

Deep Learning with Experience Ranking Convolutional Neural Network for Robot Manipulator

Hai Nguyen, Hung Manh La, and Matthew Deans

Abstract—Supervised learning, more specifically Convolutional Neural Networks (CNN), has surpassed human ability in some visual recognition tasks such as detection of traffic signs, faces and handwritten numbers. On the other hand, even state-of-the-art reinforcement learning (RL) methods have difficulties in environments with sparse and binary rewards. They require manually shaping reward functions, which might be challenging to come up with. These tasks, however, are trivial to human. One of the reasons that human are better learners in these tasks is that we are embedded with much prior knowledge of the world. These knowledge might be either embedded in our genes or learned from imitation - a type of supervised learning. For that reason, the best way to narrow the gap between machine and human learning ability should be to mimic how we learn so well in various tasks by a combination of RL and supervised learning. Our method, which integrates Deep Deterministic Policy Gradients and Hindsight Experience Replay (RL method specifically dealing with sparse rewards) with an experience ranking CNN, provides a significant speedup over the learning curve on simulated robotics tasks. Experience ranking allows high-reward transitions to be replayed more frequently, and therefore help learn more efficiently. Our proposed approach can also speed up learning in any other tasks that provide additional information for experience ranking.

I. INTRODUCTION

Reinforcement Learning (RL) methods have been applied to a variety of real-world robotics problems ranging from locomotion [1] to manipulation [2], [3]. However, one of the main challenges that limits RL's full potential for robotics applications is how to learn efficiently from sparse and binary rewards, which is very common in many robotics tasks. It is challenging as in most learning attempts, by exploration, the agents will mostly end up learning nothing from zero rewards, and they are clueless about what to improve next time. One of the best methods that we have had so far to tackle the problem is using Hindsight Experience Replay (HER) [4] in combination with an off-policy RL methods e.g., Deep Deterministic Policy Gradient (DDPG) [5], Normalized Advantage Functions (NAF) [6]. By modifying the original goal, HER allows the agent to even learn from unsuccessful transitions.

One problem with HER is that it used a hard-coded strategy (*final* or *future* strategy) [7] for randomly choosing

This material is based upon work supported by the National Aeronautics and Space Administration (NASA) Grant No. NNX15AI02H issued through the NVSGC Research Infrastructure and the NVSGC Faculty Award - CD.

Hai Nguyen and Dr. Hung La are with the Advanced Robotics and Automation (ARA) Laboratory, Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557, USA. Dr. Matthew Deans is with NASA Ames Research Center, Moffett Field, CA 94035. Corresponding author: Hung La, email: hla@unr.edu

additional goals and episodes to replay. This lack of an evaluation scheme makes it possible that the chosen transitions are poor experiences, which might slow the learning process. To solve this problem and respond to OpenAI's request for research for an automatic hindsight goal generation for HER mentioned in [7], we propose a novel idea of using an external CNN network to evaluate the training experience. Based on their ranking scores, we will decide which episodes should be restored and replayed for training.

We verified that the proposed ranking scheme can improve significantly the learning curve for a mobile robot manipulator to learn four different tasks in robotics environment of Gym [8]. Our proposed approach will also be useful for any other robotics learning problems that can provide additional information for experience ranking such as images.

II. RELATED WORK

In biological systems, experience replay has played an important role for learning behaviors. Various studies show that there is frequent experience replay in the hippocampus of rodents of both awake and sleeping animals. Research also shows that disrupting experience replay might lead to impairing spatial memory [9] [10], and some show that experience replay can contribute to better memory consolidation and retrieval [11], both are instrumental in boosting learning ability.

In machine learning, the idea of using *Experience Replay* was first introduced in [12] when it is used to speed up learning by repeating past experiences. Experience replay helps break the temporal correlation by mixing old and new experiences, and frequent experiences therefore might be replayed more often. Since then, experience replay has been widely adapted, and it becomes the norm in many success in RL research [13] [14].

However, the current way to use experience replay is to uniformly sample from first-in-first-out (FIFO) buffers of experiences [5] [14]. By doing this way, the importance of each transition is neglected, and we might replay experience that is not very helpful for the learning at all. When faced with the design choice of which experiences to restore, neuroscience research show that surprising [15] and rewarding experiences are more preferred [16].

In RL, attempts to prioritized experiences have been made in [17] where temporal difference (TD) error is used as a way to measure the priority. Our approach is reward-favored and more direct. We will look at the state of the training episode and use a pre-trained neural network to rank that episode and decide whether the associated experience is worth being stored for future learning.

