Attention-Based Experience Replay in Deep Q-Learning

Mirza Ramicic and Andrea Bonarini
Dipartimento di Elettronica,Informazione e Bioingegneria
Politecnico di Milano, Milan,Italy
{mirza.ramicic, andrea.bonarini}@polimi.it

ABSTRACT

Using neural networks as function approximators in temporal difference reinforcement problems proved to be very effective in dealing with high-dimensionality of input state space, especially in more recent developments such as Deep Q-learning. These approaches share the use of a mechanism, called experience replay, that uniformly samples the previous experiences to a memory buffer to exploit them to re-learn, thus improving the efficiency of the learning process. In order to increase the learning performance, techniques such as prioritized experience and prioritized sampling have been introduced to deal with storing and replaying, respectively, the transitions with larger TD error. In this paper, we present a concept, called *Attention-Based Experience REplay (ABERE)*, concerned with selective focusing of the replay buffer to specific types of experiences, therefore modeling the behavioral characteristics of the learning agent in a single and multi-agent environment. We further explore how different behavioral characteristics influence the performance of agents faced with dynamic environment that is able to become more hostile or benevolent by changing the relative probability to get positive or negative reinforcement.

CCS Concepts

• Computing methodologies → Reinforcement learning • Computing methodologies → Neural networks • Computing methodologies → Markov decision processes • Computing methodologies → Temporal difference learning

Keywords

Reinforcement Learning;Deep Learning;Deep Reinforcement Learning;Policy control;Congnitive Architectures.

1. INTRODUCTION

Implementation of approximation techniques widely used in supervised and unsupervised learning, namely artificial neural network architectures [1], enabled RL to cope with very large state spaces. This opened a possibility of applying RL techniques to more complex problems and gave rise to successful implementations such as playing Atari games [2, 3], which used a Deep Convolutional Neural Network to approximate the reward function.

Online agents learn from a stream of experiences: after each transition the *Temporal Difference* (TD) error is back-propagated through the neural network so that the previous approximation is updated. However, the sequence of experiences in RL can contain highly correlated samples that break the *Independent and Identically Distributed* assumption of artificial neural network architectures [4]. To reduce the temporal correlation between experiences and improve the speed ot learning, a technique called *Experience Replay* [1, 2] is used to allow an agent to reuse past experiences, therefore obtaining a more stable training of a neural network. The transitions are uniformly sampled and stored in a sliding window memory; after each transition a batch of the stored experiences are used to train the neural network.

Previous approaches have dealt with the dynamics of the replay memory mechanism in order to improve the speed of learning by focusing on the transitions that had a larger TD error in both experience sampling [5] and experience replay [4], but none was concerned about modifying the characteristics of the learning process itself.

In this paper, we are extending a biologically inspired technique of experience-replay memory, introducing the concept of attention-based working memory inspired by a cognitive mechanism system found in humans called working memory [6].

Attention-based experience replay has the ability to focus on different types of experience during sampling, thus enabling to model the behavioral differences in attention focus that can be found along the main human personality axis: extroversion/introversion. We show how different attitudes can face different environments with different performance. We also propose to consider cognitively inspired learning strategies to improve learning in environments with different characteristics.

2. THEORETICAL BACKGROUND

2.1 Reinforcement learning

A reinforcement learning process involves an agent learning from interactions with its environment in discrete time steps in order to update its mapping between the perceived state and a probability of selecting possible actions (policy). The agent performs a sequence of transitions of a Markov decision process represented by a tuple (s_t, a_t, r_t, s_{t+1}) and at each step updates its policy π_t in order to maximize the total amount of cumulative reward over the long run [7]. For this reason the optimal action-value function $Q^*(s,a)$ is defined as the maximum expected return following the policy π:

$$Q^*(s,a) = \max_{\pi} \mathbb{E}[R_t | s_t = s, a_t = a, \pi] \qquad (1)$$

ICMLC 2017, February 24-26, 2017, Singapore, Singapore
© 2017 ACM. ISBN 978-1-4503-4817-1/17/02…$15.00
DOI: http://dx.doi.org/10.1145/3055635.3056621

After each transition it is possible to update the estimation of the action-value function using Bellman equation as an iterative update in order to converge to the optimal action-value function:

$$Q_{i+1}(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q_i(s',a')|s,a\right] \qquad (2)$$

Equation (2) guarantees the convergence as $i \rightarrow\propto$, but it is impractical to use without any generalization and approximation when facing high dimensional state spaces. Instead, most practical approaches use function approximators to estimate the action-value function, which range from simple linear perceptrons to non-linear approximators such as neural networks.

