Curiosity: Models & Applications - Project

Part 3

Submitted by:

Lior Sukman	319124244
Olga Soldatenko	342480308
Yana Dervin	310684386
Yuval Goldshtein	311552368

Formalization of the Problem as a Curiosity Problem

The purpose of the curiosity algorithm is to learn which parts of the data will be the most beneficial for the learning process of the noise-generating neural network (NN) which is the learner. In other words, the goal is to learn a policy determining how to choose the pictures for the train set for the NN.
We have decided to check whether pictures of some specific labels contribute more to the learning than of other labels, and in which order these labels are best to be revealed to the learner.

1. States and Actions

This problem is analogous to the curious feature selection (Moran & Gordon, 2019), in which the curious algorithm learns which features are the most effective for the learning, only in our case the "features" to be selected are the labels.
The state space is $s \in S: \left\{s_0, l_0, l_1, ..., l_9\right\}$, where s_0 is the initial state (a state before any label is selected) and l_i corresponds to the label i being previously chosen.
The actions are the labels that we are adding to the train set $a \in A: \left\{a_0, a_1, ..., a_9\right\}$where a_i corresponds to adding the data of label i to the train set.

2. Reward

The reward of the network is based on the change of the error of the learner. The initial error is 0.985 based on the success rate of the pre-trained classifying convolutional neural network (will simply be referred to as the classifying CNN). The error is the fraction of the data correctly classified by the classifying CNN after the perturbations made by the learner were applied.
Thus, given that the error at time t is denoted as e_t, and the initial error being $e_0 = 0.985$, the reward at time $t \geq 1$ will be $r_t = e_{t-1} - e_t$ corresponding to the improvement.

3. Closing the curiosity loop

The algorithm starts with the initialization of the Q matrix we wish to learn with values of 1, this creates a greedy policy with respect to the Q-values for the first few episodes. In addition, as described in the previous subsection, $e_0$is initialized as $e_0 = 0.985$. During each episode a set of constant size is sampled and divided into training and evaluation sets (80% and 20% of the episode respectively)[1].
On each time step a new action is chosen based on the previous state. According to the chosen action, the data from the training part of the episode sample with the corresponding label is added to the data the model is trained on. After training, the

[1] Any sampling and splitting of the data in the training was balanced, meaning that the number of examples of each class was the same for all the classes. In addition, the ratio of each class was maintained when dividing the episode data to train and evaluation sets.

model is evaluated on the entire evaluation set from the episode data. Based on this evaluation an error e_t and the reward r_t are calculated. Then, using the r_t the Q matrix is updated according to Bellman's rule of optimality.
Higher reward is given when the learner reduces the error more dramatically, and lower (negative) reward is given when the error of the learner increases. This reward system yields a model that is able to choose the most informative data to train the learner on and benefits the most when a drastic improvement follows an action.

Programming

1. Running the algorithm

The code files are submitted with this report.

2. Results

The curiosity loop was run for 1000 episodes (where each episode contained 600 examples which is 1% of the entire train set) in order to go over all of the states several times.
During the learning process both decrease in error and in loss were achieved (see **figure 1** and **figure 2**). Error was calculated as the accuracy of the classifying CNN on the evaluation set, and similarly, loss was calculated as the inverse negative log-likelihood (NLL) of the classifying CNN's output on the perturbed images. In both cases, these values were averaged across all steps of the episode.

