The Paradox of Choice: On the Role of Attention in Hierarchical Reinforcement Learning

Andrei Nica[*] [1,2,3] **Khimya Khetarpal**[*] [1,2]
Doina Precup[1,2,4]
[1]McGill University, [2]Mila, [3]Polytechnic University of Bucharest, [4]Google Deepmind
khimya.khetarpal@mail.mcgill.ca, nicaandr@mila.quebec, dprecup@cs.mcgill.ca

Abstract

Decision-making AI agents are often faced with two important challenges: the depth of the planning horizon, and the branching factor due to having many choices. Hierarchical reinforcement learning methods aim to solve the first problem, by providing shortcuts that skip over multiple time steps. To cope with the breadth, it is desirable to restrict the agent's attention at each step to a reasonable number of possible choices. The concept of affordances (Gibson, 1977) suggests that only certain actions are feasible in certain states. In this work, we first characterize "affordances" as a "hard" attention mechanism that strictly limits the available choices of temporally extended options. We then investigate the role of *hard* versus *soft* attention in training data collection, abstract value learning in long-horizon tasks, and handling a growing number of choices. To this end, we present an online, model-free algorithm to learn affordances that can be used to further learn subgoal options. Finally, we identify and empirically demonstrate the settings in which the "paradox of choice" arises, i.e. when having fewer but more meaningful choices improves the learning speed and performance of a reinforcement learning agent.

1 Introduction

Decision making in complex environments can be challenging due to having many choices to consider at every time step. Learning an attention mechanism that limits these choices could potentially result in much better performance. Humans have a remarkable ability to selectively pay attention to certain parts of the visual input (Judd et al., 2009; Borji et al., 2012), gathering relevant information, and sequentially combining their observations to build representations across different timescales (Hayhoe and Ballard, 2005; Zhang et al., 2019), which plays an important role in guiding further perception and action (Nobre and Stokes, 2019; Badman et al., 2020). In this paper, we study the role of attention over action choices in reinforcement learning (RL) agents.

An RL agent interacts with an environment through a sequence of actions, learning to maximize its long-term expected return (Sutton and Barto, 2018). Temporal abstractions, e.g. options (Sutton et al., 1999) allow it to consider decisions at variable time scales. Options allow the agent to reduce the depth of its lookahead when performing planning, and to propagate credit over a longer period of time. However, in large state-action spaces, the agent is still faced with many choices at every decision step, and options can in fact worsen this problem, as choices multiply when considering different timescales. Hence, temporal abstraction can lead to a larger branching factor.

A well-known solution is to *select* a small number of choices to focus on, for instance, applying an action only when certain preconditions are met in classical planning (Fikes et al., 1972), or using initiation sets for options (Sutton et al., 1999). Recent work (Khetarpal et al., 2020b) proposed a

[*] equal contribution

NeurIPS 2022 All Things Attention: Bridging Different Perspectives on Attention Workshop.

generalization of initiation sets to *interest functions* (Sutton et al., 2016; White, 2017), which provide a way to learn options that specialize to certain regions of the state space. All these approaches can be viewed as providing an attention mechanism over action choices.

While attention has been widely studied in the field of computer vision (Hou and Zhang, 2007; Borji et al., 2012), the role of different types of attention over action choices has not been studied much, especially for temporally extended actions. In this work, we explore attention over action choices in RL agents with a special interest in affordances associated with options, which can be viewed as a form of *hard* attention. On the other hand, having soft preferences over *all* actions without eliminating any can be viewed as *soft* attention.

In this paper, we study the role of *hard* versus *soft* attention in decision making with temporal abstractions. We posit that in certain settings, restricting an agent's attention through affordances leads to fewer but more useful choices compared to using soft attention in the form of interest functions. We demonstrate empirically this *paradox of choice*: fewer choices can contribute to faster learning, resulting in more rewards. As expected, this effect is more pronounced when the agent attends to choices that lead to better long-term utility. Attending to useful choices in a given state often has a compound effect, leading to further good choices in the near future.

Our **main contributions** are:

1. We characterize affordances as *hard*-attention over temporally-extended choices in RL and investigate the role of attention in HRL. Prior frameworks for choosing temporally extended actions have been fairly naive in the RL literature. To the best of our knowledge the refined view of choice attention we propose is novel.
2. Concretely, we inspect the role of *hard* versus *soft* attention over option choices: 1) when generating training data, 2) on abstract value learning in long-horizon tasks, and 3) when increasing the number of choices.
3. We empirically demonstrate the *paradox of choice* – that fewer but more meaningful choices can be better for both learning speed and final performance for an HRL agent.

2 Preliminaries

Markov Decision Process (MDP). A finite, discrete time MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r, P \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \rightarrow Dist(\mathcal{S})$ is the transition dynamics, mapping state-action pairs to a distribution over next states. At each time step t, the agent observes a state $S_t \in \mathcal{S}$ and takes an action $A_t \in \mathcal{A}$ drawn from its policy π. A stochastic policy is a conditional distribution over actions given a state $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ and a deterministic policy is a mapping $\pi : \mathcal{S} \rightarrow \mathcal{A}$. The infinite-horizon cumulative *discounted* return is defined as: $G_t = \sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t)$, where $\gamma \in (0, 1)$ is the discount factor. The value function for π is defined as: $V_\pi(s) = \mathrm{E}_\pi[G_t | S_t = s]$ and action value function is: $Q_\pi(s, a) = \mathrm{E}_\pi[G_t | S_t = s, A_t = a]$.

