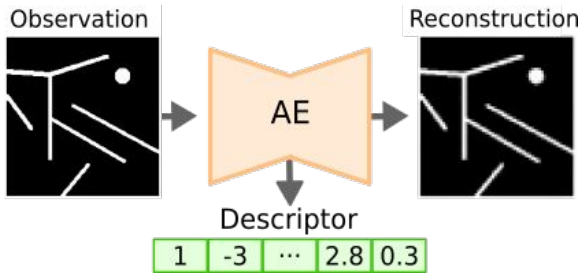
[1] Gravina, Daniele, Antonios Liapis, and Georgios N. Yannakakis. "Surprise search for evolutionary divergence." arXiv preprint arXiv:1706.02556 (2017).
[2] Stephane Doncieux and Jean-Baptiste Mouret. “Behavioral diversity with multiple behavioral distances”. In: 2013 IEEE Congress on Evolutionary Computation. IEEE. 2013, pp. 1427–1434

TAXONS: training of AE

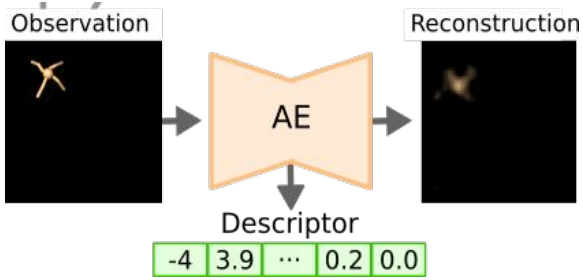


Experiment setup

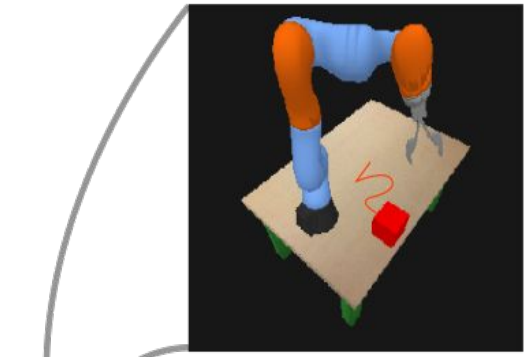
Maze

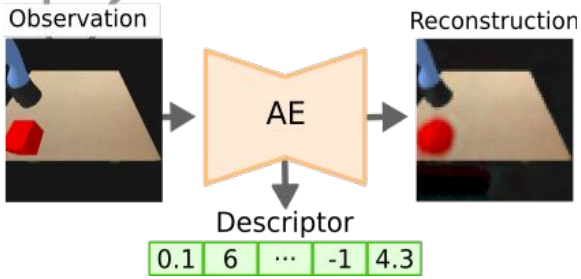


Ant



Kuka





Experimental setup

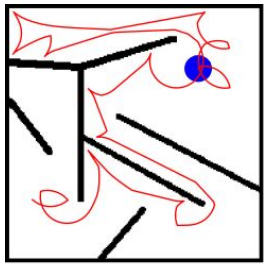
- **TAXONS**: our method using both novelty and surprise
- **TAXOS**: our method using only surprise
- **TAXON**: our method using only novelty

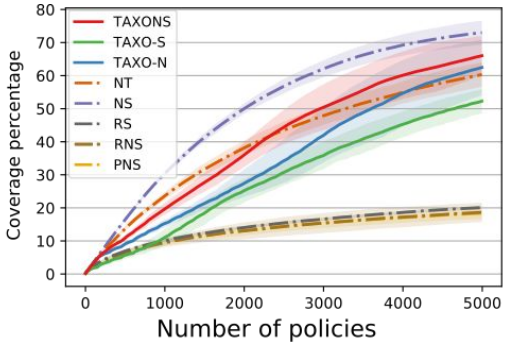
- **NT**: our method, in which the autoencoder is not trained
- **NS**: vanilla novelty search, calculates novelty on the ground truth final position of the robots (for the Maze and And) and of the box (for the Kuka)
- **RNS**: novelty search in which the outcome descriptor is a random 10D vector
- **PNS**: novelty search in which the novelty is calculated on the policy parameters

