Simple Sensor Intentions for Exploration

Tim Hertweck, Martin Riedmiller, Michael Bloesch, Jost Tobias Springenberg,
Noah Siegel, Markus Wulfmeier, Roland Hafner, Nicolas Heess
DeepMind
London, United Kingdom

***Abstract*—Modern reinforcement learning algorithms can learn solutions to increasingly difficult control problems while at the same time reduce the amount of prior knowledge needed for their application. A remaining challenge is the definition of reward schemes that facilitate exploration without biasing the solution or requiring expensive instrumentation. In this paper we focus on a setting in which goal tasks are defined as sparse rewards and exploration is facilitated via agent-internal auxiliary tasks. We introduce simple sensor intentions (SSIs) as a generic way to define auxiliary tasks that reduces the amount of prior knowledge required to define suitable rewards. Also, SSIs can be computed from raw sensor streams and thus do not require state estimation. We demonstrate that a learning system based on SSIs can solve complex robotic tasks: We show that a robotic arm can learn to grasp and lift arbitrary objects and solve a Ball-in-a-Cup task from scratch, even when only raw sensor streams are used for both controller input and in the auxiliary reward definition. A video showing the results can be found at https://deepmind.com/research/publications/Simple-Sensor-Intentions-for-Exploration.**

I. Introduction

A step stone on the path towards general AI is to minimize the amount of prior knowledge needed to set up learning systems. Ideally, we would like to identify principles that transfer to a variety of domains without task-specific adjustments. A remaining challenge is the definition of reward schemes that appropriately indicate task success, facilitate exploration without biasing the solution, and that can be implemented on robotics systems without expensive instrumentation. Sparse reward functions mitigate the bias on the final solution [11], however, an agent starting from scratch with a naive exploration strategy will most likely never encounter any learning signal. To this end Scheduled Auxiliary Control (SAC-X) [12] introduces the use of auxiliary rewards, that help exploring the environment. In the original work, the auxiliary tasks are defined with semantic understanding of the environment in mind – an important insight of [12] however is that the exact definition of auxiliary tasks can vary, as long as they jointly allow to collect rich enough data such that learning of the main task can proceed. In this work we make a step towards a more generic approach for defining auxiliary tasks, that reduces the need for task-specific semantic interpretation of sensors: A fundamental principle to enable exploration is to learn auxiliary behaviours that deliberately change sensor responses. We introduce a generic way to implement this concept and show that SSIs can aid exploration in a variety of robotic domains.

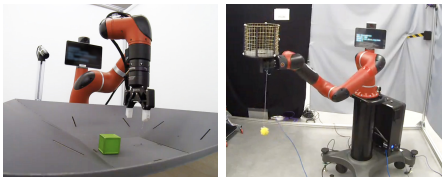

Fig. 1. Rethink Sawyer robotic arm with a Robotiq 2F-85 parallel gripper (left) and a custom made Ball-and-Cup attachment (right).

II. Simple sensor intentions

In the absence of an external reward signal, a sensible exploration strategy can be formed by learning policies that deliberately cause an effect on the observed sensor values. SSIs propose a generic way to implement this principle in the multi-task agent framework SAC-X [12]. SSIs are derived from raw sensor observations in two steps:

First step: We derive a scalar (virtual) sensor response by mapping an observation to a scalar value. For scalar observations this mapping can be the identity function, while other sensors might require pre-processing (e.g. camera images).

Second step: We define an SSI intention either by rewarding the agent for reaching a specific target sensor response, or for incurring a specific, directed change in the sensor response.

Both reward schemes do not require a semantic understanding of the environment, however a change in a sensor response is indicative for some change in the environment – by learning a policy that deliberately causes this change (a SSI) we obtain a natural way of encouraging diverse exploration.

For handling camera images, we propose to transform a pixel observation into a small amount of sensor responses by aggregating statistics of an image's spatial color distribution. As illustrated in Fig. 2, we threshold the image and calculate the mean location of the resulting binary mask along each of the image's axes, which we subsequently use as sensor values.