Hierarchical Reinforcement Learning (HRL) aims to find closed-loop policies that an agent can choose to use for some extended period of time, also known as *temporally extended actions*. Various HRL approaches have been proposed (Dayan and Hinton, 1993; Thrun and Schwartz, 1995; Parr and Russell, 1998; Dietterich, 2000; Vezhnevets et al., 2017; Nachum et al., 2018). Our work uses the options framework (Sutton et al., 1999), because it is general, and it also connects well with the notion of attention over actions, through the concept of initiation sets.

A Markovian option (Sutton et al., 1999) $\omega \in \Omega$ is composed of an *intra-option policy* π_ω, a termination condition $\beta_\omega : \mathcal{S} \rightarrow [0, 1]$, where $\beta_\omega(s)$ is the probability of terminating the option upon entering state s, and an initiation set $I_\omega \subseteq \mathcal{S}$. In the *call-and-return* option execution model, when an agent is in state s, it first examines the options that are available, i.e., for which $s \in I_\omega$. Let $\Omega(s)$ denote this set of available options. The agent then chooses $\omega \in \Omega(s)$ according to the policy over options $\pi_\Omega(s)$, follows the internal policy of ω, π_ω, until it terminates according to β_ω, at which point this process is repeated. This execution model defines a **semi-Markov decision process (SMDP)** (Puterman, 1994), in which the amount of time between two decision points is a random variable depending on the state-option pairs. Options can be learned by defining subgoal states and positively rewarding the agent for reaching them (Sutton et al., 1999; McGovern and Barto, 2001; Vezhnevets et al., 2017), in which case they are called **subgoal options**.

Affordances. Affordances are a property of the agent and its environment (Gibson, 1977; Heft, 1989; Chemero, 2003), indicating that only certain actions are feasible given certain features of the state. Affordances have been typically formalized in context of the objects (Slocum et al., 2000; Fitzpatrick et al., 2003; Lopes et al., 2007; Montesano et al., 2008; Cruz et al., 2016, 2018), e.g. a chair affords sitting, a cup affords grasping, etc. In classical AI systems, affordances have been viewed as preconditions (Fikes et al., 1972) for actions. Our view of affordances is most aligned with goal-based priors (Abel et al., 2014, 2015). In particular, we build upon the notion of action affordances defined via intents (Khetarpal et al., 2020a). Intents are defined as a target state distribution that should be achieved after the execution of an action, thereby defining a notion of action success. We will extend this concept to options, by considering intent achievement as a property that also depends on time, similarly to work on empowerment (Salge et al., 2014) or on goal conditioning (Schaul et al., 2015; Nachum et al., 2018).

3 Choice Attention for Reinforcement Learning

Consider that an agent must choose from many different action possibilities at every state as depicted in Fig. 1. In value-based methods, every choice involves computing the action-value function of all options, in order to infer which one is best. This process can be expensive, if there are many action choices. Moreover, the values of different actions are imperfect estimates, and if incorrect, they can mislead the agent into a poor decision. In temporal abstraction, these estimation errors can affect options unevenly, since they can depend on other factors that affect variance, such as the duration of the options or the termination condition. Hence, comparing options of different duration can be problematic. Using policy-based methods may appear to be a solution, but these methods still leverage value estimates in training, so the problem described still exists (albeit hidden). Finally, agents that do explicit planning using a model may further suffer from having a large number of choices, because it explodes their search space.

Attention mechanisms in animals (Hayhoe and Ballard, 2005; Tatler et al., 2011) as well as in computer vision (Xu et al., 2015; Luong et al., 2015) and in language (Brown et al., 2020) can be used to avoid the effect of noise and improve performance. In HRL agents, attention over actions can provide a similar effect. We consider and compare two approaches for implementing attention in this context.

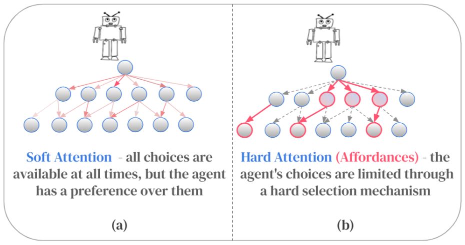

Soft Attention - all choices are available at all times, but the agent has a preference over them
(a)
Hard Attention (Affordances) - the agent's choices are limited through a hard selection mechanism
(b)

Figure 1: Soft attention vs hard attention over option choices. Soft preferences are indicated by different weights of arrows (a). Affordances, like hard attention, allow only a subset of choices, represented as solid arrows (b)

Soft-attention (Fig. 1(a)) places "soft" preferences over action choices, but does not rule out any of them. Interest functions (Khetarpal et al., 2020b) for options can be used to implement this idea. Appropriately tuned interest can reduce the noise in an agent's behavior. In contrast, an agent that understands which actions are affordable can use the information to strictly limit its attention to much fewer choices (Fig. 1(b)). Consequently, affordances can be viewed as *hard* attention (Luong et al., 2015; Xu et al., 2015). In model-free RL, this leads to a much narrower distribution of states from which to bootstrap, potentially lowering the variance of the value updates. The benefits further increase for planning, where option models allow the agent to reason "deeper", by considering a longer horizon, while affordances can reduce breadth of the search tree.

