The influence of exploratory behaviour on temporal abstraction and subgoal identification

Veronica Ioana Chelu

University Politehnica Bucharest

Motivation

Hierarchy in agent behaviour:

- High-level - context, generalization
- Low-level - abstracts details

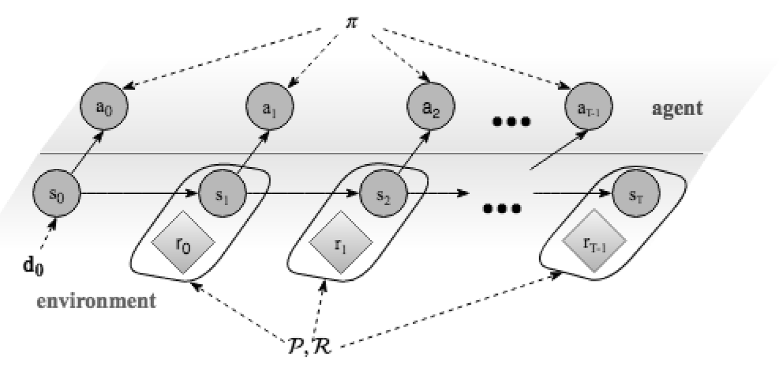
Subgoals

- Directions in latent space for the higher level policy
- Lower level policy receives pseudo-reward for maximizing progress in the direction indicated by the former.

Background

Markov Decision Process
$< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$,

- $\mathcal{S}$ is the states space.
- $\mathcal{A}$ is the action space.
- $\mathcal{P}(s'|s,a)$ is a transition probability distribution.
- $\mathcal{R}$ is the reward emission probability distribution.
- γ is the discount factor



Options framework [5, 1]

Option - a triple $O =< I_o, \pi_o, \beta_o >$:

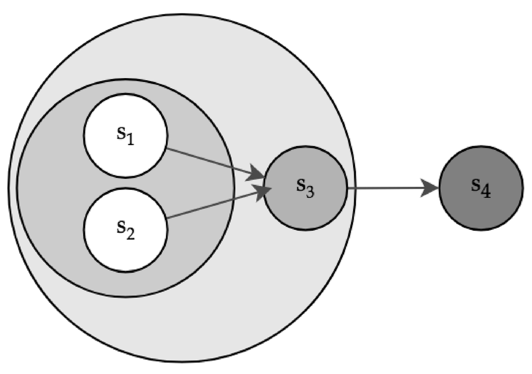
- $I_o \in \mathcal{S}$ is the initiation set of an option,
- $\beta_o : \mathcal{S} \to [0,1]$ is the stochastic termination condition of an option,
- $\pi_o : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the stochastic intra-option policy of an option.

Subgoal discovery

Successor representations & **features** [3, 2, 4]

- representation of **future timeline**
- **prediction** about the **future occurrence of the subsequent states / features** reached under a policy
- under a random policy, capture the **topology of the environment**

$$\Psi^\pi_{s_t} = \mathbb{E}_\pi[\Phi_{s_t} + \gamma \Psi^\pi_{s_{t+1}}],$$



Direction-based Option-Critic

Next observation prediction

$$d\xi = d\xi - \alpha_\xi \nabla_\theta [(s(\hat{\theta})_{t+1} - s_{t+1})^2]$$

Successor features prediction

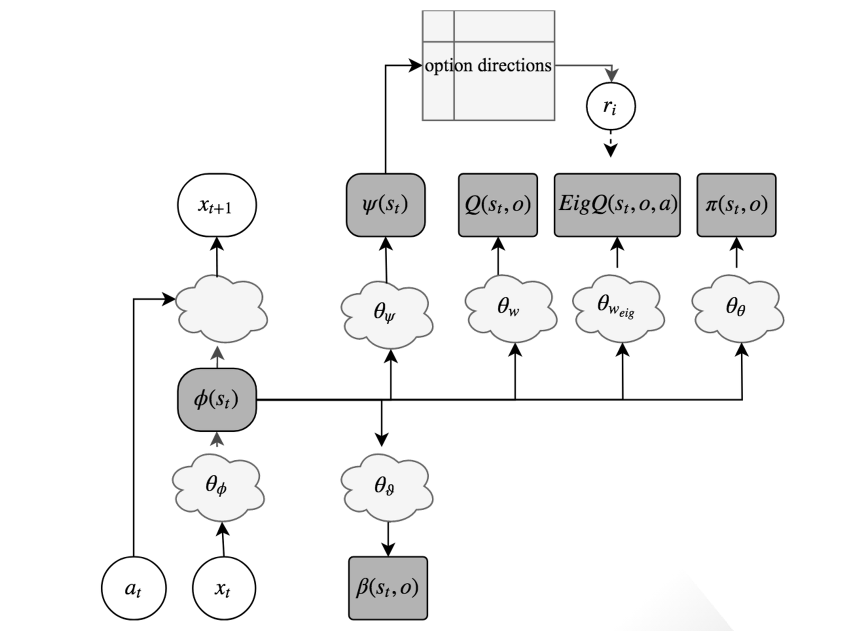
$$d\psi \leftarrow d\psi - \alpha_\psi \nabla_\psi [(\phi(s_k) + \gamma\psi_\psi(s_{k+1}) - \psi_\psi(s_k))^2],$$

