Towards Learning Agents with Personality Traits: Modeling Openness to Experience

Mirza Ramicic
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
Email: mirza.ramicic@polimi.it

Andrea Bonarini
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milan, Italy
Email: andrea.bonarini@polimi.it

***Abstract*—Recent advances in neurosciences and cognitive sciences show us that the human neocortex is not a slave to the experiences from our perception and that the memories stored in hippocampus are goal weighted during the replay of the experiences for the purpose of re-learning from them. Since temporal difference reinforcement problems that use neural networks as function approximators rely on the experience replay memory structure that is similar to the hippocampus in our work we present a novel way of using a goal weighted prioritization of the memory that is biologically inspired.Furthermore we introduce a novel prioritization criteria called Variety of Experience Index or VEI for weighting the selection of the experiences that are stored in the replay memory. Weighting the experiences based on two different extremes of VEI can behaviouraly modify the agents learning process giving us two types of learning agents that exibit the different ends of the personality trait of Openness to Experience.**

I. Introduction

Advances in *Artificial Intelligence* have sparked an interest in developing systems that perform learn and think similar to human beings. Furthermore, modern machine learning algorithms are now more than ever taking inspiration from the physiology of human brain that is especially evident in approximation tecniques of Deep Q-learning (DQN) and its replay memory mechanism.

Recent discoveries in neurosciences and cognitive sciences [1] shows us that the human neocortex is not a slave to the stream of the experiences from the environment and that the center for memories in the mammalian brain, hippocampus is goal-depended weighted in replaying stored experience for re-learning.

In this work we intend to exploit these architectural similarities with human brain by modifiing the underlying mechanisms of Reinforcement Learning (RL) and therefore simulating the physiological differences in brains which account for emergence of different personality types in humans. As a reference model of human personality traits we have taken the most widely accepted topology: Five Factor Model (FFM) which organizes personality traits in terms of five basic dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness to Experience.

As the online agents learn they sample from the stream of experiences to create a policy π which maps their perception or state input to the possible actions, we have focused on a personality dimension that most closely relates to the human perception: Openness to Experience (O). Stream of experiences are not directly propagated to the learning mechanism in order to create *Temporal Difference* (TD) error that is used to update the *Artificial Neural Network* (ANN) function approximator.To reduce the temporal correlation between experiences and improve the speed of learning, a technique called *Experience Replay* [2], [3] is used to allow an agent to reuse past experiences, therefore obtaining a more stable training of the 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.

Since some transitions are more valuable for learning than others, especially in the early stages, prioritizing on experience transitions was introduced in order to improve the general performance of learning.

Successful approaches dealt with prioritized experience sampling [4] and experience replay [5], in order to improve the speed of learning but none has been made to use prioritization of experiences in order to actually change behaviour of the agents during the learning process. Since we are focusing on agents perception, instead of using *Temporal Difference* (TD) error as a prioritization criterion we are using specific properties of agents sensed state space given by relative Shannon's entropy of the two transitioning states s_t and s_{t+1}. Agents that are more Open to Experience will favor the experience transition that lead to the increase of the relative entropy between two states while the agents that are low on the scale will favor the transitions that reduce it.

II. Theoretical Background

A. The five factor model and Openness to Experience

Personality models describe the most important factors in which human individuals differ in their emotional, attitudinal, experiental and motivational styles. Throughout the history many theory candidates have been offered but at the beggining of the 1980 most of the researchers from many different traditions agreed that there are five basic factors or personality dimensions found in natural language, theoretically based questionares and in self-reports and ratings [6], [7]. The proposed five factor model organizes the personality traits in

five dimensions, but in our work we will be focused only on one that is most related to human cognition: Openness to Experience. Individuals that score higher in Openness to Experience scale have greater permeability of consciousness and perceptive cognition and are more motivated to seek variety and experience.

B. 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 [8]. For this reason the optimal action-value function $Q^*(s, a)$ is defined as the maximum expected return following the policy π:

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

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:

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

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.

C. 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 $Q(s, a; \Theta) \approx Q^*(s, a)$.

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

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 [3].

