Goal-Conditioned Reinforcement Learning

Skill learners

Olivier Sigaud

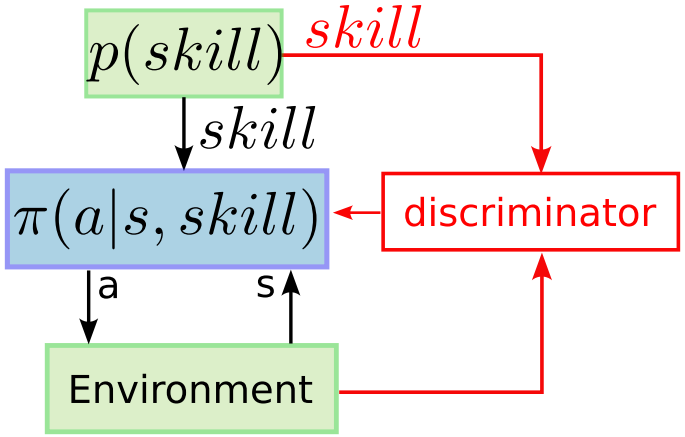
Sorbonne Université
http://people.isir.upmc.fr/sigaud

Skill discovery: generalities

- The difference between skill learners and goal reachers is in the evaluation
- In skill learners, the setter often samples skills from a fixed distribution
- But when image-based, an incremental latent skill encoder shifts the distribution
- Often, the setter emits a discriminative reward to expand skills, rather than a goal
- Generally maximizing a mutual information or empowerment criterion
- Generally followed by downstream tasks
- Secondary metrics: auxiliary task performance, zero-shot transfer perf., etc.
- The papers are often hard to compare as they do not use the same metrics

State-based skill discovery methods

VIC, DIAYN

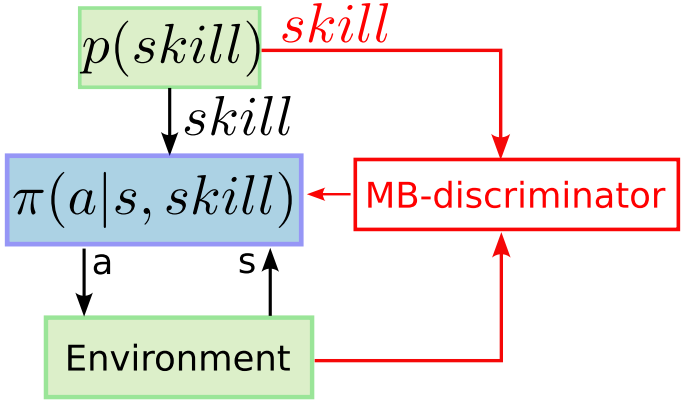


- Skills maximize their discriminability (over all states or final states)
- In canonical VIC, the setter is $p(skill)$, with p categorical (discrete set of skills)
- There is also VIC with $p(skill|s_0)$, with p continuous, and even p learned
- In Implicit VIC, $skill = s_f$ (final state), fulfilling skill $\rightarrow$ reaching goal $g = s_f$
- Limit: the number of skills N is specified in advance

Gregor, K., Rezende, D. J., and Wierstra, D. (2016) Variational intrinsic control. *arXiv preprint arXiv:1611.07507*

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2018) Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*