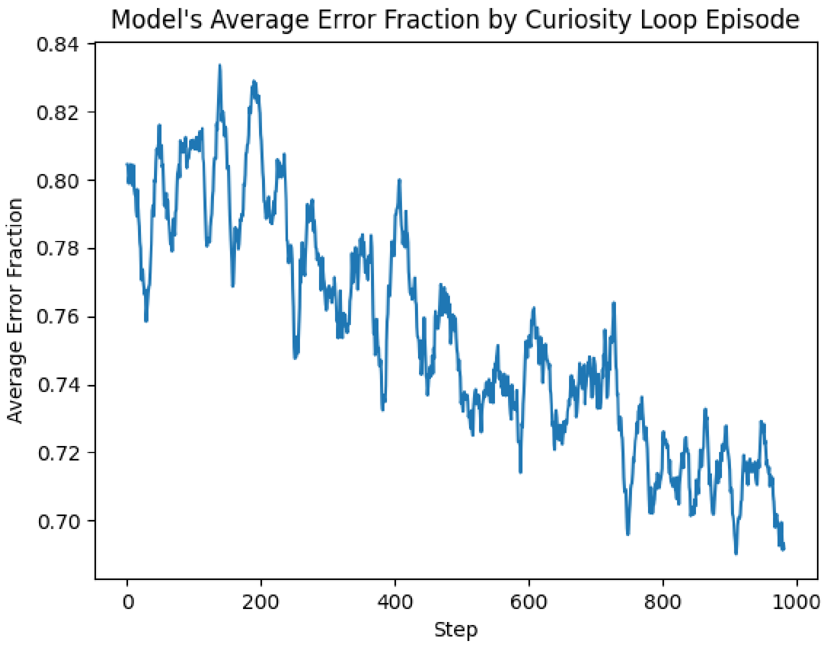


Figure 1: Average error over each episode. Average error is measured as the accuracy of the classifying CNN averaged over all steps within an episode. This graph shows a moving average of the average error with a window size of 20.

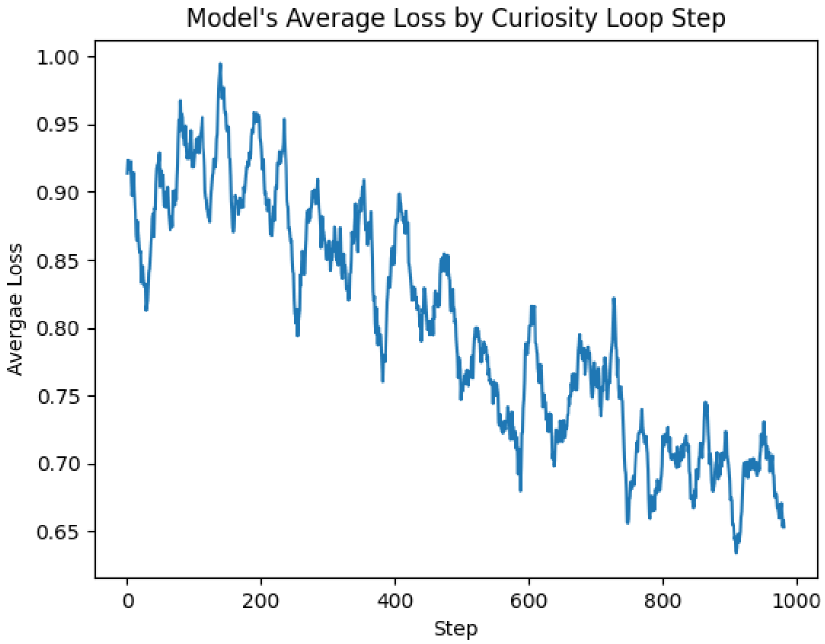
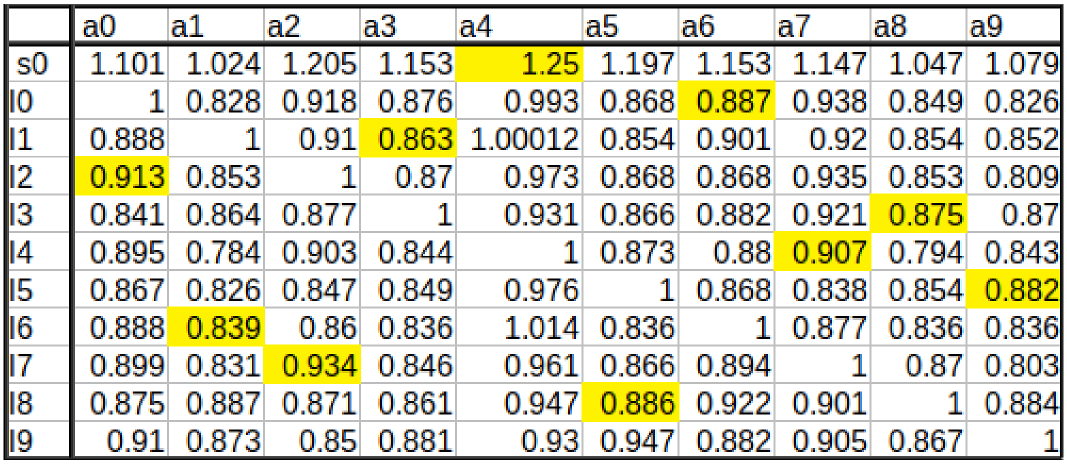


Figure 2: Average loss over each episode. Average loss is measured as the inverse NLL of the classifying CNN averaged over all steps within an episode. This graph shows a moving average of the average loss with a window size of 20.