III. Prioritization based on the Variety of Experience

An agent performs the learning process on a single transition (s_t, a_t, r_t, s_{t+1}) by first predicting its previous estimate of the Q-value for being in a state s_t and taking an action a_t. This process performs a forward pass on the neural network approximator with s_t on input, after which we select the predicted Q-value on the output a_t. TD error represents the discrepancy between the previous estimate and the expected target Q-value after the transition which is given by its newly discovered reinforcement value r_t and the discounted maximum Q-value of the next state s_{t+1}. The learning process represents an update on the estimate of the function approximator by using backpropagation rule with perceived state space features s_t on the input and TD error difference on the a_t output.
Since the experiences are constantly stored in a sliding-window memory and replayed after each transition we can modify the selection criteria for the replay memory in order to focus more transitions that lead to a higher or lower Variety of Experience and therefore model a learning agent with high and low *Openness to Experience* scale score. We introduce *Variety of Experience Index* (VEI) that represents the tendency of the agent to gain higher variety in experience and we define it as a difference of Variety of Experience of the starting state s_t and the state that an agent has transitioned to s_{t+1}.

A. Quantifying the Variety of Experience of the states

In order to quantify the amount of uncertainty and possible information gain that a state space vector can carry we have applied Shannon's entropy as a measure of diversity, also called Shannon's index. The state space vector is represented by a number M of variables that are in most cases continuous and normalized in

$$0..1$$

. In order to measure the entropy, each of the M state space variables are discretized into N bins and calculated using Equation 4, where p_i is the frequency of values belonging to the ith bin.

$$H(s_t) = -\sum_{i=1}^{M} p_i log_2 p_i \tag{4}$$

B. Model Architecture and Learning Algorithm

Previous prioritization algorithms [4] used a stochastic sampling method that falls between uniform sampling and greedy sampling based on the TD error in order to make the learning faster and more efficient.
In our approach, instead of focusing on the TD error we introduce a prioritization based on the *Variety of Experience Index* (VEI) criterion, thus able to model the behavioural characteristics of agents performing the learning.
For the purpose of modeling agents that are exibiting the behaviours on the lower and higher end of the Openness to Experience axis we define VEI_L and VEI_H criteria respectively in the Equations 5 and 6.

$$VEI_L = H(s_t) - H(s_{t+1}) \tag{5}$$

$$VEI_H = H(s_{t+1}) - H(s_t) \tag{6}$$

From Equation 5 we can see that the prioritization index for the agents that are low in Openness to Experience is

higher when an agent is performing an action that transitions it from the state with the higher entropy $H(s_t)$ to the state with the lower entropy $H(s_{t+1})$.Respectively if we use the VEI_H criterion defined in 6 the agent will give more priority to the transitions that lead to the increase of entropy between the states while sampling them into the replay memory.

Instead of greedy sampling on VEI values which can make the system prone to over-fitting because of the lack of diversity [5], we define a stochastic prioritization based on the *Variety of Experience Index* VEI where the probability of sampling the $P(i)$ transition from the sliding window experience memory D is determined from Equation 7. VEI in this case represents the priority of the transition and the β parameter determines how much prioritization is used; in the uniform case $\beta = 0$.

$$P(i) = \frac{VEI_i^\beta}{\sum_{j=1}^{j=size(E)} VEI_j^\beta} \tag{7}$$

To alleviate the selection of the values for the β parameter, which would need to be tweaked for the specific application, we introduce a more general prioritization technique based on the descriptive statistical property of quartiles that can be used in a broader sense with no additional adjustments.
In order to sample basing on the VEI criterion, in Algorithm 1, instead of the stochastic approach given by Equation 7 we use a descriptive statistic approach which takes into account the upper interquartile mean of the data stream or the third quartile value ($Q3$) of the VEI values of agents experiences stored in a sliding window memory E of capacity n. This is computed by Equation 8.

$$VEI_{Q3} = \frac{3(n+1)}{4} thVEI \tag{8}$$

Given this, we sample only the transitions with VEI higher than the upper interquartile mean VEI_{Q3} of the entropy experience memory E as shown in Algorithm 1.