The potential gains obtained by using affordances (i.e. hard attention) can further be understood through the ***paradox of choice***, an idea due to Herbert Simon and popularized by Schwartz (2004). The main idea is that having too many choices can be worse for decision making, because the optimization problem becomes harder. A related concept is that of finding "satisficing" solutions for a problem, rather than fully optimizing it. A reduced number of choices can facilitate this process. While these ideas have been mainly discussed in the context of human decision making, our work can be viewed as providing a possible implementation and evaluation for artificial RL agents. Intuitively, HRL is itself an attempt to replace exact return maximization, with a solution that is sub-optimal but easier to compute, i.e., satisficing.

4 Affordance-Aware Subgoal Options

We now introduce *affordance-aware subgoal options* in Sec. 4.1, and propose a online, model-free algorithm to learn both options and the policy over options with given subgoals (Sec. 4.2).

4.1 Problem Formulation

Let $g \in \mathcal{S}$ be an arbitrary subgoal state, which is the desired target for an option. We assume for simplicity that subgoals are given[2]. Similarly to Khetarpal et al. (2020a, 2021), we will have to define intents and affordances for options, albeit with respect to sub-goals.

Definition 1 (Subgoal-option Intent Satisfaction): *A state s satisfies intent $I_{\omega,g}$ to degree $\epsilon \in (0,1)$ if the probability of reaching g while executing ω from s, before ω terminates, is larger than ϵ.*

Consider trajectories τ generated by running ω from s until completion. On any trajectory, consider an indicator function that marks if g was achieved or not. The intent $I_{\omega,g}$ is then achieved if the expected value of this indicator function, which we call *intent completion function*, is above ϵ.

Definition 2 (Option Affordances $\mathcal{AF}_{\mathcal{G}}$): *Given option set Ω, subgoal set $\mathcal{G} \subseteq \mathcal{S}$ and $\epsilon > 0$, the affordance associated with $\mathcal{G}$, $\mathcal{AF}_{\mathcal{G}} \subseteq \mathcal{S} \times \Omega$, is a relation such that $\forall (s,\omega) \in \mathcal{AF}_{\mathcal{G}}$, state s satisfies intent $I_{\omega,g}$ to degree $\epsilon \in (0,1)$*

Definition 3 (Affordance-aware Subgoal Options): *Given a subgoal g, an affordance-aware subgoal option $\omega \in \Omega$ is composed of a tuple $\langle \mathcal{AF}_g(\omega), \pi_\omega(a|s), \beta_\omega(s) \rangle$, where $\mathcal{AF}_g(\omega)$ denotes the set of states that afford ω with respect to g, π_ω is the intra-option policy, and $\beta_\omega(s)$ is the termination condition.*

The details for computing affordances are discussed in the next section and a visual interpretation of these definitions is presented in App A.2.

4.2 Learning Algorithm

We assume that a set of subgoals is provided to the agent. These subgoals can be associated with pseudo-rewards, and we can train options that attempt to maximize these. Simultaneously, the corresponding affordances can be constructed by identifying the states in which the options achieve their intended goal. We now present the details of this process.

Core Idea. For a given option $\omega \in \Omega$ and subgoal g, the intent completion function takes a trajectory generated by the option's internal policy, π_ω, and returns a binary value, indicating whether the desired subgoal is reached on that trajectory. This offers a target to learn the affordance $\mathcal{AF}_{\mathcal{G}}(\omega)$. To facilitate learning, we would like the intent completion to reflect *gradual* change, as the agent approaches the completion of the desired subgoal. To model this temporally extended aspect, we use a *discounted*-intent completion function ($\texttt{IC}$) with discount $\gamma_{\texttt{IC}} \in [0,1)$. Note that this discount factor is different than the one used to learn option policies. Moreover, the intent completion function measures the completion for all partial trajectories of an option: for trajectory $\tau = \{s_0, s_1, \dots, s_t\}$, all partial trajectories $\{s_0, s_1\}, \{s_0, s_1, s_2\}, \{s_0, \dots, s_t\}$ are evaluated for intent completion, resulting in the following target: $\texttt{IC}^{target}(s_t, \omega_t) =$

$$\begin{cases} \gamma_{\texttt{IC}} \times \texttt{IC}^{target}(s_{t+1}, \omega_t), & \text{if } \mathbb{1}_g(\tau_{s_1:s_t}) = 0 \\ \mathbb{1}_g(\tau_{s_1:s_t}), & \text{otherwise} \end{cases} \tag{1}$$

Data Generation. The key idea to learning subgoal options in an online, model-free fashion is for the agent to generate a training dataset of option trajectories labelled with intent completion targets (Eq. 1). This is challenging due to the *chicken-and-egg* problem of knowing which option to sample, *while* the options are being learned. In applications where a wealth of data is available, our approach could leverage expert trajectories, as in imitation learning, or human-in-the-loop learning. However, here we deal with a more challenging, online scenario. To cope with online data generation, we investigate the role of different attention mechanisms, namely, uniformly random, soft, and hard

[2] Note that many approaches have been proposed to learn sub-goals, and they could be used in an outer loop for our approach.

attention. It is important to note that the attention is learned simultaneously (through Eq. 1). This poses an additional challenge: in the initial stages of learning, the attention values are almost random. To better understand this, we further investigate if knowing which options to sample facilitates learning of options themselves (see Sec. 5.1.1).