2.2 Approximation

In a function approximation with neural networks, at each iteration, the weights Θ are updated by performing a gradient descent on the loss functions $L_i(\Theta_i)$ according to Equation (3) therefore improving the previous estimate of the optimal action-value function $Q(s,a;\Theta) \approx Q^*(s,a)$.

$$\nabla_{\Theta_i} L_i(\Theta_i) = \big(y_i - Q(s,a;\Theta_i)\big)\nabla_{\Theta_i} Q(s,a;\Theta_i) \qquad (3)$$

where $y_i = r + \gamma \max_{a'} Q(s',a';\Theta_{i-1})$ is the target for iteration.

Temporal difference learning combined with a deep neural network for approximation of action-value function is called *Deep Q-Learning*, or DQL [2].

3. ATTENTION-BASED REPLAY MEMORY

3.1 Cognitively Inspired Architectures

Studies have showed that human cognitive processes utilized during the interaction with the environment are mediated by a memory buffer called working memory [6]. The working memory keeps a temporary storage of the perceived information needed to perform a complex cognitive task: it acts as a connecting mechanism between perception and long term memory.

Experiments have identified that the differences between individuals in the capacity of working memory [8] and the breadth of attention generally influence the way they are focusing their attention and creative abilities [9]. The term “breadth of attention”, in this context, refers to a sort of cognitive bandwidth, i.e., the number and scope of stimuli that one is attending at a time.

Extroverted individuals tend to have a broader breadth of attention than the introverted ones, which, in turn, tend to focus their attention to a narrower subset of stimuli in order to reduce the cognitive load of having a higher basal arousal level [10, 11].

3.2 Model Architecture and Learning Algorithm

Using only uniform sampling as a way to store experiences in the replay memory proved to have limitations such as that some of the valuable experiences might never be replayed [5]. Attention-based replay memory keeps the uniform sampling and extends it by additionally sampling the experiences that emerged from a specific type of interaction. For the purpose of mapping the transition to a specific, goal-oriented interaction, we extend the experience description tuple with a transition type indicator $e_t = (s_t, a_t, r_t, c_t, s_{t+1})$.

The modification to the uniform sampling replay memory algorithm is that, in addition to sampling every Sth sample, we sample the experiences that match the subset of transition types, called F (for focus of attention), as shown in Algorithm 1.

Modifying the scope of F makes it possible to model the agents with different behavioral characteristics in both goal and trait oriented way, thus making them more adapted to learn in different environments.

Algorithm 1 DQL with attention-based replay memory

```
Initialize replay memory D with capacity N and sampling frequency $S$
Initialize and set transition types index $C = \{c_1, c_2, \ldots, c_n\}$ and attention focus index $F \subset C$
Initialize action-value function Q with random weights
for episode = 1, M do
    Initialize sequence $s_1 = \{x_1\}$ and pre-processed sequenced $\phi_1 = \phi(s_1)$
    for t = 1, T do
        With probability $\varepsilon$ select a random action $a_t$
        otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \Theta)$
        Execute action $a_t$, observe reward $r_t$ type of transition $t_t$ and image $x_{t+1}$
        Set $s_{t+1} = s_t, a_t, x_{t+1}$ and pre-process $\phi_{t+1} = \phi(s_{t+1})$
        if $i$ mod $S = 0$ then
            Store transition $(\phi_t, a_t, r_t, c_t, \phi_{t+1})$ in $D$
        end if
        for each f in F do
            if $c_t = f$ then
                Store transition $(\phi_t, a_t, r_t, c_t, \phi_{t+1})$ in $D$
            end if
        end for each
        Sample random batch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $D$
        set$y_i = \begin{cases} r_j, & \text{terminal}\phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \Theta), & \text{non terminal} \end{cases}$
        Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \Theta))^2$ according to Equation 3.
    end for
end for
```