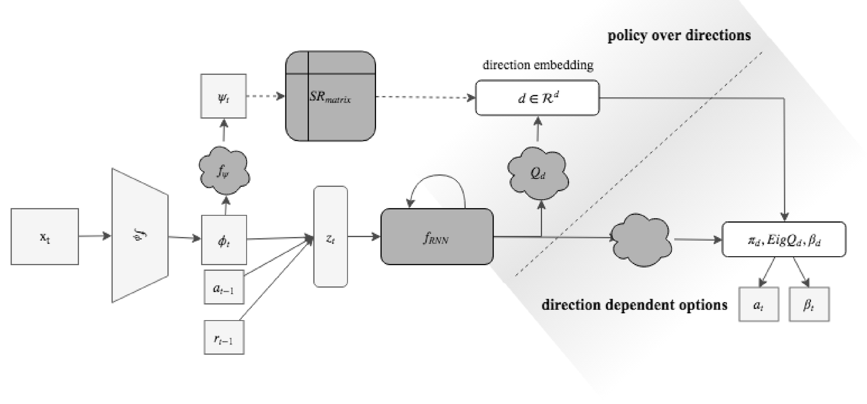
Mixed pseudo-reward signal

$$r_{mix}(s,a,o) = \alpha * r_i(s,s',o) + (1-\alpha) * r_e(s,a)$$

Critics, intra-option policies and termination conditions

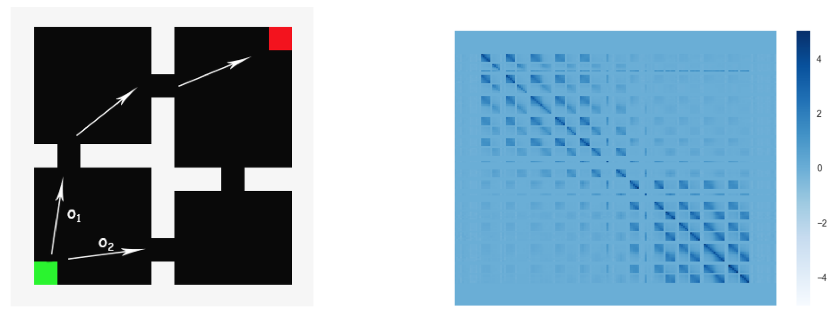
$$dw \leftarrow dw - \alpha_w \nabla_w [(R - Q_w(s_k,o_k))^2]$$
$$d\theta \leftarrow d\theta + \alpha_\theta \nabla_\theta [log\pi_\theta(a_k|s_k,o_k)](R - Q_w(s_k,o_k))$$
$$d\vartheta \leftarrow d\vartheta + \alpha_\vartheta \nabla_\vartheta [\beta_\vartheta(s_k)](Q_w(s_k,o_k) - V_w(s_k) + \eta)$$
$$dw_{eig} \leftarrow dw_{eig} - \alpha_{w_{wig}} \nabla_{w_{eig}} [(R_{mix} - EigQ_{w_{eig}}(s_k,o_k))^2]$$



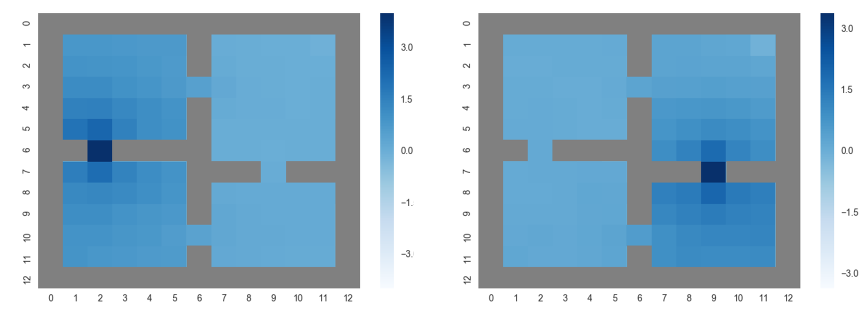


Experiments

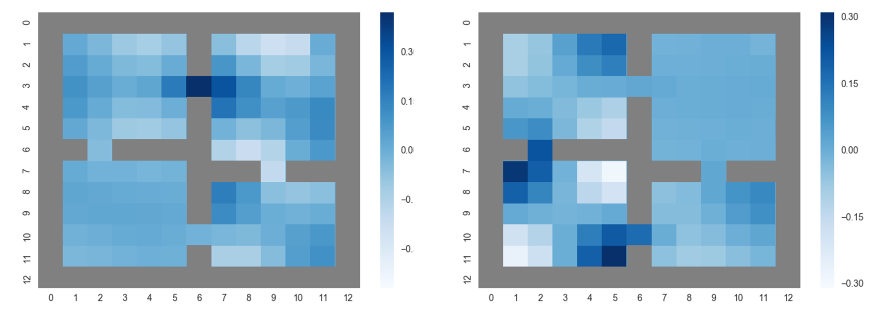
Autonomous discovery of bottleneck and salient information states using **one-hot states**.