III. BACKGROUND

A. Reinforcement Learning Formalism

Considering the standard reinforcement learning setting when an agent interacts with an environment. The agent continually makes decisions, and the environment responds to the chosen actions and bring agents to new states. At each step, the agent receives the current environment's state $s_t \in \mathcal{S}$ and bases on the current policy π to select an action $a_t \in \mathcal{A}$. The next time step, the agent receives a numerical reward $R_{t+1} \in \mathbb{R}$ with a reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and itself being in another state S_{t+1} . The goal of the agent is to maximize the reward over time through learning a policy, which can help to choose actions that lead to maximizing the accumulated reward. The dynamics of the environment is defined by the state-transition probability $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$.

A deterministic policy will map states to actions: $\pi : \mathcal{S} \rightarrow \mathcal{A}$. Every episode starts with sampling an initial state s_0 , and at every timestep t the agent selects an action based on the current policy and its current state: $a_t = \pi(s_t)$. Then it receives the reward $r_t = r(s_t, a_t)$, and the environment's new state is sampled from the distribution $p(\cdot | s_t, a_t)$.

A discounted sum of future rewards is $R_t = \sum_{i=t}^{\infty} \gamma^{i-t} r_i$ with a discount rate $\gamma \in [0, 1]$, and the agent's goal is to maximize its expected return starting from the beginning $\mathbb{E}_{s_0}[R_0 | s_0]$.

The action-value function describing the value of taking action a in state s under the policy π is defined as $Q^\pi(s, a) = \mathbb{E}[R_t | s_t = s, a_t = a]$. An optimal policy π^* i.e. any policy π^* s.t. $Q^{\pi^*}(s, a) \geq Q^\pi(s, a)$ for every $s \in \mathcal{S}, a \in \mathcal{A}$, and any policy π^* share the same *optimal Q-function* denoted as Q^* . It satisfies the *Bellman equation*:

$$Q^*(s, a) = \mathbb{E}[R(s, a) + \gamma \max Q^*(s', a')].$$

B. Deep Q-Networks (DQN)

DQN [14] is a model-free RL algorithm for discrete action spaces. In DQN, neural networks are used to approximate the action-value function $Q(s, a)$. Two important features of DQN are *experience replay* and *target network*.

Experience replay is used to store previously transitions, which are defined as tuples (s_t, a_t, r_t, s_{t+1}) where s_t is the current state, a_t is the action, r_t is the reward, and s_{t+1} is the next state. Tuples will then be added to a *replay buffer* from which batch of transitions will be sampled for training. The usage of experience replay provides uncorrelated samples for training and also improve data efficiency [18].

The network is then trained using mini-batch gradient descent on the loss $\mathcal{L} = \mathbb{E}(Q(s_t, a_t) - y_t)^2$, where $y_t = r_t + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a')$, with the tuple (s_t, a_t, r_t, s_{t+1}) sampled from the replay buffer.

In order to make this optimization procedure stabler, the target y_t is usually computed using a separate *target network*, which changes at a slower pace than the main network. After that, the weights of the *target network* will be periodically updated to the main network to increase the stability.

C. Deep Deterministic Policy Gradient - DDPG

Although DQN has achieved good performance in higher dimensional problems such as Atari games [13], it only works with discrete action spaces. DDPG [5] is developed based on DQN, and it is a model-free off-policy RL algorithm for continuous action spaces.

DDPG belongs to the actor-critic structure where there are two neural networks: an *actor* $\pi : \mathcal{S} \rightarrow \mathcal{A}$ and a *critic* $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The critic's job is to evaluate the current policy by estimating the Q-value from the current state and the action outputted from the actor.

Firstly, action is generated by the target policy with added noise for action exploration, e.g., $a_t = \pi(s) + \mathcal{N}_t$. The critic is trained in a similar way as the Q-function in DQN with the target y_t being computed using actions produced by the actor, e.g., $y_t = r_t + \gamma Q(s_{t+1}, \pi(s_{t+1}))$. The actor is trained using mini-batch gradient descent applied on the loss $\mathcal{L} = -\mathbb{E}_s Q(s, \pi(s))$.

D. Hindsight Experience Replay - HER

The idea behind HER is to mimic human ability to learn from failures. HER allows learning from all episodes, even if in those episodes the agent did not achieve the original goal. Instead, HER considers the final state that the agent finally reached to be a modified goal. While in the standard experience replay, only the transition $(s_t || g, a_t, r_t, s_{t+1} || g)$ with the original goal g is stored, HER also keeps the transition $(s_t || g', a_t, r', s_{t+1} || g')$ with modified goal g' . By adding additional goals, the RL algorithm can still have a learning outcome because it can learn to achieve a goal even that goal is not the original target.