4. EXPERIMENTAL SETUP

To evaluate the proposed model we have adopted a learning environment that consists of moving good/bad food pieces and multiple agents [12]. Food pieces are generated with a random speed and direction, and move in a constrained environment by bouncing on the walls. Agents can move in the same environment and should learn to touch (eat) good food pieces and to avoid bad food pieces.The goal of each agent is to consume as much good food pieces as possible, either directly or by interacting with other agents that can share food, while, in turn, try to avoid the bad food sources. After being consumed, new food pieces of the same type of the consumed ones are re-generated with a random speed and direction, thus keeping the distribution of food constant. Agents receive reinforcement of +1 for consuming good food pieces and -1 for consuming bad ones.

The state space is continuous and intentionally high-dimensional for the purpose of increasing the entropy and consequently the diversity of possible experience transitions. Each agent has 40 directional sensors and each of them can perceive 6 features: type of sensed object (good food, bad food, agent), as well as the continuous values for range and the velocity of the object detected; this gives a total of 240 state space inputs for each agent.

As a function approximator we are using a deep neural network to approximate $Q(s,a;\Theta) \approx Q^*(s,a)$. To reduce the computational complexity of having multiple forward passes each time, we want to find an action that maximizes the state-action function $\mathrm{argmax}_a Q(s,a)$; the network takes the state vector s as an input and predicts $Q(s,a)$ for each possible action.

We have adopted the original Q-learning update with a learning rate α set to a low value (0.05) because of the nature of the approximator, and discount factor $\gamma = 0.9$. The default capacity of the replay memory buffer D included 9000 experiences. For comparison with our proposed algorithm we performed reference experiments where we uniformly sampled experiences every 7th transition. With regards to our experimental environment this sampling frequency provided a balance between the transitions that were sampled uniformly and the ones that were sampled on the basis of attention focus.

We also performed experiments in a multi-agent setting. The multi-agent environment differed from the single-agent one in size and amount of food generated to accommodate up to 7 agents learning simultaneously. Agents in a multi-agent environment had a possibility of social interaction by sharing food with other agents in proximity, as detected by their sensors. If a single agent consumed a positive food piece it shared the full reinforcement reward of +1 to each of the agents found within its range.

5. EXPERIMENTAL RESULTS

In the experiments, we have compared three types of agents implementing different types of focus of *ABERE*, with the baseline uniform sampling already proposed in literature, under three different configurations of the environment. The transitions were given a focus type only if they resulted in an interaction, i.e., either a food piece has been consumed or an agent has been perceived. To differentiate between the interactions we have defined three focus types in $C = \{consume - good, consume - bad, social\}$. If the transition resulted in a consumption of good food, it was labeled as *consume-good*, if bad food was consumed it was labeled as *consume-bad*, and if it resulted in either sharing or receiving food through social interaction it was labeled as *social*.

Table 1 shows which agent personality type is associated with which subset of C.We call this subset *Attention focus* F as it represents the set of type labels on which Algorithm 1 additionally focuses while sampling from the stream of experiences. For instance, the *Introverted – Brave* agent focuses on consuming good food, i.e., it samples experiences labeled as *consume-good*. Analogously for the others.