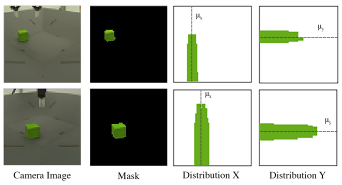

Fig. 2. Transformation used for deriving scalar responses from images.

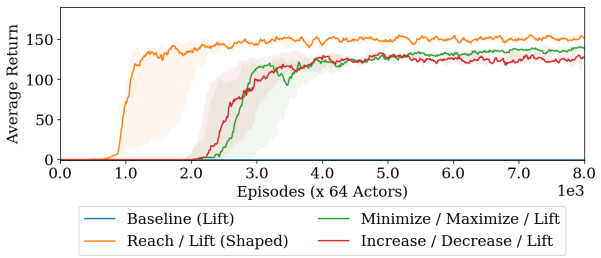

Fig. 3. 'Lift' learned from pixels in simulations.

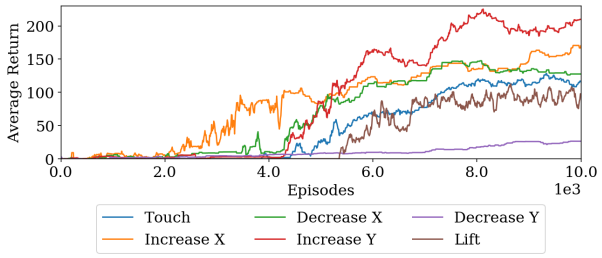

Fig. 4. 'Lift' learned from pixels on a real robot.

III. Learning SSIs for Active Exploration

In this work, we employ SAC-X to learn SSIs. The set of tasks is given by the reward functions for each of the SSIs that we want to learn, as well as the sparse, externally defined goal reward. The transition distribution, for which the policy and Q-function are learned, is obtained from a replay buffer that is filled by executing both the policy for the goal task as well as all other available exploration SSIs. Episodes are divided into multiple sequences, with a scheduler choosing a task-policy to execute for each sequence (see [12]).

IV. Experiments

We apply SSIs in the context of robotic experiments in simulation and on a real robot. We show, that by using SSIs, complex manipulation tasks can be solved: grasping and lifting objects, stacking two objects and solving a Ball-in-a-Cup task end-to-end from raw pixels. In all experiments we employ a Rethink Sawyer robotic arm as shown in Fig. 1. The goal reward is sparse (i.e. binary) in all setups. The agent's observations are proprioceptive sensors and raw camera images. The action space for manipulation is five dimensional and uses continuous cartesian velocities, while Ball-in-a-Cup uses four dimensional continuous joint velocities (see [13]).

A. Learning to grasp and lift

Using appropriate SSIs, the agent successfully learns to approach, grasp and eventually lift an object, which is very unlikely to happen, if only the sparse task reward is used (see Fig. 3). In addition, we highlight the following ablations:

1) Learning success does not depend on a particular camera pose: We find that while varying the camera pose has an influence on learning speed, successful learning is possible for a wide variety of camera positions.

2) The SSI color channel does not necessarily need to specify a single object: Even if multiple objects with the same color are present in the scene or if (small) parts of the background have the same color as the object, lifting can be learned successfully.

3) One can use a much more general selection of the color channel: Rewards for different color channels can be totalled to form 'aggregate SSIs'. Using this method, an agent can learn to lift arbitrary colored objects placed in the workspace.

4) The SSI method is not restricted to pixels, but works with a general set of (robot) sensors: SSIs can additionally be applied to basic sensors like the touch sensor, the joint angles or the joint velocities. In conjunction with SAC-Q we can show that the 'Lift' task can be learned in a more general setup with 22 auxiliary SSIs.

B. Learning to stack

Learning to stack poses additional challenges: The scene is more complex since there are two objects, reward is given only if one object is placed above the target object, and the target object can move. Without SSIs the agent is not able to learn the task but with SSIs, learning is possible from raw sensor information and an external task reward only.