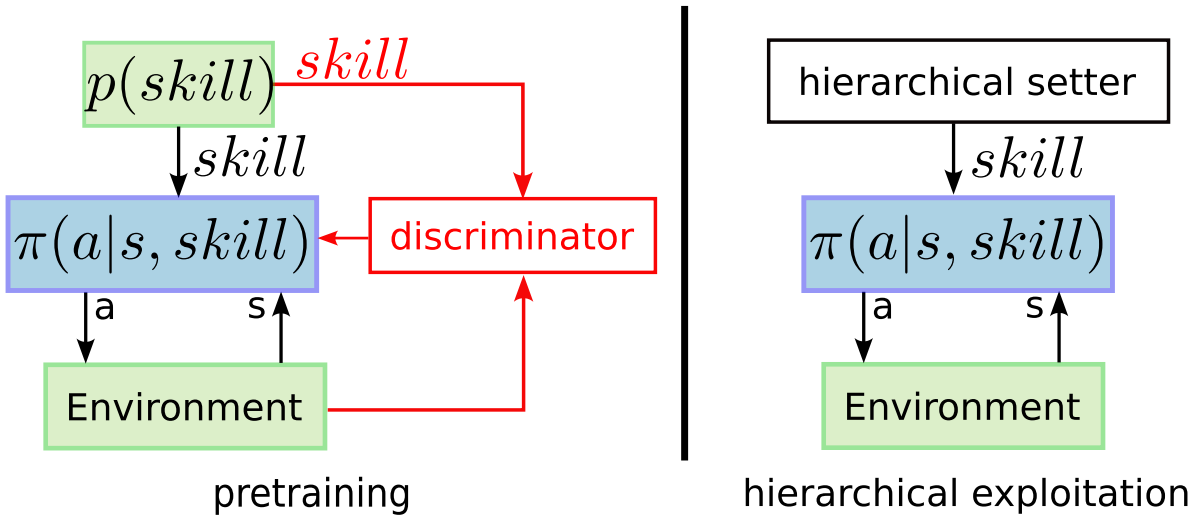
DADS



- Skills maximize their discriminability over states, but also the predictability of the next state given the skill
- It is easier to learn a dynamics model for each skill than a global dynamics model

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. (2019) Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657* (ICLR 2020)

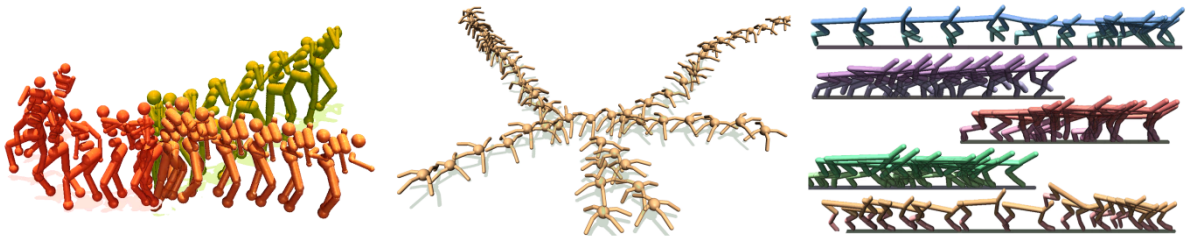
Hierarchical attempt: VALOR



- Shows that VIC and DIAYN can be framed into a common framework
- Shows that the learned skills are not satisfactory for complex hierarchical tasks

Achiam, J., Edwards, H., Amodei, D., and Abbeel, P. (2018) Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*

Use of skills



- Once trained, the skills can be exploited in several ways:
  1. To bootstrap a target task policy: try the N skills, train the one that performs best on the desired task
  2. To bootstrap imitation. Use the discriminator to find the fittest skill
  3. To bootstrap a hierarchical policy: find which skill to use in which context
- 1. does not work if the reward is sparse
- In these works, HER makes no sense (no behavioral/achieved goals)

Pretraining perspective of unsupervised RL: further work

Table 1: Comparing methods for pretraining RL in no reward setting. VISR (Hansen et al. 2020), APT (Liu & Abbeel 2021), MEPOL (Mutti et al. 2020), DIYAN (Eysenbach et al. 2019), DADS (Sharma et al. 2020), EDL (Campos et al. 2020). Exploration: the model can explore efficiently. Off-policy: the model is off-policy RL. Visual: the method works well in visual RL, e.g., Atari games. Task: the model conditions on latent task variables z. * means only in state-based RL.

Algorithm	Objective	Exploration	Visual	Task	Off-policy	Pre-Trained Model
APT	$\max \mathrm{H}(s)$	✓	✓	✗	✓	$\pi(a\|s), Q(s,a)$
VISR	$\max \mathrm{H}(z) - \mathrm{H}(z\|s)$	✗	✓	✓	✓	$\psi(s,z), \phi(s)$
MEPOL	$\max \mathrm{H}(s)$	✓*	✗	✗	✗	$\pi(a\|s)$
DIAYN	$\max -\mathrm{H}(z\|s) + \mathrm{H}(a\|z,s)$	✗	✗	✓	✗	$\pi(a\|s,z)$
EDL	$\max \mathrm{H}(s) - \mathrm{H}(s\|z)$	✓*	✗	✓	✓	$\pi(a\|s,z), q(s'\|s,z)$
DADS	$\max \mathrm{H}(s) - \mathrm{H}(s\|z)$	✓	✗	✓	✗	$\pi(a\|s,z), q(s'\|s,z)$
APS	$\max \mathrm{H}(s) - \mathrm{H}(s\|z)$	✓	✓	✓	✓	$\psi(s,z), \phi(s)$

$\psi(s)$: successor features, $\phi(s)$: state feature (*i.e.*, the representation of states).