Table 1. Personality types and attention focus

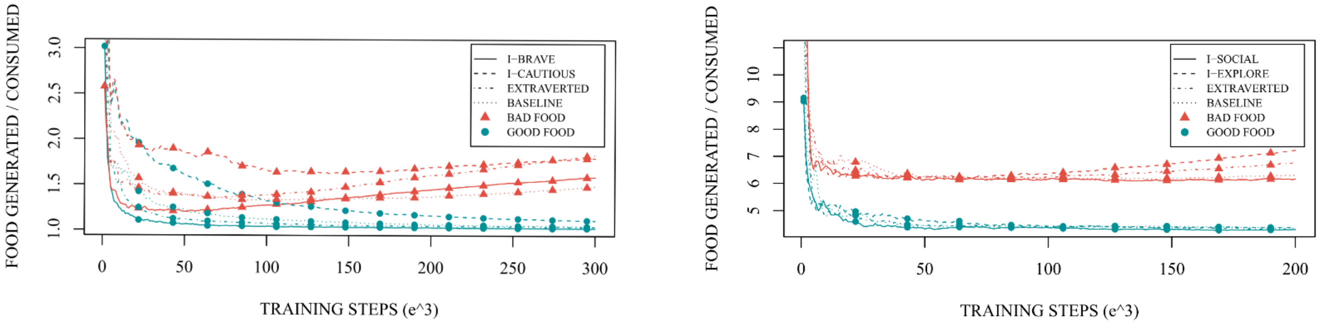
Single Agent Environment	
Agent personality type	*Attention focus (F)*
Introverted - Brave	consume-good
Introverted - Cautious	consume-bad
Extroverted	consume-good, consume-bad
Baseline	-
Multi Agent Environment	
Agent personality type	*Attention focus (F)*
Introverted - Social	social
Introverted - Explore	consume-good, consume-bad
Extroverted	social, consume-good, consume-bad
Baseline	-

5.1 Efficiency Comparison

In this section we evaluate the efficiency of agents with different configurations of *ABERE* with respect to the ability to consume good food pieces and avoid the bad ones in the environment with an equal distribution of good and bad food pieces. The aim is to compare the behavioral differences of the agents and their effect on the performance by two different criteria: ability to avoid the bad pieces of food and the ability to consume the good ones.

In Figure 1 we compare these criteria for each type of agent as a ratio between generated and consumed food pieces. Figure 1a shows the average results of 10 experiments done under the same settings for each of the defined agent type, while Figure 1b depicts analogous results averaged over 7 agents of the same type interacting in a multi-agent environment.

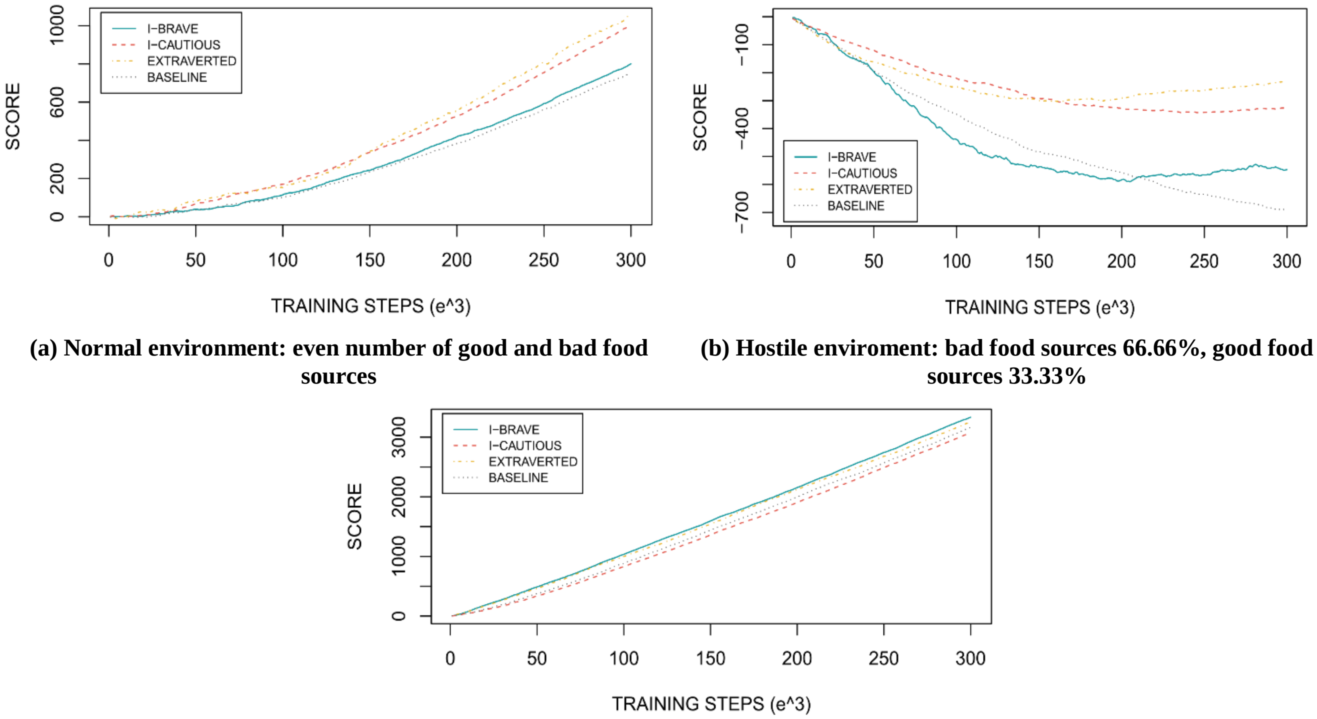
From Figure 1 we can notice that the efficiency of the agents differs depending of the agent type in both single and multi-agent environment. Introverted-Cautious agent type showed to be the most efficient in avoiding bad food sources followed by Extroverted type, while Introverted-Brave outperformed every other type in consuming good food sources. From these results, it seems that focusing on a given aspect pushes to efficiently develop a policy that takes better into account that aspect. We can also notice that ABERE agents generally perform better than the non-focused ones.



