Self-Development Framework for Reinforcement Learning Agents

Andrea Bonarini  
Department of Electronics and Information, Politecnico di Milano  
piazza Leonardo Da Vinci, 32  
20133 Milan, Italy  
bonarini@elet.polimi.it

Alessandro Lazaric  
Department of Electronics and Information, Politecnico di Milano  
piazza Leonardo Da Vinci, 32  
20133 Milan, Italy  
lazaric@elet.polimi.it

Marcello Restelli  
Department of Electronics and Information, Politecnico di Milano  
piazza Leonardo Da Vinci, 32  
20133 Milan, Italy  
restelli@elet.polimi.it

Patrick Vitali  
Department of Electronics and Information, Politecnico di Milano  
piazza Leonardo Da Vinci, 32  
20133 Milan, Italy  
patvi@tin.it

Abstract—We present SMILe (Self-Motivated Incremental Learning), a new learning framework where an agent learns a set of abilities needed to face several tasks in its environment, by following a biologically inspired, self–motivated approach that loops over three main phases. In the babbling phase, the agent randomly explores the environment, in a way similar to what animal puppies do. This provides information about the effects of action on the environment. In the motivating phase, the agent identifies what is interesting in the environment and develops an intrinsic motivation in achieving situations with highest interest. In the skill acquisition phase, the agent learns the skills needed to reach the most interesting state, guided by a self-generated reinforcement function. Once a new skill is available the babbling phase can start again with the enlarged set of abilities, and learning continues all the life long. We present results on a gridworld abstraction of a robotic environment to show how SMILe makes it possible to learn skills that enable the agent to perform well and robustly in many different tasks.

Index Terms—Reinforcement Learning, Self-Motivation, Self-Development, Intrinsically Motivated Reinforcement Learning.

I. INTRODUCTION

In this paper, we present SMILe (Self-Motivated Incremental Learning), a learning framework where an agent incrementally learns general abilities through direct interaction with the environment, only guided by self-generated interest. This biologically inspired approach integrates ideas coming cognitive sciences and intrinsically motivated reinforcement learning.

In recent years, studies on the inner mechanisms of human development, pursued in many different areas (such as robotics, machine learning, neuroscience, psychology, developmental sciences) converged to a new field, commonly referred to as developmental robotics [?], [?], whose goal is to enable robots to solve complex tasks in dynamical environments. Traditionally, a designer must specifically program the set of skills for a robot, needed to accomplish a given task. Often these skills are tuned to perform a predefined task on a specific environment, and can be hardly reused if the task or the environment changes. On the other hand, developmental robotics focuses on the study of frameworks in which the agent does not directly address any specific problem, but develops a set of basic skills up to very general abilities that can be used to solve many different tasks.

Because of the complexity of its goal, developmental robotics has many different facets [?]; in this paper, we focus on a subset of them. In particular, we refer to developmental robotics as an incremental process where the agent organizes its initial skills through spontaneous exploratory phases and self-motivated learning. Therefore, we focus only on high-level behavior and we assume that the agent is provided with abstractions on actions and states. Issues related to the use of real sensors and actuators will be addressed in future work.

In particular, self-motivated learning proved to be one of the most challenging parts of the development process and many frameworks have been proposed to deal with this problem [?], [?], [?]. One of the most promising approaches is intrinsically motivated reinforcement learning [?] whose goal is to autonomously develop a hierarchy of skills through a process guided by an intrinsic motivation, without any commitment to achieve a specific task. The learned skills are general enough to be used in many different tasks.

The framework we introduce in this paper extends the intrinsically motivated reinforcement learning model to a more general, biologically inspired, development process, in which the notion of interest is not hardwired in the algorithm, but autonomously extracted from characteristics of the environment. The proposed development process iterates through three different phases: babbling, motivating and skill acquisition. The idea is that, through a playful exploration of the environment (babbling phase), the agent identifies interesting situations and generates an intrinsic motivation (motivating phase) aimed at learning to get into these situations (skill acquisition phase). This process incrementally increases the skills of the agent, so that new interesting configurations can be experienced. The proposed decomposition of the development process, although not so sharp in the biological
beings, has been considered suitable for artificial systems, as confirmed by experimental results.

Beyond the design of an iterative learning process that incrementally acquires more and more complex skills, in this paper we introduce the definition of a general methodology for the autonomous evaluation of the interest of a state in relation to the dynamics of the environment and the agent capabilities, and the use of an indifference criterion that prevents the agent from learning many useless tasks when no interest state can be actually identified.