Attention Mechanism. We train a convolution neural network (see Sec. 5 for details), referred to as $\texttt{IC}^{net}$, which takes as input a state and option id and predicts the discounted intent completion ($\texttt{IC}^{target}$). To obtain *hard*-attention, i.e. ***affordances***, we consider a *hard*-threshold on the predicted value. The threshold value is a hyper-parameter which controls the number of affordable options, based on the distance to the corresponding subgoal, i.e. subgoal reachability. A threshold of k means that values of discounted intent completion lower than γ_{IC}^{k} are replaced by 0s, sparsifying the set of options that the agent considers. The effect is two-fold: 1) options which are further from completing their associated subgoal are deemed less affordable and 2) the option duration is in effect controlled according to the choice of subgoal-intent specification. With a well designed choice of intents, obtaining affordances by thresholding can significantly reduce the size of the policy class used by the agent, thereby reducing the problem complexity, especially in large action spaces. This approach relies on having good subgoals, as they ultimately induce both the option policies and the intents. In this paper, we rely on domain knowledge to generate good subgoals.

Choice of Threshold & Sensitivity. We chose the threshold k by analyzing the normalized predicted values of $\texttt{IC}^{net}$. We observed that a hard-threshold of $k = 90$ separates the two distributions of affordable and non-affordable subgoals with high accuracy. Thus, we use this value of threshold for all our experiments. We provide more details in the App A.4. Next, We investigate the sensitivity to varying values of $k \in \{(5, 15, 90, 140\}$. The results corroborate the intuition that lower values of k (e.g. 5, 15) lead to fewer affordable options, including no affordable options for states farther in time (due to our design of affordances via discounted intent). However, larger values of k (e.g 90, 140) result in more options being affordable, which facilitates generating option trajectories with better coverage of the state space. We demonstrate a detailed evaluation of threshold values, $\texttt{IC}^{net}$ predictions and training sensitivity to them in the App A.4.

To obtain *soft*-attention, i.e. ***interest***, we compute a softmax over $\texttt{IC}^{net}$ predictions for all options given a state. While our implementation of interest over option choices (i.e. soft-attention) is closely related to the interest function introduced in interest-option-critic (IOC, Khetarpal et al., 2020b), there are two key differences: 1) soft attention in our approach is learned using downstream-task-agnostic intent completion targets, as opposed to a task-specific reward function in IOC, and 2) we use a softmax over predicted discounted intent completion values, instead of gradient-based updates to parameterized interest as in IOC. App. Alg. 3 shows the computation of the attention mechanisms. One could also consider softmax with temperature function, as it facilitates transitioning from soft-attention to hard attention. See detailed analysis in the App A.5.3.

Algorithm 1 Subgoal Option Learning (Simultaneous use & learning of attention in blue)

```
Require: Intents $\mathcal{I}$, Environment $E$, Intent completion function $I_{\omega,g}$, Training iterations $mtrain$.
Require: Number of options, Maximum option length $n_{max}$, Attention type.
Initialize parameters $z, \theta$ for $\texttt{IC}_z^{net}$, $\pi_{\omega,\theta}$, $\beta_{\omega,\theta}$
for $m = 1 \ldots mtrain$ do
    $\mathcal{D}, \mathcal{D}_\tau \leftarrow$ DataGeneration with attention using Alg. 2.
    Update $\pi_{\omega,\theta}(a|s)$ via PPO towards $I_{\omega,g}$ as reward. ($\mathcal{D}$)
    Update $\beta_{\omega,\theta}(s')$ with BCE Loss via $I_{\omega,g}$ as target ($\mathcal{D}$)
    Update $\texttt{IC}_z^{net}(s, \omega)$ with MSE loss towards Eq. 1 ($\mathcal{D}_\tau$)
end for
```

Option Training. Given a dataset of subgoal-option trajectories with intent completion labels, we compute and propagate the gradients for the respective losses as in Alg. 1. All components of options are parameterized and trained as follows. We use a common neural network backbone and different fully connected layers for options policies, terminations and value estimations. Option policies are trained with proximal policy gradient methods (PPO) (Schulman et al., 2017)), with the objective of maximizing intent completion. Since we would like the termination parameterization to learn to predict termination condition for all options at the same time, it is trained with standard binary cross entropy (BCE) loss. Finally the $\texttt{IC}^{net}$ is trained using mean-squared-error (MSE) loss towards the intent completion target, for all partial option trajectories (Eq. 1). The predictions from the $\texttt{IC}^{net}$ are used to compute the respective attention mechanisms described above. Implementation and training details are provided in App A.3.