(a) Single agent environment

(b) Multi-agent environment with food sharing

Figure 1: Differences in ratio of generated and consumed food sources amounts between *ABERE* focus variations over first 300K learning steps.



(a) Normal environment: even number of good and bad food sources

(b) Hostile enviroment: bad food sources 66.66%, good food sources 33.33%

(c) Benevolent environment: bad food sources 33.33%, good food sources 66.66%

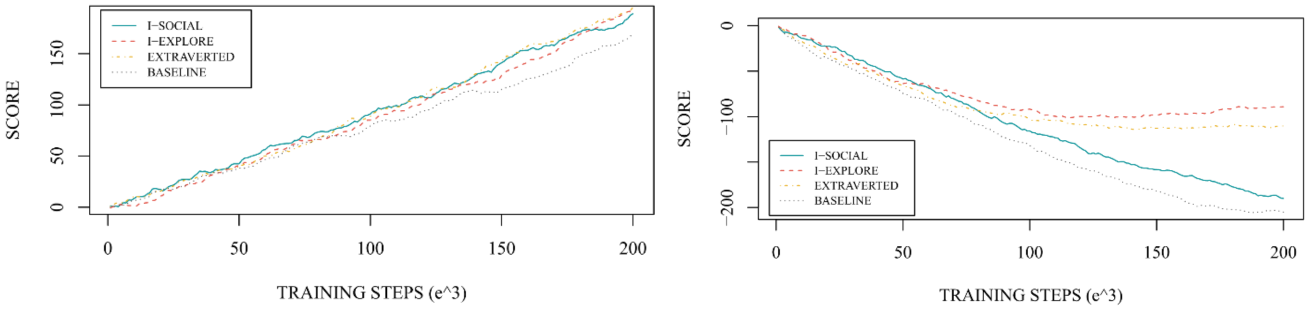
Figure 2: Differences in average score/reward between agents with *ABERE* focus variations learning in a single agent enviroment over first 300K learning steps.

5.2 Performance in different environmental conditions

In the second experiment, our intention was to explore how can differences in agent personality type impact on the performance under different environmental conditions. We wanted to answer the question: Can some personality type be more capable than others to learn in a specific environment?

We have modified the equal ratio between the generated good and bad food pieces for the purpose of creating more hostile or more benevolent environment. Benevolent environment generated 2/3 of good food pieces and 1/3 of bad, while the hostile environment had a distribution of 2/3 bad food pieces and only 1/3 good. Results from single agent simulation as depicted in Figure 2 show that the Extroverted agent was performing best in both normal and hostile environments, while Introverted-Brave type better adapted to the environment that contained more good food. It seems that the broader attention span of the Extroverted agent gave it an advantage in the environments that contained higher amount of bad food points. Focusing on both positive and negative experiences allowed the Extroverted agent to learn a policy that was equally efficient in avoiding the bad food points as it was in consuming the positive ones.