The Q-table that was obtained during the learning process is presented in **table 1**. We can see that all of the values on the s_0 are larger than the initial value of 1 , which means that all of the labels can contribute to the learning of the noise-generating network. Yet, some values for states other than s_0 are lower than the initial value of 1, which indicates low to negative decrease in error[2]. As expected, the diagonal contains the initial value of 1 as each action can only be performed once.

According to the results, the optimal policy is $l_4 \rightarrow l_7 \rightarrow l_2 \rightarrow l_0 \rightarrow l_6 \rightarrow l_1 \rightarrow l_3 \rightarrow l_8 \rightarrow l_5 \rightarrow l_9$.

[2] note that the initialization results in a somewhat arbitrary scale and as will later be shown, some actions with values lower than 1 still result in improvement



	a0	a1	a2	a3	a4	a5	a6	a7	a8	a9
s0	1.101	1.024	1.205	1.153	1.25	1.197	1.153	1.147	1.047	1.079
l0	1	0.828	0.918	0.876	0.993	0.868	0.887	0.938	0.849	0.826
l1	0.888	1	0.91	0.863	1.00012	0.854	0.901	0.92	0.854	0.852
l2	0.913	0.853	1	0.87	0.973	0.868	0.868	0.935	0.853	0.809
l3	0.841	0.864	0.877	1	0.931	0.866	0.882	0.921	0.875	0.87
l4	0.895	0.784	0.903	0.844	1	0.873	0.88	0.907	0.794	0.843
l5	0.867	0.826	0.847	0.849	0.976	1	0.868	0.838	0.854	0.882
l6	0.888	0.839	0.86	0.836	1.014	0.836	1	0.877	0.836	0.836
l7	0.899	0.831	0.934	0.846	0.961	0.866	0.894	1	0.87	0.803
l8	0.875	0.887	0.871	0.861	0.947	0.886	0.922	0.901	1	0.884
l9	0.91	0.873	0.85	0.881	0.93	0.947	0.882	0.905	0.867	1

Table 1: The Q-table, marked yellow the actions chosen at each state (i.e. maximal Q-value).

We have compared the acquired policy to the random one (each action is chosen randomly from the available actions). In this section we partitioned the training data to a train set of size 50,000 and an evaluation set of size 10,000. At each iteration we trained the learner on the training set and chose the best model out of 50 epochs using the evaluation set[3]. Following this part, the model was evaluated on a test set of 10,000 examples.

Figure 3 shows the graph of error on the test set at every policy step, averaged across 10 trials (note that the errors were calculated as discussed before). As expected, after the last step the learners reach similar error rates. In addition, we can see that for the first two steps the computed policy is preferable, although further on, the policy does not have an advantage over the random choice and even performs worse.

This may be due to the fact that the model does not take into consideration the history, but only the previous state. It is possible that using a more complex model with recurrent layers (e.g. LSTM or GRUs) could solve this problem and show superior results, although such a model will require more computing power and additional hyperparameters and so is out of scope for this project.

[3] This process was done to assure best generalizability.