5 Empirical Analysis

To investigate the role of attention in decision making with temporal abstraction, we empirically tackle the following questions: **Q1.** Does hard attention result in improved sample efficiency during online data generation? Yes (See Sec. 5.1.1). **Q2.** Does an HRL agent with affordances (hard attention) perform better in long-horizon sparse reward tasks? Yes. In harder tasks, hard attention outperforms soft attention, no attention, and a flat agent with pseudo-reward, potentially due to improved credit assignment. See results on MiniGrid supporting this in Sec. 5.1.3. **Q3.** Does hard attention help when increasing the number of choices? Yes (See Sec. 5.1.4). **Q4.** Does our analysis scale to continuous control? Yes (See Sec. 5.2).

Motivation for choice of domains. We choose MiniGrid and Fetch domain to allow for 1. specification of intentions, 2. varying complexity in terms of number of options and tasks, 3. isolation and focused investigation of the role of attention, and 4. to demonstrate that our approach scales to both discrete and continuous domains. Implementation details along with the hyperparameters used for all the experiments[3] are provided in App A.3.

5.1 Discrete Domain: MiniGrid

Environment & Task Specification. We use the MiniGrid (Chevalier-Boisvert et al., 2018), adapted for different configurations of sub-goals. The agent receives a $2D$ fully observable, egocentric, multi-hot encoded view of the map. The action space is: turn left, turn right, move forward, pick up. A task is specified by the number of unique objects. Task is completed when the agent collects all the objects on the map in the correct sequence, upon which the agent receives a positive reward. At the beginning of each episode, a task is sampled at random from a distribution of tasks. There are no common objects between tasks. The objects can differ in type, colour, and disappear from the map once collected.

5.1.1 On The Role Of Attention In Data Generation

Experimental Setup. We first evaluate the role of different types of attention while training option policies i.e. during pre-training. We use 4 tasks, each with 3 unique objects to be collected at the beginning of the episode. The agent has to learn 12 options, from which only a maximum of 3 can achieve their subgoal at any given time. A task is sampled randomly, rewarding each option only when it completes its intended subgoal. A new option has to be selected based on the choice of the sampling method.The option policy, termination, and attention mechanism are learned simultaneously. During the training of option policies, we also train the $\texttt{IC}^{net}$, which predicts the discounted intent completion. This poses the challenge of online data generation for learning both the options and attention, at the same time (see Sec. 4.2). To gain a better understanding of the role of attention during sampling, we compare and highlight differences when data is generated using different attention mechanisms i.e. 1) **hard-attention**: hard thresholding of the predicted discounted intent, 2) **soft-attention**: normalized predicted discounted intents and 3) **no-attention**: sampling uniformly randomly from all choices.

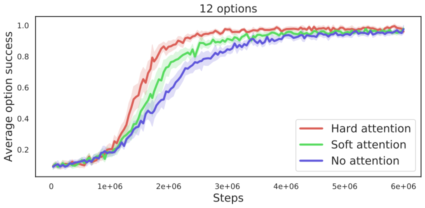


Figure 2: Option policy evaluation shows that sampling with hard-attention results in faster policies convergence. We plot the average of 10 runs and 95% confidence intervals.

Results. For a fair comparison of the quality of the learned options, the option policies are evaluated after each update by comparing their success rate in achieving their corresponding subgoal intent. For evaluation we roll out a set of 64 ground truth affordable options from random initial states. We observe in Fig. 2 that sampling from *hard* attention improves the learned option policies, compared to sampling with *soft* attention or no attention. Leveraging affordances as *hard* attention yields a sample efficient approach to generate training data. $\texttt{IC}^{net}$ learns faster than the option policies, corroborated

[3] Code for all the experiments is available on *

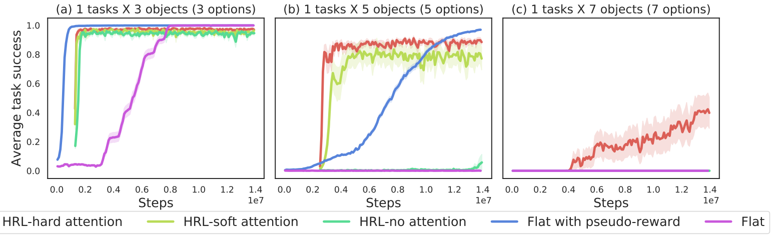


Figure 3: Long Horizon Sparse Reward Tasks with progressively increasing horizon, via 3 to 7 sequential objects. With increasing difficulty, the gap between the hard and soft attention increases. The lines depict the average of 10 runs and 95% confidence intervals.

also by further analysis of its precision during training in App A.5.1. As a result, knowing which option is strictly affordable narrows the data distribution to fewer but more useful choices. This speeds up the option learning process, as the dataset required by Alg. 1 contains more samples of successful option trajectories. Instead, soft-attention dilutes the preferences across all option choices. Despite the success of *hard*-attention in our empirical evaluation, it is intuitive to see that *soft*-attention might be a better choice in some scenarios. See App A.1 and A.5 for discussion and additional experiments respectively.