The rest of the paper is organized as follows. In the next section we briefly review works in psychology and neuroscience that inspired our proposal. Section ?? we describes SMILe, a new framework for intrinsically motivated learning. Section ?? provides some preliminary experimental results on a gridworld that simulates a simple robotic environment. Finally, we discuss related works in Section ??, and conclude our work by outlining future work in Section ??.

II. BIOLOGICAL BACKGROUND

As stated in [?], one of the most promising directions to achieve the ambitious goal of autonomy in robotics, is the definition of a suitable lifelong development process. This consists of an open-ended learning process in which an agent pursues self-motivated goals and develops highly re-usable skills. According with many different research results in neuroscience, psychology and philosophy, a similar mechanism could be traced in the developmental process of many biological entities, which represents the main source of inspiration for the studies in developmental robotics.

Many approaches in developmental robotics refer to the studies by Piaget [?], and, in particular, to his research on children’s very early phase of development. Piaget showed that childish development can be considered as an incremental process of acquisition of new abilities in which children modulate the complexity of their activities in association with the increasing complexity of their cognitive and morphological structures. Starting from this evidence, Barto et al. [?] propose a framework in which the agent is able to develop a hierarchical collection of skills given a set of basic abilities. According to the common perspective shared by many works in developmental robotics [?], a key factor of the self-development process is the existence of an intrinsic motivation that provides an internal reward, independent from any externally imposed goal, that guides the behavior of the agent to the development of new, useful skills.

The introduction of intrinsic motivation in the learning process requires the modification of the actor-critic reinforcement learning model. While in the actor-critic model the critic, that is part of the environment, provides the agent with rewards and punishments according to its behavior, in the intrinsic motivated model the reinforcement signal is generated by the agent itself, as suggested by many neuroscience studies [?]. Therefore, the learning process does not aim at the achievement of a goal defined by an external designer, but at the identification of a strategy to accomplish a task autonomously generated by the agent. This model has been adopted in many frameworks, such as those by Schmidhuber [?], [?] and Barto et al. [?], [?].

Another important contribution to the comprehension of the mechanisms attending human development comes from the research carried out by Berlyne [?] about the notion of curiosity and its influence on behavior and the rising of intrinsic motivation. Berlyne asserts that, in absence of a particular aim, human behavior is partly determined by an innate will of exploring what is perceived as interesting. Psychologists define curiosity as a form of motivation that promotes exploratory behavior to learn more about a source of uncertainty. Schmidhuber [?] relates interest to the current knowledge of the observer and proposes quantitative definitions of novelty and surprise.

In summary, life-long learning in biological systems seems to be characterized by a progressive, self-motivated development that leads to the incremental acquisition of more and more complex skills.

III. SMILE

Inspired by the model just summarized, SMILe adopts a life-long self-motivated developmental learning process that aims at acquiring a set of skills useful to face different tasks.

As already proved in many studies as [?], [?], [?] Reinforcement Learning (RL) is one of the most suitable frameworks to deal with learning problems in developmental robotics. Furthermore, the incremental development of simple skills into complex activities can be efficiently described using Hierarchical Reinforcement Learning (HRL) [?], as suggested in [?]. HRL problems are generally formalized as a Semi-Markov Decision Process (SMDP). In particular, in the option framework [?], an SMDP is defined by tuple $\langle S, O, P, R \rangle$, where $S$ is the set of states (i.e. perceptions), $O$ is the set of options (i.e. skills available to the agent), $P(s, o, s')$ is the transition model, that is the probability to get to state $s'$ taking the option $o$ is state $s$, and $R(s)$ is the reward function that returns how good is state $s$. Unlike traditional RL approaches, in intrinsically motivated learning the reward is the result of a self-motivation to the achievement of autonomously identified goals.
Formally, a skill is represented as an option \( o \), i.e. as a tuple \( (\pi_o, I, \beta) \), where \( \pi_o : S \times O \rightarrow [0, 1] \) is the control policy that describes the probability to execute an option when the agent is in a specific state, \( I \subset S \) is the set of states where the option is defined and \( \beta(s) \) is the probability for an option to finish at state \( s \). When the development process starts, the agent has an initial set of basic options \( O^0 \), and, at each iteration \( k \), the set of options is incrementally modified adding the option learned in the skill acquisition phase: \( O^k = O^{k-1} \cup \{o^k\} \).

Each skill is learned by a self-motivated process that continually iterates on three main phases (see Fig. ??):

- **Babbling**: the agent playfully interacts with the environment to get aware of the relationships between its actions and the environment dynamics.
- **Motivating**: the agent evaluates which is the most interesting situation it has experienced during the exploration performed during the babbling phase.
- **Skill acquisition**: the agent tries to learn the skill that allows it to reach the interesting situation.