Algorithm 1 selectively stores the transitions after each update step based on VEI criterion. The criterion stores the transitions that have the *Variety of Experience Index* VEI value higher than the upper interquartile mean of the n latest VEI samples from E given by the $VEI > VEI_{Q3}$ conditional.
After each transition a random batch of the previous transitions is selected from the replay memory D in order to perform additional training on the approximator.

Algorithm 1 Deep Q-learning with VEI prioritizations

Initialize replay memory D with capacity N and VEI experience memory E
Initialize action-value function Q with random weights and agent type $T_a = (L, H)$
for episode = 1, M **do**
  for t = 1, T **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \arg\max_a Q^*(s_t, a; \Theta)$
    Execute action a_t, observe reward r_t and state s_{t+1}
    if $T_a = L$ **then**
      Calculate the transition value VEI based on Equation 5 and add it to the sliding window memory E
    end if
    if $T_a = H$ **then**
      Calculate the transition value VEI based on Equation 6 and add it to the sliding window memory E
    end if
    Calculate upper interquartile mean VEI_{Q3} of the last n samples from E using Equation 8
    if $VEI > VEI_{Q3}$ **then**
      Store transition (s_t, a_t, r_t, s_{t+1}) in D
    end if
    Sample random batch of transitions (s_t, a_t, r_t, s_{t+1}) from D

$$\text{set } y_i = \begin{cases} r_i, & \text{terminal } s_{i+1} \\ r_i + \gamma \max_{a'} Q(s_{i+1}, a'; \Theta), & \text{non terminal} \end{cases}$$

    Perform a gradient descent step on $(y_i - Q(s_i, a_i; \Theta))^2$ according to Equation 3
  end for
end for

IV. Experimental setup

To evaluate the proposed model we have applied the algorithm to two different enviroments: *Waterworld* and *Puckworld* from ReinforceJs framework [9]. The first, more complex simulation, *Waterworld* represents an environment with moving good and bad food pieces. Food pieces are generated at a random position with 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 an agent is to consume as much good food pieces as possible, 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 position, speed, and direction, thus keeping the distribution of food constant. Agents receive reinforcement +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 30 directional sensors and each of them can perceive 5 continuous variables: distance of sensed object (good food, bad food, wall), the first two of which have the two additional attributes: speed in x direction and speed in y direction; this gives a total of 150 state space inputs for each agent.
Second simulation: *Puckworld* consists of a much simpler environment with two points that are changing positions. Agent wants to stay as close to the good (green) point while avoiding bad (red) point as conditioned by the proportional

reinforcement value. The good (green) point is static and it's instantiated at a random position, but the bad (red) point is constantly moving in a random direction which gives it a tendency to be found in states with higher entropy values. The state space is smaller but still continuous taking into account agents own location and speed and the locations and speed of both green and red points which gives a total of 12 continuous variables. The reward is based on the agents distance to the green point (lower is better). However, if the agent is in the vicinity of the moving red target it gets a negative reward proportional to its distance from the red point.

As a function approximator for both simulations we are using a deep neural network with weights Θ to approximate $Q(s, a; \Theta) \approx Q^*(s, a)$. To reduce the computational complexity of having multiple forward steps each time, we want to find an action that maximizes the state-action function $\arg\max_a Q(s, a)$; the network takes the state vector s as an input and predicts $Q(s, a)$ for each possible action.

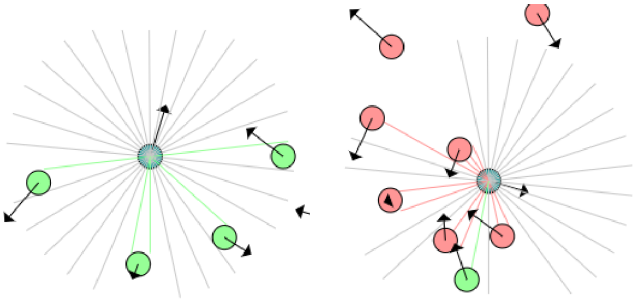
We have adopted the original Q-learning update fourmula with a learning rate α set to a low value (0.05) because of the nature of the approximator, and discount factor $\gamma = 0.9$. The default capacity of the replay memory buffer D included 7000 experiences and the entropy experience memory capacity n was set to 500.