5.1.2 On The Role Of Attention In Leveraging Previously Acquired Skills

Next, we train HRL agents at the SMDP level using PPO algorithm to learn an optimal policy over (pre-trained) options to solve the downstream task. We first learn a set of subgoal-options via Alg. 1 in an environment E using different types of attentions. The agent now samples an option based on the learned policy over options ($\pi_\Omega(\omega|s)$) modulated by this attention (Alg. 3) as follows:

$$\pi_\Omega(\omega|s) \propto f_{att}(s,\omega)\pi_\Omega(\omega|s) \tag{2}$$

The sampled option policy ($\pi_\omega(a|s)$) runs until the learned termination ($\beta_\omega(s)$). $\pi_\Omega(\omega|s)$ is learned using the Generalized Advantage Estimator (GAE): $\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{l=0}^{T}(\gamma^m\lambda)^l\delta_{t+1}^V$, where $\delta_t^V = r_t + \gamma^m V(s_{t+1}) - V(s_t)$, r_t is the task-specific reward, λ is a hyper-parameter for horizon, δ_t^V is the temporal-difference residual, and $\hat{A}^{GAE}$ the advantage estimation. Both δ_t^V and $\hat{A}^{GAE}$ are discounted via γ^m in order to account for the option duration m.

We compare the following methods (See App A.3.4 for implementation details) 1) A **flat agent** (PPO) using primitive actions, 2) An **HRL agent with no attention**, 3) An **HRL agent with *hard*-attention**, and 4) An **HRL agent with *soft*-attention**. The HRL agents use pre-trained options with the same attention as in the downstream task and the maximum option pre-training steps are determined based on the option policy performance plateau. We now investigate two settings, namely long horizon sparse reward tasks (Sec. 5.1.3)) and increasing number of option choices (Sec. 5.1.4)).

5.1.3 On The Role Of Attention In Long horizon sparse reward tasks

Experimental Setup. The environment is comprised of only one task, but with different numbers of objects to be collected per task (i.e. $3, 5, 7$). The tasks are progressively harder due to the delayed reward with increasing number of objects.

Results. In Fig. 3 we compare the aforementioned methods alongside an additional baseline of the flat agent with a pseudo-reward for whenever the agent collects any of the objects, akin to the subgoal information. We observe that as the total task length is progressively longer, the HRL agent copes better in this more challenging situation as compared to the flat agent (purple) due to faster reasoning at the SMDP level. More importantly, we notice that the HRL agent with *hard*- attention (red) consistently outperforms HRL agent with *soft*- attention due to the agent's ability to 1) strictly restrict its attention to much fewer choices and 2) to cope with imperfect option policies. Our interpretation of this result is that limiting an agent's attention to fewer but meaningful choices potentially results in improved credit assignment particularly in such long-horizon sparse reward tasks. While the flat agent with pseudo-reward outperforms the options agent in easy tasks (Fig 3a.), in more challenging longer horizon tasks (Fig 3c.), it is unable to learn even with significantly high number of training steps. In this harder task the HRL agent with *hard*- attention is the only one able to learn.

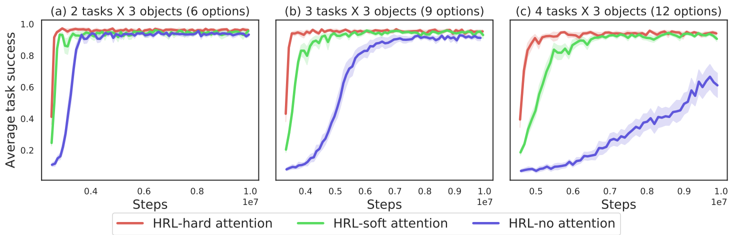


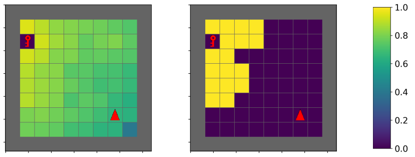
Figure 4: SMDP value learning with increasing number of choices (using pre-trained options). The gap between hard and soft attention increases as more choices are to be considered demonstrating the paradox of choice. We plot the average of 10 runs with 95% confidence intervals.

5.1.4 Increasing number of choices

Experimental Setup. Next, we investigate the role of attention with growing number of choices, by varying only the number of tasks. The agent is tasked to collect a fixed number of objects (3 in this case) per task, in the correct order. Leveraging the pre-trained options, we learn a policy over options in order to solve the task (as detailed in Sec. 5.1.2).

Results. Fig.4 depicts the average task success for increasing number of choices from left to right, namely $6, 9$, and 12 options. We observe that as the decision maker has more choices to consider, restricting the agent's action possibilities strictly (such as via *hard*-attention) results in better performance highlighting the paradox of choice. We anticipate with growing number of choices, knowing what is affordable as a *hard*-attention criteria enables agents to be much more sample efficient than agents that dilute attention over all choices (green) and agents that do not attend selectively (purple).

5.1.5 Qualitative Analysis



(a) $\texttt{IC}^{net}$ (b) Affordances

Figure 5: Qualitative Analysis shows (a) predictions from $\texttt{IC}^{net}$ for *pick-up-the-key*. (b) A hard thresholding on $\texttt{IC}^{net}$ predictions generates option affordances (threshold $\gamma^{10}_{\texttt{IC}^{net}}$).

We analyze the predicted discounted intent completion target (Eq. 1) and affordances qualitatively via a heatmap in MiniGrid in Fig. 5. We observe that learned subgoal-intent completion is peaked in states around the subgoal such as near the key (Fig. 5a). Higher values are denoted in brighter yellow while the lower ones in darker colors. See additional examples in App A.4. To obtain option affordances, we apply a hard threshold on the $\texttt{IC}$ network predictions. Here we plot the affordances for a threshold of 10 steps, restricting the option to be affordable for intent completion values $\geq \gamma^{10}_{IC}$.