A. Babbling Phase

One of the crucial activities in the development process of puppies and babies is self-exploration [?]. Through self-explorative acts, they become aware of their own capabilities with respect to the surrounding environment, understand the consequences of the actions they have autonomously selected, and learn to control and exploit the dynamics of their bodies. In analogy to vocal babbling, this experiential process has been called body babbling [?].

Moving from these observations, we introduced in SMILe, at the beginning of each iteration, a babbling phase. The acquisition of a new skill starts with a self-explorative phase in which the agent, for a certain time, randomly executes one skill, choosing from the set of admissible skills \( O^k \). The choice of taking actions according to a uniform probability distribution over the action space is consistent with the fact that at this time the agent has neither an extrinsic nor an intrinsic motivation. The goal of the babbling phase is to collect information about the environment dynamics that can be used in the next phase to determine whether there is any interesting skill that is worth to learn. In particular, at each iteration \( k \), the agent builds an estimate (even partial) of the state transition model \( \hat{P}_{\pi_k}(s, s') \) (computed from the transition model \( P(s, o, s') \)), that is the probability of moving from \( s \) to \( s' \) according to the random policy \( \pi_k \). Since the state transition probabilities do not depend only on intrinsic characteristics of the environment, but also on the abilities of the agent, everytime a new skill is learned, the capabilities of the agent to control the environment dynamics change and the state transition probabilities must be recomputed.

In the development process the agent learns skills with increasing complexity, that require more time steps to finish with respect to simple actions. Therefore, at each iteration the number of steps in the babbling phase must be increased depending on the skills currently available to the agent.

B. Motivating Phase

As discussed in Section ??, there is a huge body of evidence about the central role played by intrinsic motivation in each development process. Intrinsic motivation leads organisms to visit regions of the environment with particular characteristics without any extrinsic motivation. In this way, they may increase their competence to control the environment, by acquiring a broad set of skills that can be reused for different goals. In particular, the studies of Piaget [?] and Berlyne [?] suggest that intrinsic motivation may be generated by several factors: surprise, incongruity, uncertainty, and novelty. All these factors act together to determine an intrinsic interest associated to different situations. A similar process is implemented in SMILe motivating phase, where an interest value is associated to each state visited during the babbling phase on the basis of the information contained in the estimated state transition probabilities \( \hat{P}_{\pi_k}(s, s') \).

Since interest may arise from different sources, here we propose two measures of the interest of a state:

- **Unbalance**: 
  \[
  \rho_U(s) = (1 - p_{in}(s)) - p_{in}(s)(1 - p_{out}(s)),
  \]
  where \( p_{in}(s) = \frac{1}{|S|} \sum_{s' \in S} P_{\pi_k}(s', s) \) and \( p_{out}(s) = \sum_{s' \notin S} P_{\pi_k}(s, s') \). The first term of Equation ?? is the probability of not moving into state \( s \) in one step following \( \pi \), given that the agent starts from a random state. To this term, we subtract a second term that represents the probability to reach \( s \) in one step starting from a random location and then to remain in \( s \) for another step. The intuition behind Equation ?? is that states that, under a random policy, are difficult to be reached or that, once reached, can be easily left, are relevant as subgoals for many complex tasks whose solution needs the agent to pass through states that cannot be easily reached without a specific skill. Using this definition \( \rho \) ranges in \([-1, 1]\).

- **Differential entropy**: 
  \[
  \rho_E(s) = H_{out}(s) - H_{in}(s),
  \]
  \[
  = - \frac{\sum_{s' \in S} \hat{P}_{\pi_k}(s, s') \log |S| \hat{P}_{\pi_k}(s, s')}{\pi_k(s)} + \frac{\sum_{s' \notin S} \hat{P}_{\pi_k}(s', s) \log |S| \hat{P}_{\pi_k}(s', s)}{|\pi_k(s)|}.
  \]

The differential entropy gives a measure of how much a state is connected with the other ones. The first term (\( H_{out}(s) \)), that is the entropy of probability distribution of the output transitions, assumes high values when, taking a random action in \( s \), there is a near-uniform probability to reach any other state. On the other hand, the second term (\( H_{in}(s) \)) takes
into consideration the number of states that have a non-null probability of arriving in $s$. The differential entropy measure evaluates as interesting those states that are reachable from a few states and that allow to reach several others.