Figure 3 shows the results from a simulation that included 7 agents interacting by sharing food sources, each of them learning separately. For the normal environment configuration Introverted-Social and Extroverted types were best performing probably because their social focus allowed them to make better use of the available good food points by sharing. Introverted-Explore type outperformed others in a hostile environment mostly because its narrow focusing on the food points rather than social interaction allowed it to be more efficient in avoiding the bad food points.

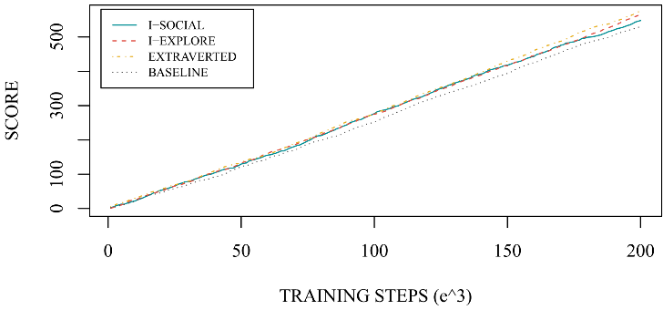


(a) Normal environment: even number of good and bad food

(b) Hostile enviroment: bad food sources 66.66%, good food

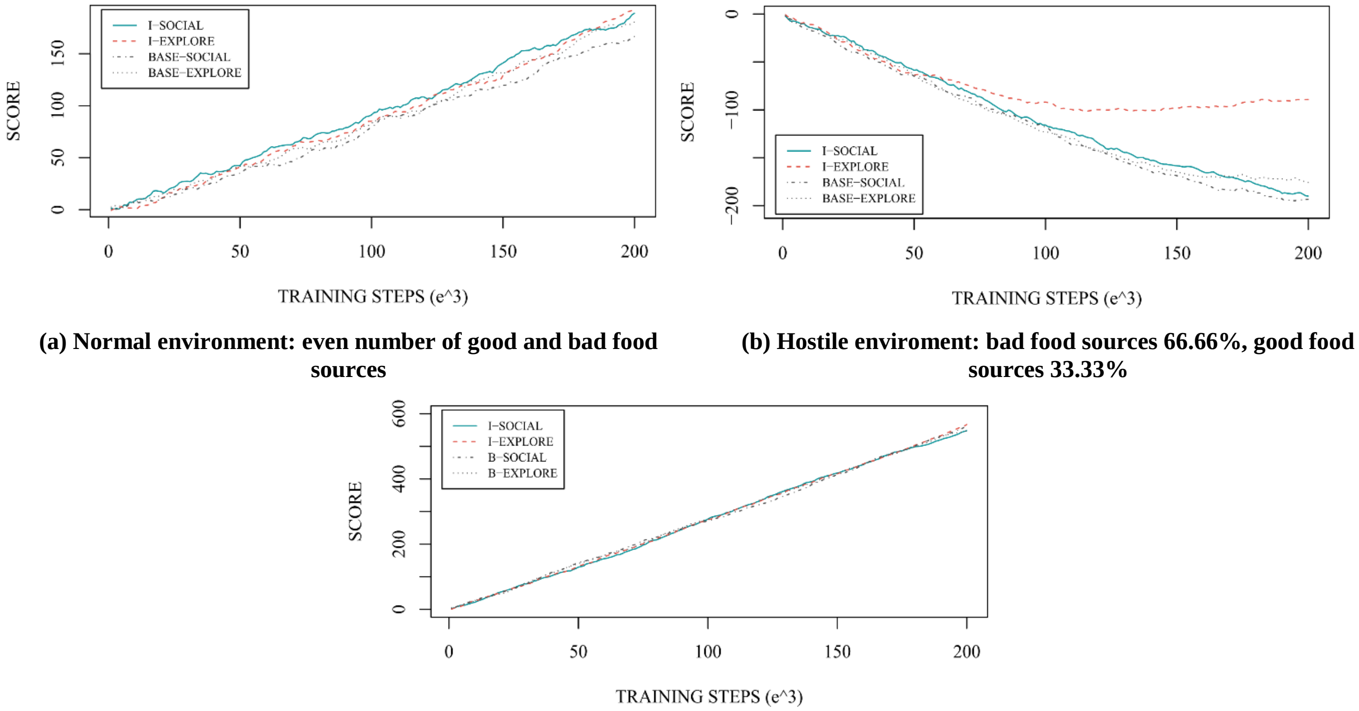
sources

sources 33.33%



(c) Benevolent environment: bad food sources 33.33%, good food sources 66.66%

Figure 3: Differences in average score/reward between agents with *ABERE* focus variations learning in a multi-agent enviroment over first 300K learning steps.