C. Ball-in-a-Cup

As an example of the generality of SSIs, we employ the same set of SSIs used before, to learn the dynamic Ball-in-a-Cup task [13]. Dynamic tasks in general exhibit additional difficulties (e.g. timing or reaching and staying in possibly unstable regimes of the configuration-space) and, as a result, learning to catch the ball purely from pixels is out-of-reach for an agent, that only employs the sparse catch reward. With SSIs however, the agent can successfully learn the task.

V. Related Work

Transfer from additional tasks has a long-standing history in reinforcement learning to accelerate exploration and learning [15, 10]. Auxiliary tasks have been investigated as manually chosen to help in specific domains [12, 4, 8, 9, 3] and as based on agent behaviour [1]. In comparison to methods using auxiliary tasks mostly for representation shaping by sharing a subset of network parameters across tasks [8, 9], SSIs share data between tasks which directly uses additional tasks for exploration. Recent work on diversity has demonstrated the importance of the space used for skill discovery [14]. SSIs provide a perspective on determining valuable task spaces with limited human effort. Finding automated curricula to accelerate learning gains relevance in this context [2, 7, 5, 6]. In this work, we rely on task scheduling similar to Riedmiller et al. [12] in order to optimize the use of training time.

VI. Conclusion

Learning to deliberately change sensor responses is a promising exploration principle in settings, where it is difficult or impossible to experience an external task reward purely by chance. We introduce the concept of simple sensor intentions (SSIs) that implements this principle in a generic way within the SAC-X framework. Our approach requires less prior knowledge than the broadly used shaping reward formulation, that relies on task insight for the definition and state estimation for the computation of rewards. In case studies, we demonstrate the application of SSIs to various robotic tasks showing, that SSIs are general – that no or only minor adaptations between tasks are required – and that yet, SSIs provide for meaningful exploration in various domains.

An extended version of the paper can be found at https://arxiv.org/abs/2005.07541.

References

[1] Marcin Andrychowicz, Dwight Crow, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5055–5065, 2017.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.

[3] Serkan Cabi, Sergio Gomez Colmenarejo, Matthew W. Hoffman, Misha Denil, Ziyu Wang, and Nando de Freitas. The Intentional Unintentional Agent: Learning to solve many continuous control tasks simultaneously. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, Proceedings*, pages 207–216, 2017.

[4] Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. In *International Conference on Learning Representations (ICLR)*, 2017.

[5] Sébastien Forestier, Yoan Mollard, and Pierre-Yves Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *arXiv preprint arXiv:1708.02190*, 2017.

[6] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1311–1320. JMLR. org, 2017.

[7] Nicolas Heess, Dhruva Tirumala, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, Ali Eslami, Martin Riedmiller, et al. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*, 2017.

[8] Max Jaderberg, V Mnih, W M Czarnecki, T Schaul, J Z Leibo, D Silver, and K Kavukcuoglu. Unreal: Reinforcement learning with unsupervised auxiliary tasks. In *ICLR 2017*, 2017.

[9] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J. Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, and Raia Hadsell. Learning to Navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

[10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[11] Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *ICML*, volume 98, pages 463–471. Citeseer, 1998.

[12] Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International Conference on Machine Learning*, pages 4341–4350, 2018.

[13] Devin Schwab, Jost Tobias Springenberg, Murilo Fernandes Martins, Michael Neunert, Thomas Lampe, Abbas Abdolmaleki, Tim Hertweck, Roland Hafner, Francesco Nori, and Martin A. Riedmiller. Simultaneously learning vision and feature-based control policies for real-world ball-in-a-cup. In Antonio Bicchi, Hadas Kress-Gazit, and Seth Hutchinson, editors, *Robotics: Science and Systems XV, University of Freiburg, Freiburg im Breisgau, Germany, June 22-26, 2019*, 2019. doi: 10.15607/RSS.2019.XV.027. URL https://doi.org/10.15607/RSS.2019.XV.027.

[14] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

[15] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.