In [4], they experimented with different strategies for choosing goals to use with HER: *final* strategy - goals corresponding to the final state of the environment and *future*: k random states which come from the same episode as the transition being replayed and were observed *after* it. However, both strategies neglect the importance of each episode, and there is actually no mechanism to measure.

IV. DDPG + HER AND EXPERIENCE RANKING CNN

A. Algorithm

The idea is to improve the sampling of additional goals of HER by introducing new information - ranks of new trajectories as shown in Figure 1. Additional goals selected by HER will then contain only episodes that have high rewards. We do not account for the loss of diversity here because HER itself allows a number of percentage of normal episodes to pass through [4] by specifying a hyper-parameter k controlling the ratio between HER replays and regular replays (e.g. $k = 4 - 4$ times). The output of HER selection algorithm will include both ranked experiences and normal ones.

The modification to DDPG + HER is described in Algorithm 1. In the algorithm, we change HER's random sampling regardless of importance of transitions by using experience ranking to only allow episodes in which the agent achieve high reward to be stored. If the rank of the episode is low, the

Algorithm 1 DDPG + HER modified with ER-CNN

```

1: Initialize DDPG
2: Initialize replay buffer  $R$ 
3: for episode = 1,  $M$  do
4:   Sample a goal  $g$  and an initial state  $s_0$ 
5:   for  $t = 0, T-1$  do
6:     Sample an action  $a_t$  using DDPG behavioral policy
7:     Execute the action  $a_t$  and observe a new state  $s_{t+1}$ 
8:   end for
9:   for  $t = 0, T - 1$  do
10:     $r_t := r(s_t, a_t, g)$ 
11:    Store the transition  $(s_t || g, a_t, r_t, s_{t+1})$  in  $R$ 
12:    Sample a set of additional goals for replay  $G := \mathbb{S}$  (current episode)
13:    for  $g' \in G$  do
14:      Retrieve the rank of the transition from ER-CNN           ▷ ER-CNN to rank the experience
15:      if rank  $\geq$  threshold then                                ▷ Choose highest ranked transitions to store
16:         $r' := r(s_t, a_t, g')$ 
17:        Store the transition  $(s_t || g', a_t, r', s_{t+1})$  in  $R$           ▷ HER
18:      else
19:        Next loop
20:      end if
21:    end for
22:  end for
23:  for  $t = 1, N$  do
24:    Sample a minibatch  $B$  from the replay buffer  $R$ 
25:    Perform one step of optimization
26:  end for
27: end for

```



Fig. 1: An overview of the proposed approach for DDPG+HER with Experience Ranking.

algorithm jumps to the start of the next *for* loop to sample new experiences.

B. Convolutional Neural Network

In this section, we briefly describe common knowledge about CNN before continuing with the specific CNN structure that we used to rank the training experiences.

CNN uses various underlying layers consisting of convolution layers, pooling layers and fully connected layers to predict any sort of recommendation, objects in image or video, or some natural language processing for voice recognition [19] [20].

Convolution layers perform the convolution operation between an input image and a set of filters (kernels) whose weights are updated during training [21]. First, a convolution

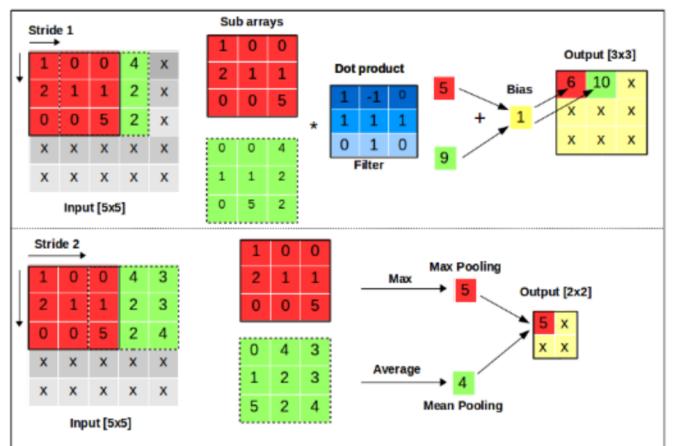


Fig. 2: Illustration of convolution and pooling layer.

layer will perform dot product between a sub-array of input (receptive fields) and a filter, whose size is often smaller than the input array. After that, the results will be summed, and a bias term will be added. The way convolution layers work is demonstrated in Figure 2.