5.2 Continuous Control: Fetch Domain

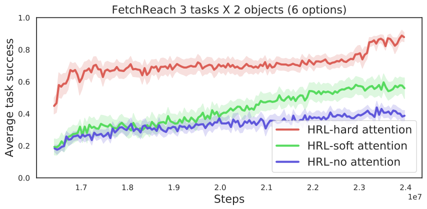


Figure 6: SMDP value learning in Fetch Reach. We plot the average of 10 runs with 95% confidence intervals.

Experimental Setup. To demonstrate choice attention in HRL scale for continuous control robotics environment, we consider tasks with multiple goals per task in an adaptation of the FetchReach-v0 domain (Schulman et al., 2017). The agent must learn to move a robotic gripper close to a set of target positions in the correct order. All goal positions are represented in the input observation by either a valid coordinate from the state space or a null coordinate for unavailable targets. A sparse reward is used for both option policy training (intent completion) and downstream task (See Sec. 5.1.2). See Appendix A.3.5 for implementation details.

Results. Fig.6 shows that affordance aware subgoal options have an advantage in learning even when considering 6 options in continuous space. Although the option policies are limited from pre-training (lower maximum downstream reward possible), affordance aware SMDP learning can still show a major advantage in performance in significantly lesser time. See App A.5.4 for option policy and HRL training evaluation for different number of goals.

6 Related Work

Attention is a widely studied concept in the vision and language community. For e.g., when considering images, *soft*-attention places learned multiplicative weights over *all* patches in the input image, Bahdanau et al. (2014); soft-attention has access to the entire image. On the other hand, *hard*-attention only selects one patch of the image to attend to (Luong et al., 2015; Xu et al., 2015). This already established distinction can be juxtaposed with the characterization of attention that we use in our work in terms of action choices. Specifically, *hard*-attention via affordances only allows certain option choices and strictly limits the agent's focus to parts of the option space. *Soft*-attention, on the other hand, places preferences over *all* actions.

Learning options has been extensively explored in the literature (Parr and Russell, 1998; Thrun and Schwartz, 1995; Dayan and Hinton, 1993; Dietterich, 2000; Stolle and Precup, 2002; McGovern and Barto, 2001). Our work can be viewed as leading to options with initiations sets (Sutton et al., 1999). Khetarpal et al. (2020a) proposed end-to-end gradient based discovery to learn parameterized interest functions, a generalization of initiation sets, by optimizing a task specific reward signal. In contrast, we consider learning options with given subgoals, akin to using pseudo-rewards (Kulkarni et al., 2016).

Our work is related to *hierarchical affordance competition* (Pezzulo and Cisek, 2016), where interactive decision making entails a competition between representations of available actions (affordances) biased by the desirability of their predicted outcomes (intents). Here, we consider that we are given desirable outcomes of an *intentional action* via the specification of subgoals. The affordances, as a *hard*-attention mechanism, facilitate reasoning across different time scales, by predicting the expected future outcomes of options using the discounted intent completion. Ahn et al. (2022) interpret value functions as affordance functions that capture the log likelihood that a particular skill will be able to succeed in a state in order to ground large language models in the real-world. In RL, the closest to our work is Xu et al. (2020); a method that incrementally builds environment models around the affordances of parameterized motor skills. Unlike their work, we propose a model-free online algorithm. We also assume that subgoal information is known apriori and utilized as intents. More recent, concurrent work (Khetarpal et al., 2021) considers a model-based approach in building partial option models based on affordances and is complementary to our work.

7 Discussion and Future Work

We studied the role of choice attention in RL agents. To this end, we presented the notion of subgoal option affordances through the lens of *hard*-attention. This study demonstrated the paradox of choice, i.e. fewer but more meaningful choices can be better for both learning speed and final performance, compared to *soft*-attention over all option choices.

We have focused on two types of domains, discrete (MiniGrid) and continuous (Fetch). While relatively small, they allow us to do careful quantitative and qualitative analysis, to control the number and difficulty of the tasks. In addition, this facilitated isolating the role of affordances as hard attention, while being able to vary the complexity in terms of number of options and tasks. We expect our approach to be especially useful in tasks where intent achievement can be easily available through domain knowledge, such as robotics or dialog management tasks.

While we assume prior knowledge of the subgoals, note that learning intents, affordances and options simultaneously would have made it harder to study hard vs soft attention. The choice of intents/subgoals presented provided a ground truth for the hard attention, which facilitated the evaluation and clarity that this work brings in the matter, and suggests a clear path towards future work. Indeed, this constraint could be relaxed through a two-step iterative algorithm, 1) subgoal discovery, e.g. using an information theoretic objective, such as in Harutyunyan et al. (2019), 2) iterative learning of affordance-aware subgoal options.

Acknowledgments and Disclosure of Funding

The authors would like to thank Emmanuel Bengio, Joelle Pineau for detailed feedback, Zafarali Ahmed, Martin Klissarov for valuable comments on a draft of this paper. A special thank you to Emmanuel Bengio for helping us edit the paper. The authors also thank the anonymous ICML, NeurIPS, and AAAI reviewers.