These measures define the concept of interest of a state on the basis of information about its input and output transition probabilities, without taking into account the characteristics of the surrounding states; for this reason we call them local interest functions. Using the estimated state transition probabilities and a local interest function, we define the global interest function with the following Bellman-like equation:

$$I^k(s) = \rho^k(s) + \gamma \sum_{s' \in S} \hat{P}^k(s, s')I^k(s').$$

(3)

In this way, the interest of a state depends, not only on the characteristics of its local transitions, but also on the interest of the states that may be reached from it. The discount factor $\gamma \in [0,1)$ determines how much distant states should influence the interest of the current state. To compute $I(s)$ we can use an iterative policy evaluation algorithm [7]. The formulation of the interest function $I(s)$ is such that it can represent a large set of the aspects of the concept of interest [7] depending on the specific definition of local interest $\rho(s)$ that is used.

Once $I^k(s)$ has been computed, the agent self-determines its next goal by choosing the most interesting state $\pi^k = \arg \max_s I^k(s)$, and produces an intrinsically motivated reward function that simply returns a positive reward when the agent achieves state $\pi^k$ and null otherwise. Learning to reach this state is the goal of the third phase.

It is possible to show that, using the definition of local interest previously introduced, the acquisition of new skills decreases the interest in goal states (boredom effect), thus preventing the agent from choosing them again. After some iterations of the development process, the interest function should tend to flatten until no state with relevant interest can be actually identified in the motivating phase. To prevent the extraction of new useless goals, SMILE does not acquire a new skill and steps back to the babbling phase whenever the interest function becomes too flatten, thus indicating the absence of interesting goals. More details about this stopping condition can be found in [7].

C. Skill Acquisition Phase

During a development process, an organism starts with simple skills and acquires more and more complex abilities. Each time a new skill is learned, it may be used to simplify the following learning processes, thus progressively increasing the complexity of tasks that can be successfully solved.

Recently, the idea of hierarchically decomposing complex problems into simpler sub-problems has been successfully exploited also in RL with the introduction of formalisms for managing temporally extended actions [7]. Several of these approaches work with fixed hand-coded decompositions, even if some proposals have been advanced to dynamically decompose a given goal into simpler sub-goals [7, 8].

In SMILE, once the motivating phase has identified the goal state $\pi^k$ and generated the intrinsically motivated reward function $R^k(s)$, the agent starts the skill acquisition phase in which it learns the policy of a new option $o^k$ whose goal is $\pi^k$. The policy of the new option is learned according to the option learning algorithm described in [8]. At each time step, the action value function $Q(s,o)$, that is the estimation of the amount of reward the agent can obtain by taking option $o$ in state $s$, is updated according to the following update rule:

$$Q(s,o) \leftarrow (1-\alpha)Q(s,o) + \alpha \left[ \tilde{r} + \gamma^i \max_{o' \in O^{k-1}} Q(s',o') \right]$$

(4)

where $\alpha$ is a learning step size, $i$ is the number of steps taken by option $o$ to meet its termination condition, and $\tilde{r}$ is the discounted reward accumulated from $s$ to $s'$ in $i$ steps according to reward function $R(s)$. In learning a skill, the agent, in addition to basic actions, may benefit also from the previously acquired skills, thus forming a hierarchy of skills.

Once the skill acquisition is finished, the new option $o^k$ is created and added to the set of options $O^{k-1}$. This new option is characterized by a deterministic policy that can be directly derived from the action value function $Q(s,o)$ by choosing in each state $s$ the option $o$ that maximizes its value. The termination condition $\beta(s)$ is set to 1 for $s = \pi^k$ and to 0 elsewhere. For what concerns the initial set, it can be limited to a subset of the state space $S$ composed by the states that have been most visited in the skill acquisition phase.

After the acquisition of a new skill, the development process starts a new iteration activating a new babbling phase, in order to experience how the new skill modifies the agent interaction with the environment. This leads to the definition of a new interest function that defines a new learning goal, thus obtaining an incremental process in which the agent increases its capabilities of controlling the environment.

IV. EXPERIMENTAL ACTIVITY

The experiment we discuss is a version of the Playworld proposed in [7]. The Playworld is an abstraction of a real environment characterized by two rooms with a door in between, two panels and a charger (Fig. ??). The panels are in the room at left: the light panel switches the light on and
incremental process described in Section ?? In the second stage, five different goals are imposed by an external designer providing an extrinsic reward function.

In the first stage, the relevant events we can expect the robot to find are: light on, light off, open door, close door, charge. The upper graph of Fig. ?? shows the events occurred in the babbling phase at first iteration. In this phase the robot takes actions among its basic skills at random and it succeeds only in switching the light on and off a few times. Immediately after the babbling phase, the robot computes the interest function using the local interest ρ(·) (due to lack of space, we present results obtained only with the unbalance measure) and, according to its model estimation, learns a new skill. The lower graph of Fig. ?? shows the changes introduced by the skills learned after five iterations. As it can be noticed, the skills developed in the previous iterations bias the random exploration so that the robot succeeds in activating more complex events (e.g., open the door and charge). This suggests that SMILe could be used to make a robot to autonomously discover interesting configurations in the environment and to develop self-motivation in learning new skills.