- Not covered here: APT, MEPOL, VISR, EDL, APS...
- Image-based skill discovery is becoming prevalent with fundational models
- The pretraining perspective too

Liu, H. and Abbeel, P. (2021) APS: Active pretraining with successor features. In *International Conference on Machine Learning*, pages 6736–6747. PMLR

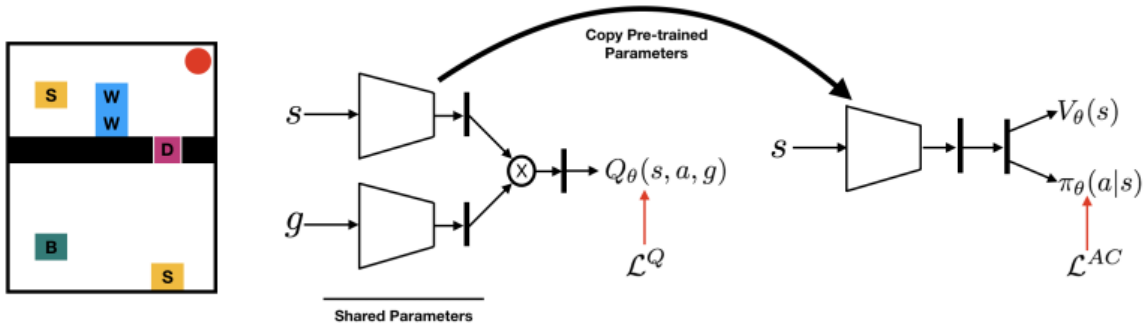
Selected algorithms

Algorithm	Reference	Downstream task	Type of Setter
VIC	[Gregor et al., 2016]	Standard RL (S)	Fixed
ASAP	[Mankowitz et al., 2016]	HRL (H)	Fixed
SNN4HRL	[Florensa et al., 2017]	HRL	Fixed
DIAYN	[Eysenbach et al., 2018]	S, Imitation, HRL	Fixed
VALOR	[Achiam et al., 2018]	HRL	Fixed
DADS	[Sharma et al., 2019]	HRL	Fixed
EDL	[Campos et al., 2020]	Not studied	Fixed
MISC	[Zhao et al., 2020]	Standard RL	MI-based
LSD	[Park et al., 2022]	Standard RL	MI-based
CIC	[Laskin et al., 2022]	Standard RL	Constrative encoder
GEAPS	[Wu and Chen, 2022]	GCRL	MI-based
HSD-3	[Gehring et al., 2021]	HRL	Hierarchical
CSD	[Park et al., 2023a]	Standard RL	MI-based
CESD	[Bai et al., 2024]	Standard RL	Ensembling-based