References

Abel, D., Barth-Maron, G., MacGlashan, J., and Tellex, S. (2014). Toward affordance-aware planning. In *First Workshop on Affordances: Affordances in Vision for Cognitive Robotics*.

Abel, D., Hershkowitz, D. E., Barth-Maron, G., Brawner, S., O'Farrell, K., MacGlashan, J., and Tellex, S. (2015). Goal-based action priors. In *Twenty-Fifth International Conference on Automated Planning and Scheduling*.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., et al. (2022). Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Badman, R. P., Hills, T. T., and Akaishi, R. (2020). Multiscale computation and dynamic attention in biological and artificial intelligence. *Brain Sciences*, 10(6):396.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Borji, A., Sihite, D., and Itti, L. (2012). Salient object detection: A benchmark. computer vision—eccv 2012: the 12th european conference on computer vision; 2012 oct 7-13; florence, italy.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological psychology*, 15(2):181–195.

Chevalier-Boisvert, M., Willems, L., and Pal, S. (2018). Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid.

Cruz, F., Magg, S., Weber, C., and Wermter, S. (2016). Training agents with interactive reinforcement learning and contextual affordances. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):271–284.

Cruz, F., Parisi, G. I., and Wermter, S. (2018). Multi-modal feedback for affordance-driven interactive reinforcement learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Dayan, P. and Hinton, G. E. (1993). Feudal reinforcement learning. In *Advances in neural information processing systems*, pages 271–278.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303.

Fikes, R. E., Hart, P. E., and Nilsson, N. J. (1972). Learning and executing generalized robot plans. *Artificial Intelligence*, 3:251–288.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning about objects through action-initial steps towards artificial cognition. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, volume 3, pages 3140–3145. IEEE.

Gibson, J. J. (1977). The theory of affordances. *Hilldale, USA*, 1(2).

Harutyunyan, A., Dabney, W., Borsa, D., Heess, N., Munos, R., and Precup, D. (2019). The termination critic. *arXiv preprint arXiv:1902.09996*.

Hayhoe, M. and Ballard, D. (2005). Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194.

Heft, H. (1989). Affordances and the body: An intentional analysis of gibson's ecological approach to visual perception. *Journal for the theory of social behaviour*, 19(1):1–30.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE.

Khetarpal, K., Ahmed, Z., Comanici, G., Abel, D., and Precup, D. (2020a). What can i do here? a theory of affordances in reinforcement learning.

Khetarpal, K., Ahmed, Z., Comanici, G., and Precup, D. (2021). Temporally abstract partial models. *Advances in Neural Information Processing Systems*, 34.

Khetarpal, K., Klissarov, M., Chevalier-Boisvert, M., Bacon, P.-L., and Precup, D. (2020b). Options of interest: Temporal abstraction with interest functions. *AAAI*.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*.

Lopes, M., Melo, F. S., and Montesano, L. (2007). Affordance-based imitation learning in robots. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1015–1021. IEEE.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

McGovern, A. and Barto, A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density.

Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. (2018). Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3303–3313.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Nobre, A. C. and Stokes, M. G. (2019). Premembering experience: a hierarchy of time-scales for proactive attention. *Neuron*, 104(1):132–146.

Parr, R. and Russell, S. J. (1998). Reinforcement learning with hierarchies of machines. In *Advances in neural information processing systems*, pages 1043–1049.

Pezzulo, G. and Cisek, P. (2016). Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, 20(6):414–424.

Pineau, J. (2019). The machine learning reproducibility checklist.

Plappert, M., Andrychowicz, M., Ray, A., McGrew, B., Baker, B., Powell, G., Schneider, J., Tobin, J., Chociej, M., Welinder, P., et al. (2018). Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*.

Puterman, M. (1994). Markov decision processes. 1994. *Jhon Wiley & Sons, New Jersey*.

Salge, C., Glackin, C., and Polani, D. (2014). Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1312–1320, Lille, France. PMLR.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schwartz, B. (2004). The paradox of choice: Why more is less. Ecco New York.

Slocum, A. C., Downey, D. C., and Beer, R. D. (2000). Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In *From animals to animats 6: Proceedings of the sixth international conference on simulation of adaptive behavior*, pages 430–439.

Stolle, M. and Precup, D. (2002). Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*. Springer.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Mahmood, A. R., and White, M. (2016). An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17:73:1–73:29.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5.

Thrun, S. and Schwartz, A. (1995). Finding structure in reinforcement learning. In *Advances in neural information processing systems*, pages 385–392.

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pages 3540–3549. PMLR.

White, M. (2017). Unifying task specification in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3742–3750.

Xu, D., Mandlekar, A., Martín-Martín, R., Zhu, Y., Savarese, S., and Fei-Fei, L. (2020). Deep affordance foresight: Planning through what can be done in the future.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Zhang, R., Walshe, C., Liu, Z., Guan, L., Muller, K. S., Whritner, J. A., Zhang, L., Hayhoe, M. M., and Ballard, D. H. (2019). Atari-head: Atari human eye-tracking and demonstration dataset. *arXiv preprint arXiv:1903.06754*.