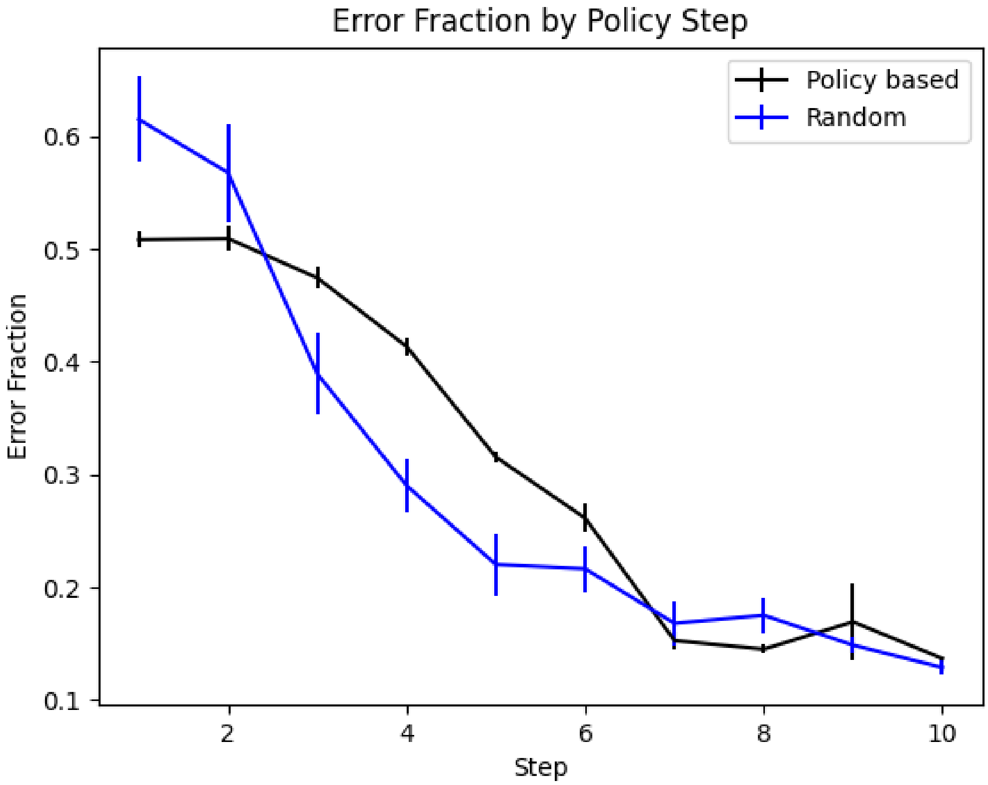


Figure 3:The error fraction against step (at each step the new label is added to the training set). Each point represents the average over 10 trials and is shown with standard error bars.

It is interesting to see that in both cases (policy based and random) after only a single step (i.e. training only on one class) the error decreased significantly (and much more than that class presence in the data). This suggests a common mechanism for the creation of perturbations for all classes. This is in line with previous research in this field (e.g. the fast gradient sign method that was presented in the first part of the project). Nevertheless, according to the Q-table presented in **table 1**, we do see variability between different classes that indicates that some classes are more beneficial for the learning process. It is reasonable to assume that this is a result of the average value of pixels (more signal may allow better learning) but this stands at odds with the fact that for $s_0$both 1 and 8 have similar Q value (1.024 and 1.047 respectively) yet they are represented by different signal intensity[4]. Hence, we assume the differences in the informative value of different examples are more complicated and will require further research. It is worth emphasizing that according to this graph most classes had some informative value as the graph is approximately monotonic[5] indicating that complete training is better when possible,

[4] Note that this observation is intuitive (i.e. higher signal for 8 compared with 1) yet it was checked and confirmed.

[5] note that some increase in error is present after specific actions and according to the Q-table some actions when followed by others do not reduce error rate or only reduce it minorly.

yet it is not clear whether additional examples of a specific class will be proven better than the addition of an unseen class.

Bonus - Mapping the Problem to the Brain and to Psychology[6]

First, we will discuss the mapping of the agent's role to the brain and psychology. From the neuroscience perspective it is reasonable to assume that the dopaminergic system will take part in the learning process as the problem contains some sort of a reward in the form of reduced error and improved performance over time. This means that ventral tegmental area (VTA) and substantia nigra (SN) complex will be active. These areas were shown to be active during question presentation (under assumption of a curious state), and thus we expect them to be active prior to the action selection. We also expect before an action is taken higher activity in the nucleus accumbens which is activated by the dopaminergic signals from the VTA/SN complex. Another expected observation is some activity of the hippocampus (HC) while encoding the new information (around the time of action selection and receival of reward), yet we note that the HC is commonly related to long term memory and looking at the learning as a short time process, consolidation of the memory is not required and hence the role of the HC in the context of encoding is minor. That said, the HC is also related to novelty detection and was shown to have effect in exploratory behavior (Voss et al., 2011).
In addition, after reward receival, we expect some sort of prediction error to be calculated, this was shown to be related to an activation in the ventral and dorsal striatum (Delgado et al., 2000; Knutson et al., 2000; Pagnoni et al., 2002; O'Doherty et al., 2003, 2004; McClure et al., 2003, 2004; Rodriguez et al., 2006), and was shown not only to be mediated by dopaminergic activity, but also is correlated with successful learning (Schonberg et al. 2007). Calculation of prediction errorwas also shown to cause activation in more nuclei of the basal ganglia including the globus pallidus (Joel, Niv & Rupin, 2002). Based on the review by Rushworth et al. (2011), the frontal cortex is engaged in reward-related learning and decision making in this context, specifically, the areas with an important role are the lateral orbitofrontal cortex, the ventromedial prefrontal cortex and adjacent medial orbitofrontal cortex, anterior cingulate cortex, and the anterior lateral prefrontal cortex. The functions of these areas are diverse and include expectations of loss and gain, value evaluation, comparison between the values of different options, creating action-reward association and exploratory behavior.