(a) Normal environment: even number of good and bad food sources

(b) Hostile enviroment: bad food sources 66.66%, good food sources 33.33%

(c) Benevolent environment: bad food sources 33.33%, good food sources 66.66%

Figure 4: Differences in average score/reward of agents with behaviour modulated by *ABERE* focus types (*I*) and behaviours induced by implicit modification of the reinforcement function (*B*).

5.2.1 Implicit vs. Explicit Goal Directed Behavior

In the next batch of experiments, we wanted to compare the difference between goal-oriented behavior that is modulated implicitly by *ABERE* and the behavior that was explicitly influenced by different reinforcement values. Two additional “baseline” agent types were defined that used only uniform sampling replay memory and differed only in their reinforcement functions. Baseline social agent was given double value of reinforcement for making a social contact relative to the food, while the baseline exploratory type had double reinforcement for food consumption. From Figure 4 we can see the difference in performance between attention-based approaches of modeling social and exploratory behaviors (I-SOCIAL,I-EXPLORE) and the baseline ones (BASE-SOCIAL,BASE-EXPLORE). It is evident that in the hostile environment the ABERE exploring agent is better suited to learn faster to avoid bad food, while in the other situations the performance of the different agents is comparable, which means that, at least for these experiments, attention-based replay memory gives the agents the possibility to successfully face different environments, without requiring any special design of the reinforcement function. In particular, in at least one combination, the ABERE agents where even able to perform better than the one with modified reinforcement function.

6. CONCLUSION AND FUTURE WORK

We presented a novel approach of replay memory sampling combined with Deep Q-learning called *Attention-Based Experience REplay*. Experimental results have shown that *ABERE* can outperform state of the art approaches on at least some of the environment variations, or have a similar performance. The

ABERE approach makes thus possible to define the focus of attention for an agent and have it performing well in different environments, without the need of re-designing the reinforcement function.

Being able to select the focal experiences by different criteria opens a lot of possibilities for modeling a stream of replay experiences that can potentially give rise to complex behavioral patterns. In future work, we will focus on changing the classification criteria taking into account properties other than interaction type such as other attributes of the agents state space. We also plan to work to model the focus of attention on the characteristics of the environment, so to be able to define a priori the most suitable focus for a given environment.

7. REFERENCES

[1] Lin, L.-J. *Reinforcement learning for robots using neural networks*. Fujitsu Laboratories Ltd, 1993.

[2] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[3] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K. and Ostrovski, G. Human-level control through deep reinforcement learning. *Nature*, 518, 7540 (2015), 529-533.

[4] Schaul, T., Quan, J., Antonoglou, I. and Silver, D. Prioritized experience replay. *arXiv preprint arXiv:1511.05952* (2015).

[5] Zhai, J., Liu, Q., Zhang, Z., Zhong, S., Zhu, H., Zhang, P. and Sun, C. *Deep q-learning with prioritized sampling*. Springer, 2016.

[6] Baddeley, A. D. and Hitch, G. Working memory. *Psychology of learning and motivation*, 8 (1974), 47-89.

[7] Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[8] Engle, R. W. Working memory capacity as executive attention. *Current directions in psychological science*, 11, 1 (2002), 19-23.

[9] Kasof, J. Creativity and breadth of attention. *Creativity Research Journal*, 10, 4 (1997), 303-315.

[10] Eysenck, M. *Attention and arousal: Cognition and performance*. Springer Science & Business Media, 2012.

[11] Lieberman, M. D. Introversion and working memory: Central executive differences. *Personality and Individual Differences*, 28, 3 (2000), 479-486.

[12] karpathy/reinforcejs: 2016. https://github.com/karpathy/reinforcejs. Accessed: 2016- 12-04.