- The list is far from exhaustive
- See the references below

Image-based skill discovery methods

Many goals

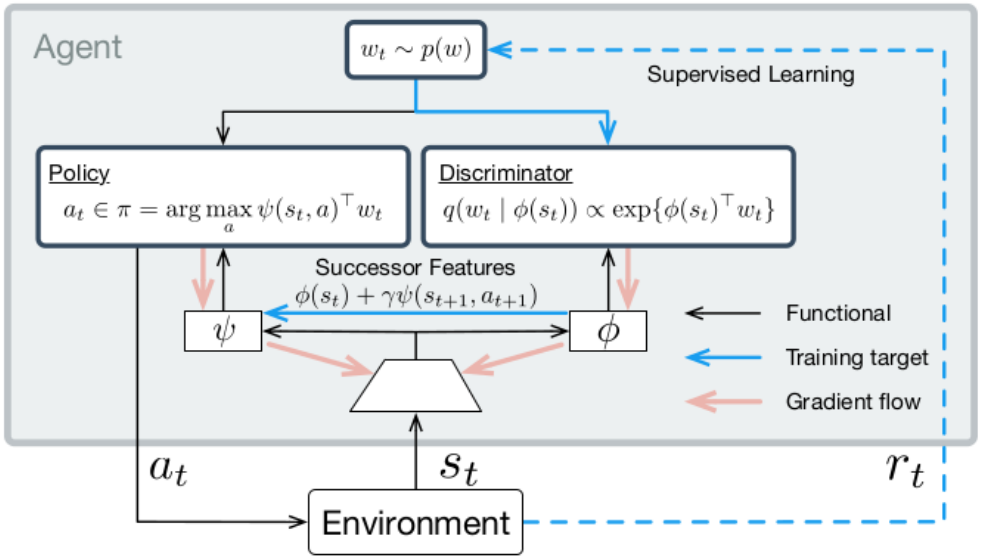


- An extension of DG-LEARNING [Kaelbling, 1993] using neural network approximators, learning from images and sampling many goals instead of all goals.
- Not truly a goal-conditioned policy
- The pre-trained goal-aware policy is reused in a goal-free context
- Works on symbolic images

Kaelbling, L. P. (1993) Learning to achieve goals. In *IJCAI*, pages 1094–1099

Veeriah, V., Oh, J., and Singh, S. (2018) Many-goals reinforcement learning. *arXiv preprint arXiv:1806.09605*

VISR

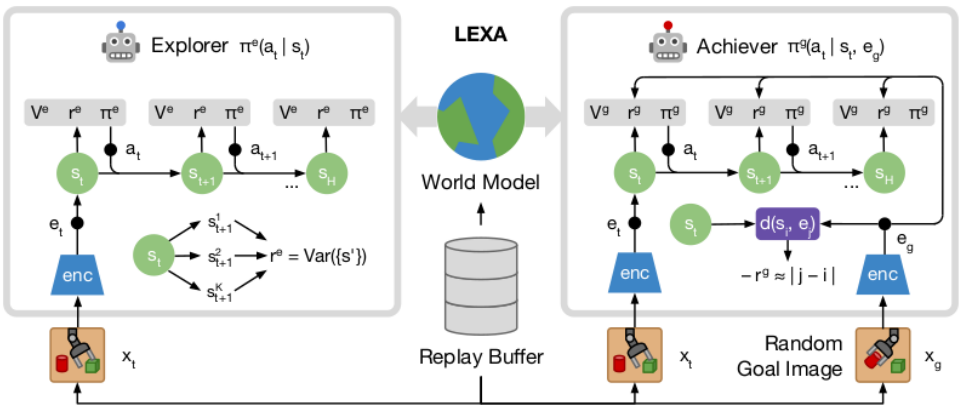


- An extension of VIC or DIAYN to the case of images, using successor features from [Barreto et al., 2017]
- Successor features provide efficient generalization to similar tasks
- Evaluated on ATARI games

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. (2017) Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30

Hansen, S., Dabney, W., Barreto, A., Van de Wiele, T., Warde-Farley, D., and Mnih, V. (2019) Fast task inference with variational intrinsic successor features. *arXiv preprint arXiv:1906.05030*

Latent Explorer Achiever (LEXA)



- Performs unsupervised pretraining and zero-shot evaluation
- Separates an explorer and an achiever as solver
- Leverages a world-model, making profit of the DREAMER family

Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. (2021) Discovering and achieving goals with world models.
In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*

Selected algorithms

Algorithm	Reference	Downstream task	Type of Setter	Image encoder
Many goals	[Veeriah et al., 2018]	Standard RL	uniform + LP-based	CNN
VISR	[Hansen et al., 2019]	Standard RL	MI-based	Successor features
APS	[Liu and Abbeel, 2021a]	Standard RL	MI-based	Successor features
APT	[Liu and Abbeel, 2021b]	Standard RL	MI-based	Constrative encoder
LEXA	[Mendonca et al., 2021]	Zero-shot transfer	Explorer + Achiever	RSSM (CNN)
METRA	[Park et al., 2023b]	Standard RL	Fixed	CNN
DODON'T	[Kim et al., 2024]	Standard RL	Tutor-based	CNN

- The list is far from exhaustive
- See the references below

Any question?