From a psychological point of view, the problem can be described with operant learning terminology. In operant learning, the learner faces a stimulus (or state) S, with possible response (or action) R and some reward O. According to Thorndike's law of effect (1911), if following a response R in the presence of some stimulus S there is a satisfying event, the association between the stimulus and the response would increase (i.e. the S-R association). According to this logic, given a state S the learner will respond with response R that has the highest S-R association of all possible responses (action selection). In addition, R-O associations are also created, indicating some expectation for specific reward following a response, similar to the aforementioned information from the neuroscience perspective.

[6] Any unreferenced information that is presented here is based on the lecture of Matthias Gruber.

Apart from that, it is possible to look at the problem under the scope of problem solving as a probabilistic problem in order to understand the possible strategies[7]. In probabilistic learning the learner faces various options (in our case possible actions), each one will yield some reward in a probabilistic manner unknown to her (in our case this is the error rate, which is probabilistic due to random initialization of the networks).

In probabilistic learning, there are three noteworthy strategies. The first is maximization, in this approach, after the first few responses, the learner stabilizes on a specific response for the remainder of the session. Another strategy is alteration, in which the response is changed frequently in order to cover as many possibilities as possible. The third hypothesis is testing, in which the probabilistic principle is denied by the learner and she aims at finding the rules behind the task.

Each of these strategies can be applied by a human learner in the process (balance between the maximization and alteration strategies are implemented in many reinforcement learning algorithms as the exploration-exploitation tradeoff), yet it is clear that if each of the first two strategies would be used exclusively the results are likely to be relatively inferior. A curious human learner may lean towards the hypothesis testing strategy, aiming at finding the rules behind the problem. In this case both responses that were found beneficial will be repeated but also new ones and new combinations of responses.

Another learning process to discuss is the learner's task (i.e. the noise generation). In this case we will explain why a human learner will most likely fail the task. In this case, the learner's input is an image of a handwritten digit. Visual input reaches the primary visual cortex (V1) of the brain in the occipital lobe. From V1 the information progresses and undergoes further processing, it then continues through two pathways (Goodale & Milner, 1992), the ventral (what) pathway and the dorsal (where/how) pathway. The processing in the brain still surpasses in many fields the performance of state of the art artificial neural networks, thanks to its complex structure and holistic processing[8]. While we have shown throughout our project that it is possible to reduce a trained network's accuracy to chance level by adding generated noise, we have also shown multiple examples demonstrating the fact that these perturbations minorly affect human perception. In light of these arguments, we assume this problem can not be solved efficiently by a human. It is worth emphasizing that there could be different formalizations of the problem and/or parallel problems which may be more suitable given the human's brain processing, yet they might be less intuitive.

[7] The information in this paragraph is based on the course: "Cognitive Psychology: Problem Solving and Creativity" which is passed in the psychology department of Tel Aviv University.

[8] You may refer to gestalt theory for some rules (although not complete in the context of holistic processing).

References

Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA (2000) Tracking the hemodynamic responses to reward and punishment in the striatum. *J Neurophysiol* 84:3072–3077.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1), 20-25.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks*, *15*(4-6), 535-547.

Knutson B, Westdorp A, Kaiser E, Hommer D (2000) FMRI visualization of brain activity during a monetary incentive delay task. Neuroimage 12:20 –27

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38:339 –346.

McClure SM, York MK, Montague PR (2004) The neural substrates of reward processing in humans: the modern role of FMRI. *Neuroscientist* 10:260 –268.

Moran M, Gordon G. “Curious Feature Selection.” Information Sciences, vol. 485, 2019, pp. 42-54.

O’Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304:452–454.

O’Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38:329 –337.

Pagnoni G, Zink CF, Montague PR, Berns GS (2002) Activity in human ventral striatum locked to errors of reward prediction. *Nat Neurosci* 5:97–98.

Rodriguez PF, Aron AR, Poldrack RA (2006) Ventral-striatum/nucleus accumbens sensitivity to prediction errors during classification learning. *Hum Brain Mapp* 27:306 –313.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, *70*(6), 1054-1069.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860-12867.

Voss, J. L., Gonsalves, B. D., Federmeier, K. D., Tranel, D., & Cohen, N. J. (2011). Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. *Nature neuroscience*, 14(1), 115-120.