A. *Variety of Experience criterion comparisons*

In order to evaluate how does *Variety of Experience Index* of the transition VEI relate to the behavioural characteristics of the *Openness to Experience* personality trait we are comparing transition examples from the *Waterworld* experimental setup with highest values of *Variety of Experience Index* between two different agent types: VEI_L and VEI_H. For the purpose of evaluation each of the detected objects and agent is depicted with its speed vector that represents the composition of its x and y speed components as described in the state space.

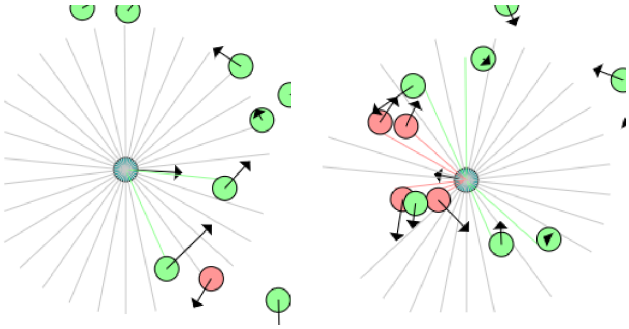
Figure 1 showcases some of the transitions with high value of VEI_L which represents the prioritization criterion for the type of agent that is associated with low level of *Openness to Experience*. These agents favorize the transitions which have entropy values of the starting state s_t higher than of the end state s_{t+1}. This is behavioraly evident in the transitions shown in Figures 1a and 1b where we can see that an agent has the tendency to move away from the experience that is represented by the moving food clusters.

On the other end of the VEI spectrum from Figures 2a and 2b we can see agents transitions that have the tendency to move toward the experience since the entropy values of the end state s_{t+1} are higher than the first s_t. These transitions are prioritized in the agent type that is high on the *Openness to Experience* scale or VEI_H.



(a) $VEI_L = 0,1528$ (b) $VEI_L = 0,2083$

Fig. 1: Transitions with high VEI_L values



(a) $VEI_H = 0,2872$ (b) $VEI_H = 0,3064$

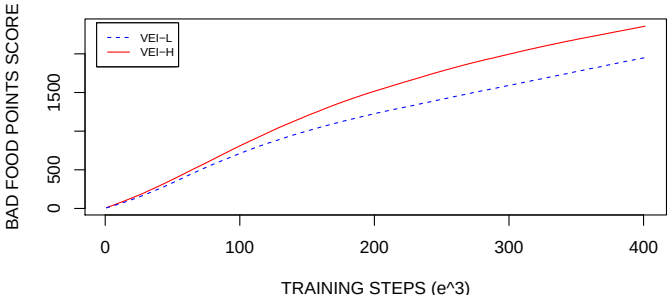
Fig. 2: Transitions with high VEI_H values

V. Experimental results

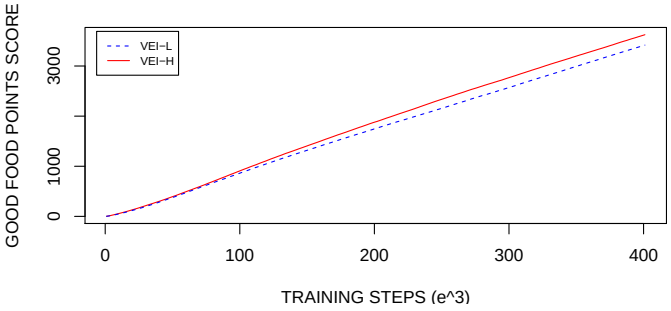
In the experiments, we have compared two types of prioritized sampling algorithms with two different prioritization criteria: VEI_L and VEI_H associated with agents that are on the lower and higher end of *Openness to Experience* axis respectively. Figure 3 shows the comparison between the two different prioritization criteria applied to more complex *Waterworld* experimental setup; We can see that the two agent types act behavioraly differerent from both bad total score shown in Figure 3a and good total score shown in Figure 3b. Agents that are high on the *Openness to Experience* scale marked by VEI_H score more on both good and bad food points thus demonstrating the behaviour caracteristics of taking more risk due to the tendency of moving towards area with more experience, in this case food sources. Agents that are lower in *Openness to Experience* VEI_L behave differently from their counterparts by scoring much lower values of bad food sources but also lower values of good food sources making them behave in a more cautious way. They personality trait of lower *Openness to Experience* gives them advantage in this setup as their overall score is higher then the agents that are conditioned higher on the trait.