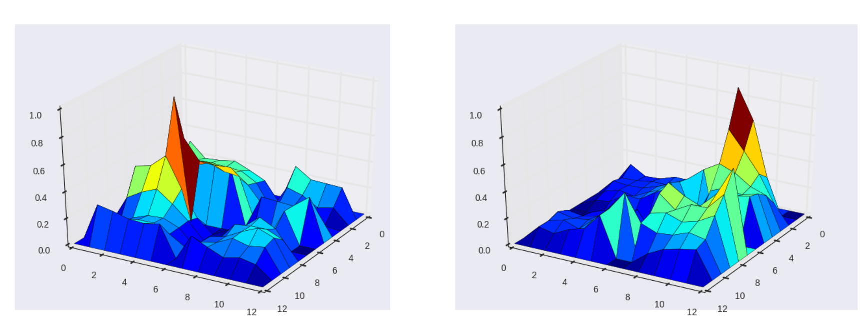
SR vectors plotted over the environment.



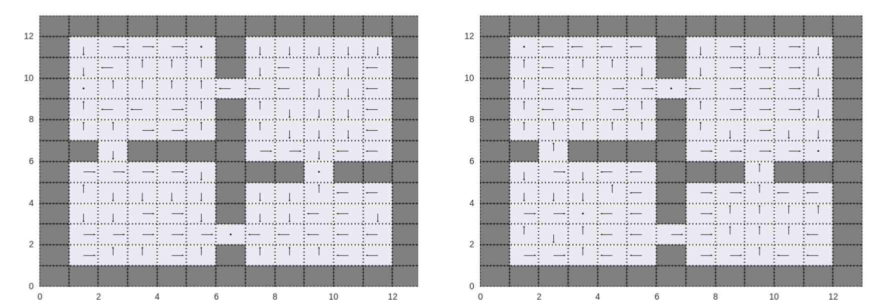
Eigenvectors of the SR matrix plotted over the environment.



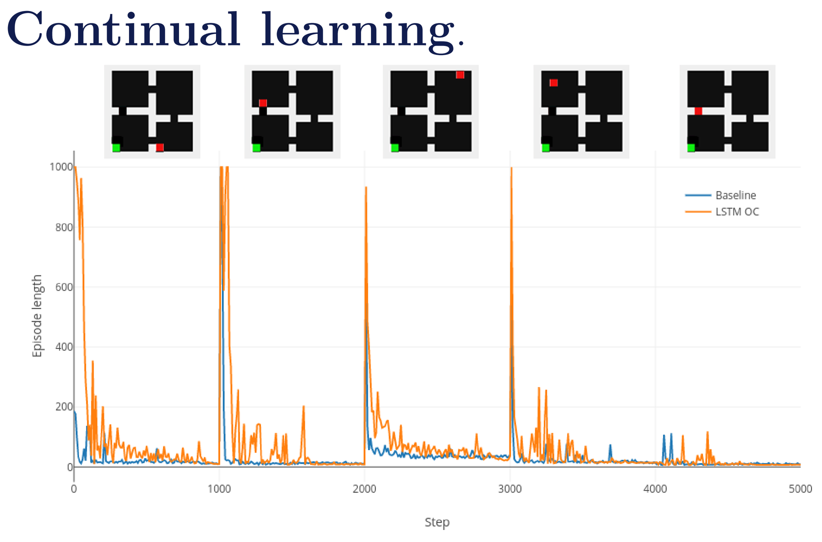
Autonomous discovery of bottleneck and salient information states using **function approximation** for state features.
Value functions trained with policy iteration using pseudo-reward given by the direction indicated by the eigenvectors of the SR matrix.



Policies trained using the same procedure.



Continual learning.



Conclusion

Sample efficiency

- Learn from multiple tasks
- Construct a hierarchy of abstract behaviour spanned over a temporal window

Subgoal discovery

- Construct a latent space where each state represents a future timeline/rollout
- Ground abstract behaviour using a basis of this space
- Decodable options
- Can learn options from the real behavior

References

[1] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. *arXiv preprint arXiv:1609.05140*, 2016.

[2] A. Barreto, R. Munos, T. Schaul, and D. Silver. Successor features for transfer in reinforcement learning. *CoRR*, abs/1606.05312, 2016. URL `http://arxiv.org/abs/1606.05312`.

[3] M. C. Machado, C. Rosenbaum, X. Guo, M. Liu, G. Tesauro, and M. Campbell. Eigenoption discovery through the deep successor representation. *CoRR*, abs/1710.11089, 2017. URL `http://arxiv.org/abs/1710.11089`.

[4] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman. The hippocampus as a predictive map. *bioRxiv*, 2016. doi: 10.1101/097170. URL `https://www.biorxiv.org/content/early/2016/12/28/097170`.

[5] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, 112(1-2):181–211, Aug. 1999. ISSN 0004-3702. doi: 10.1016/S0004-3702(99)00052-1. URL `http://dx.doi.org/10.1016/S0004-3702(99)00052-1`.