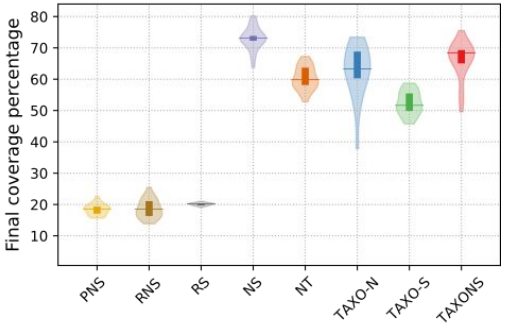
- **RS**: random search in which the policies are all randomly generated

Results

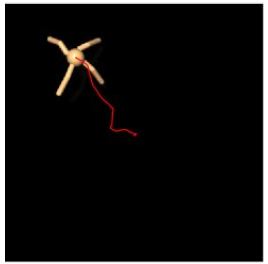
Maze

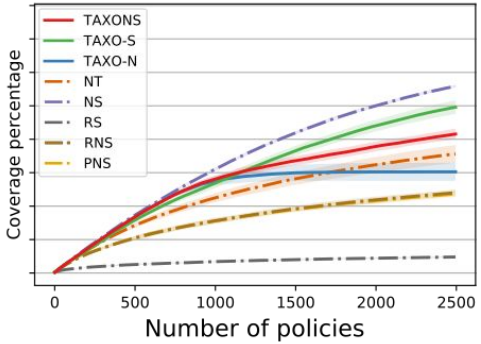


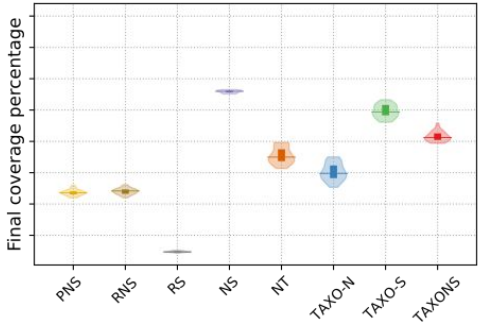




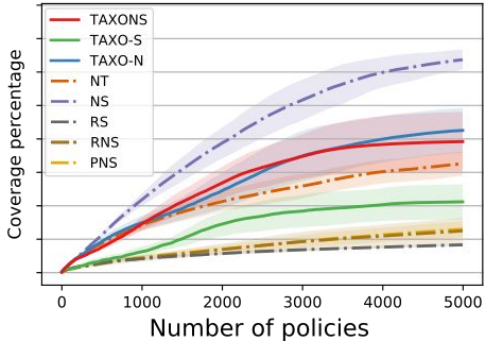
Ant

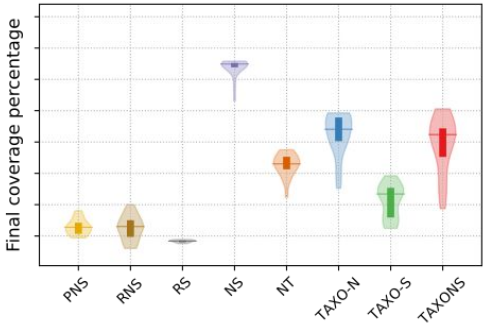






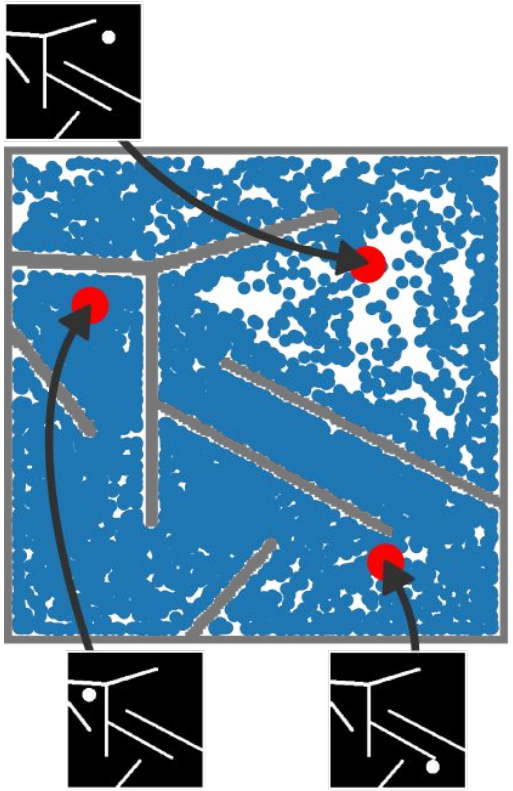
Kuka



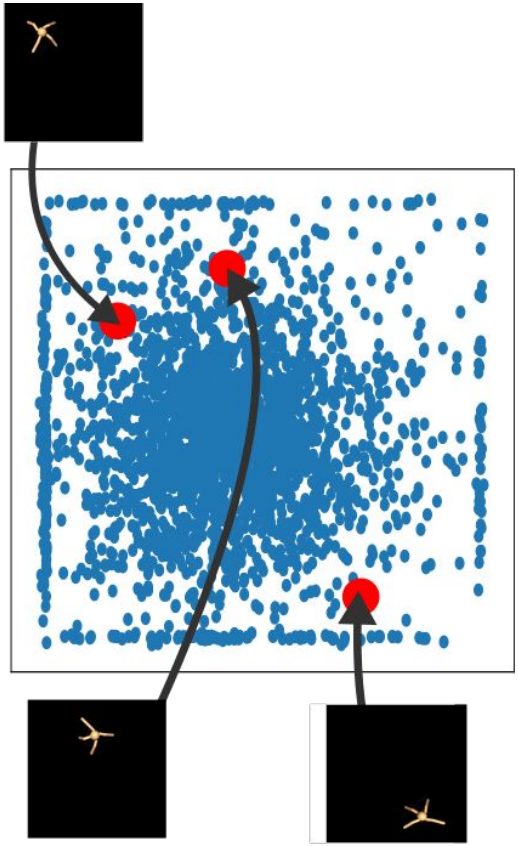


Results

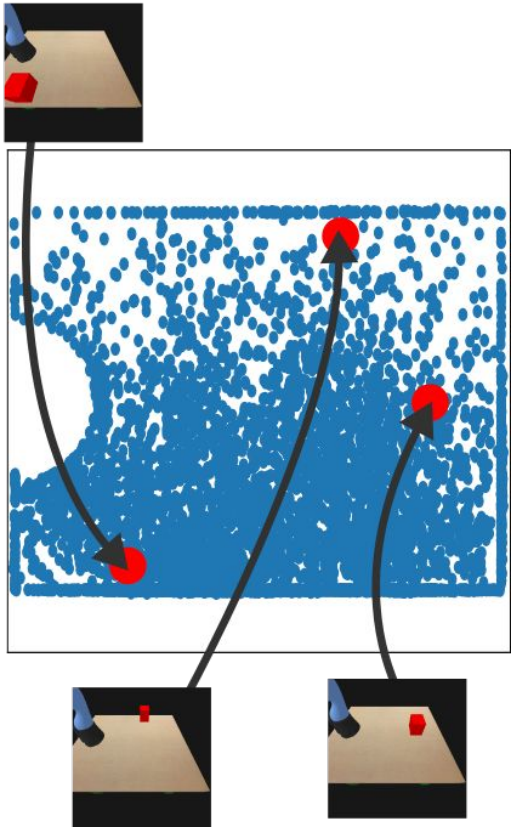
Maze



Ant



Kuka



Conclusions

- Autonomously building the low-dimensional outcome space from high-dimensional observations reduces the amount of information needed at design time
- Combining two evaluation metrics makes the exploration process more robust
- No reward signal is needed
- Policies found can the be used later to solve tasks given in the environment

Main assumption: the last observation is enough to characterize the behavior of a policy

Unsupervised Learning and Exploration of Reachable Outcome Space

giuseppe.paolo@softbankrobotics.com

https://sites.google.com/view/gpaolo

Get the paper! →

Thanks!

Robotics Reading Group

Giuseppe Paolo