Figure 4 shows the similiar comparison, the difference being only the Y axis of the graph which shows the average distance to the good and bad points instead of the score. From Figure 4b we can see that the agents with lower *Openness to Experience* score much better by being more close to the good point in average than the agents that are high in the same trait but also score less efficiently in the Figure 4a by being also more close to the bad point.In *Puckworld* environment states with higher entropy contain moving bad red point while the lower ones contain usually just the static good green one, therefore

explaining the better score of the VEI_H on the distance from the bad moving target.



(a) Total score of bad food sources



(b) Total score of good food sources

Fig. 3: Comparison between average reward values of VEI_L and VEI_H types of agents for good and bad food sources in 20 learning epochs of *Waterworld* simulation environment, over first 400K learning steps

VI. Conclusion

We presented a novel approach of behaviourally modifyng the characteristics of learning agents just by favorizing on a specific types of experiences during the sampling into the replay memory.This technique proved to be an efficient way of exibiting a specific personality trait in a learning agent without modifying any other properties of the algorithm or reinforcement function. The novelty of the approach emphasizes the use of replay memory in a biologically inspired goal oriented approach.

References

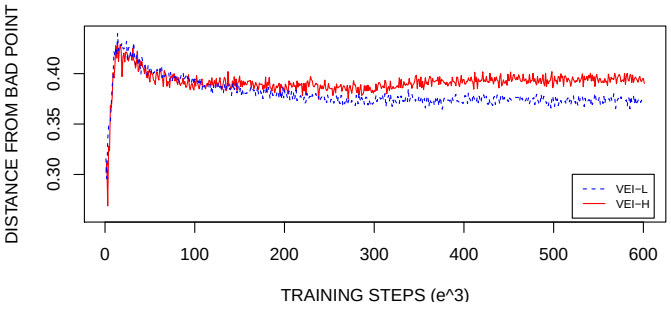
[1] D. Kumaran, D. Hassabis, and J. L. McClelland, "What learning systems do intelligent agents need? complementary learning systems theory updated," *Trends in cognitive sciences*, vol. 20, no. 7, pp. 512–534, 2016.

[2] L.-J. Lin, "Reinforcement learning for robots using neural networks," DTIC Document, Tech. Rep., 1993.

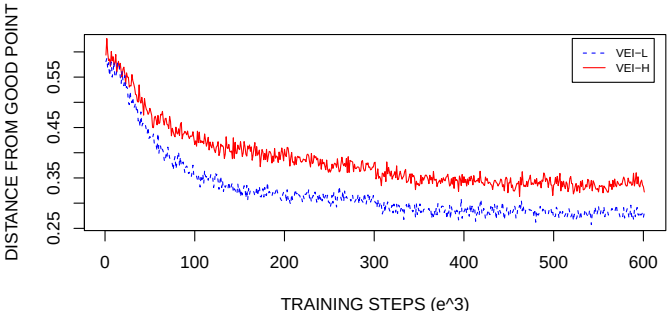
[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[4] J. Zhai, Q. Liu, Z. Zhang, S. Zhong, H. Zhu, P. Zhang, and C. Sun, "Deep q-learning with prioritized sampling," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 13–22.

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



(a) Total score of bad food sources



(b) Total score of good food sources

Fig. 4: Comparison between average distance values of VEI_L and VEI_H types of agents from good and bad targets in 20 learning epochs of *Puckworld* simulation environment, over first 600K learning steps

[6] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[7] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[9] Karpathy, "Reinforcejs framework," https://github.com/karpathy/reinforcejs, 2013, accessed: 2016-12-04.