In the second stage, the skills developed in the previous stage are tested to evaluate whether they can be reused in different tasks. In particular, we compare the performance of a robot that exploits the new skills to that of a robot that uses Q-Learning [?], on five different tasks:

- **Task1:** charge
- **Task2:** charge, move to upper left corner of right room
- **Task3:** charge, move to upper left corner of left room
- **Task4:** charge, move to left room and close the door
- **Task5:** charge, move to left room, close the door, switch the light off

While **Task2** and **Task3** are not strictly related to any salient event, the other tasks require that the robot achieves configurations relevant for the Playworld environment.

In the comparison, we adopted the same learning parameters for both Q-Learning and SMILe (learning rate α = 0.6, ε-greedy exploration with ε = 0.2, discount factor γ = 0.95). Each 1000 learning episodes, the extrinsic reward function is changed according to the task that must be accomplished and the learning robot should be able to adapt its policy to the new task without restarting the learning from scratch.

**Fig. ??** shows the number of steps per learning episode. The first 2100 episodes, labeled as **Self-Development** in the graph, represent the first stage of the experiment in which the SMILe robot autonomously identifies six different goals for which one new skill is learned at each iteration. On the other hand, in the first stage the Q-Learning robot does nothing, since no extrinsic reward is provided. The second stage starts with the introduction of a positive extrinsic reward for achieving the charger. While Q-Learning robot can only use the basic skills, SMILe robot exploits the skills learned.
in the first stage and succeeds in finding the optimal policy to reach the charger in less episodes than those needed by Q-Learning. Similarly, SMILe succeeds in exploiting its skills even for changing tasks, while Q-Learning took more time to adapt to new extrinsic reward functions.

Furthermore, in Fig. ?? we compare the total number of steps for both the algorithms and we report their difference. In the first stage, SMILe takes almost 250000 steps to explore the environment and to learn the new skills, while no steps are taken by the Q-Learning robot. Notwithstanding the initial loss, the total number of steps needed by SMILe after the accomplishment of Task1 is less than that of Q-Learning. The advantage of SMILe becomes even more relevant at the end of the second stage when Q-Learning took almost twice as much of steps than SMILe. This comparison shows that SMILe, even though it requires potentially expensive exploration of the environment, leads to the development of useful skills that can be profitably reused in many different tasks. In particular, the number of steps saved during the extrinsically motivated learning stage is greater than those used in the first stage already in the first goal.

Further experiments on other problems are reported in [?].

V. RELATED WORKS

The learning framework proposed in SMILe is related to many works in the field of RL. The development process proposed in SMILe is closely related to the IML framework described in [?]. While the automatic development of a hierarchy of skills is similar, the solution proposed in [?] does not address the problem of the definition of a criterion for the extraction of salient events, that are hardwired in the agent for each problem. On the other hand, in SMILe we provide a general definition of interest function that can evaluate many different characteristics in the environment.

Many techniques for task decomposition using automatic discovery of subgoals have been proposed [?], [?], [?]. Since the subgoal discovery process is strictly related to the specific task (except for [?] where a task-independent criterion), the skills learned for the achievement of intermediate goals are not likely to be reused for other purposes.

Several approaches [?], [?] identified the error in prediction of environment dynamics, or the decrease of this error [?], as the main driver for self-motivation. Even if in SMILe the interest function is based on an estimate of the transition model as well, the proposed definition is more general, since it allows to catch many different features of the environment depending on the chosen local interest function. Furthermore, the use of iterative policy evaluation makes the agent able to propagate local interest to surrounding regions, thus identifying global interesting configurations.

Many works in developmental robotics [?], [?], [?] proposed the exploitation of RL techniques in the learning process. While they deal with robotic applications and enable robots with basic learning capabilities, SMILe is focused on the higher-level learning issues related to the development of complex hierarchies of reusable skills.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented SMILe, a new self-development reinforcement learning framework that incrementally acquires more and more complex skills through an iterative three phases learning process. Experimental results show the effectiveness of the skills learned by SMILe when operating in environments where different tasks may arise, thus providing the agent with a good degree of autonomy.

Currently, we are investigating the use of function approximation techniques to scale to large, high dimensional domains. Future work includes the integration of SMILe with developmental robotics approaches in real robotic tasks.