Send mail to: `Olivier.Sigaud@upmc.fr`

Achiam, J., Edwards, H., Amodei, D., and Abbeel, P. (2018).

Variational option discovery algorithms.
arXiv preprint arXiv:1807.10299.

Bai, C., Yang, R., Zhang, Q., Xu, K., Chen, Y., Xiao, T., and Li, X. (2024).

Constrained ensemble exploration for unsupervised skill discovery.
arXiv preprint arXiv:2405.16030.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. (2017).

Successor features for transfer in reinforcement learning.
Advances in neural information processing systems, 30.

Campos, V., Trott, A., Xiong, C., Socher, R., Giró-i Nieto, X., and Torres, J. (2020).

Explore, discover and learn: Unsupervised discovery of state-covering skills.
In *International Conference on Machine Learning*, pages 1317–1327. PMLR.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2018).

Diversity is all you need: Learning skills without a reward function.
arXiv preprint arXiv:1802.06070.

Florensa, C., Duan, Y., and Abbeel, P. (2017).

Stochastic neural networks for hierarchical reinforcement learning.
arXiv preprint arXiv:1704.03012.

Gehring, J., Synnaeve, G., Krause, A., and Usunier, N. (2021).

Hierarchical skills for efficient exploration.
Advances in Neural Information Processing Systems, 34:11553–11564.

Gregor, K., Rezende, D. J., and Wierstra, D. (2016).

Variational intrinsic control.
arXiv preprint arXiv:1611.07507.

Hansen, S., Dabney, W., Barreto, A., Van de Wiele, T., Warde-Farley, D., and Mnih, V. (2019).

Fast task inference with variational intrinsic successor features.
arXiv preprint arXiv:1906.05030.

Kaelbling, L. P. (1993).

Learning to achieve goals.
In *IJCAI*, pages 1094–1099. Citeseer.

Kim, H., Lee, B., Lee, H., Hwang, D., Kim, D., and Choo, J. (2024).

Do's and don'ts: Learning desirable skills with instruction videos.
arXiv preprint arXiv:2406.00324.

Laskin, M., Liu, H., Peng, X. B., Yarats, D., Rajeswaran, A., and Abbeel, P. (2022).

Unsupervised reinforcement learning with contrastive intrinsic control.
Advances in Neural Information Processing Systems, 35:34478–34491.

Liu, H. and Abbeel, P. (2021a).

APS: Active pretraining with successor features.
In *International Conference on Machine Learning*, pages 6736–6747. PMLR.

Liu, H. and Abbeel, P. (2021b).

Behavior from the void: Unsupervised active pre-training.

Mankowitz, D. J., Mann, T. A., and Mannor, S. (2016).

Adaptive skills adaptive partitions (asap).
Advances in neural information processing systems, 29.

Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. (2021).

Discovering and achieving goals with world models.
In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*.

Park, S., Choi, J., Kim, J., Lee, H., and Kim, G. (2022).

Lipschitz-constrained unsupervised skill discovery.
In *International Conference on Learning Representations*.

Park, S., Lee, K., Lee, Y., and Abbeel, P. (2023a).

Controllability-aware unsupervised skill discovery.
arXiv preprint arXiv:2302.05103.

Park, S., Rybkin, O., and Levine, S. (2023b).

Metra: Scalable unsupervised RL with metric-aware abstraction.
In *The Twelfth International Conference on Learning Representations*.

Sharma, A., Gu, S., Levine, S., Kumar, V., and Hausman, K. (2019).

Dynamics-aware unsupervised discovery of skills.
arXiv preprint arXiv:1907.01657.

Veeriah, V., Oh, J., and Singh, S. (2018).

Many-goals reinforcement learning.
arXiv preprint arXiv:1806.09605.

Wu, L. and Chen, K. (2022).

Goal exploration augmentation via pre-trained skills for sparse-reward long-horizon goal-conditioned reinforcement learning.
arXiv preprint arXiv:2210.16058.

Zhao, R., Gao, Y., Abbeel, P., Tresp, V., and Xu, W. (2020).

Mutual information-based state-control for intrinsically motivated reinforcement learning.
arXiv preprint arXiv:2002.01963.