Pooling layers often follows preceding convolution operations, and they reduce the spatial size of the input array in so-called down-sampling process. They do not have any trainable parameters. There are two options for pooling

layers: max pooling takes the maximum value from the input array, and mean pooling chooses the average values as shown in Figure 2.

Fully connected layers have full connection to all activation units of the previous layer and combine learned features from previous lower layers to make a decision based on a high-level reasoning.

In our experiment of a mobile robot manipulator learning to open a door, we used two CNNs: an experience ranking CNN (ER-CNN) for assisting HER by performing evaluation on door-opening training experiences and another for detecting the door handle. In both networks, we use the same baseline structure as shown in Figure 3, the only difference between them is the last Convolution layer whose number of outputs will depend on the number of classes. Parameters for all layers used for the two CNNs are shown in Table I.

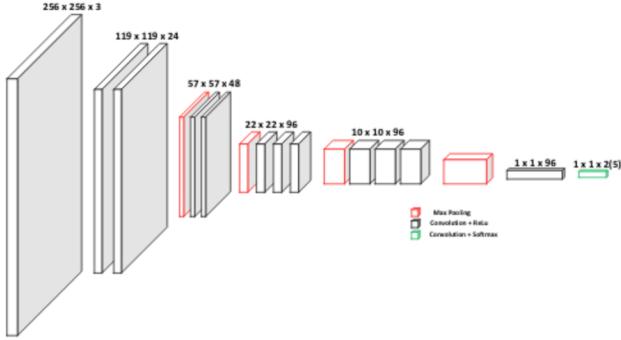


Fig. 3: CNN base line structure.

Layer Type	Filter Size	Number	Stride
Convolution + ReLU	20x20	24	2
Max Pooling	7x7	N/A	N/A
Convolution + ReLU	15x15	48	2
Max Pooling	4x4	N/A	N/A
Convolution + ReLU	10x10	96	2
Convolution + Softmax	1x1	2(5 for ER-CNN)	2

TABLE I: Parameters of 2 CNN networks.

1) *CNN for Experience Ranking*: The purpose of the ER-CNN is to rank the robot's experience during episodes of training. Only episodes with the score being higher than a threshold will be fed to the experience replay buffer. We rank each episode by directly looking at the images showing the states of the door and the manipulator at the end of that training episode. More specifically, if the door's hinge is open more than a 0.05 rad at the end of that episode, the experience is in Group 0 (highest score) and if not, its group (1-4) will be decided based on the distance from the robot's end-effector to the door handle.

This grouping process will be autonomously performed on images captured at the end of training episodes to generate labeled data for training the ER-CNN. The trained ER-CNN will then be able to predict the rank of any episode from images taken at the end of that episode. In this scenario, the trained ER-CNN's role is similar to a human expert

supervising the whole training process and giving score for each episode.

2) *CNN for Door Handle Detection*: The second CNN will try to detect door handle from images, and the manipulator will push open the door when the detection is positive. The final layer of this CNN will output only two outcomes: detecting the door or not. To further train the network for a better accuracy, we also take the advantage of the image stream coming from the ER-CNN. We perform online training on the pre-trained CNN by applying stochastic gradient descent with the learning rate being slowly decreased for better convergence near the minimum.

V. EXPERIMENT AND RESULTS

We used Gym [8] (Robotics Fetch robot environment) and MuJoCo physics simulator [22] to perform our experiment. We based on DDPG + HER implementation by OpenAI Baselines [23] and modified the *FetchReach-v1* environment to simulate the door opening task. We also verified the improved performance using other three robotics tasks in Gym including: *FetchPush-v1*, *FetchSlide-v1* and *FetchPick&Place-v1* as shown in Figure 4. We created the model for AUBO i5 robot, which we have in our Advanced Robotics and Automation (ARA) lab as shown in Figure 5. It is 6-DoF robotic arm, and it can move using the 4-wheeled mobile base but we keep it static in our experiment. Its joints are controlled by velocities and positions commands. Since the actual robot has built-in gravity compensation, we assume no gravity in the simulation. The complete simulation environment can be seen in Figure 6. At each training episode, the location of the door handle (the door also moves along) will be uniformly sampled inside a 30cm x 30cm x 30cm cube with the center fixed in a location. The robot is trained to push open the door using DDPG + HER assisted by ER-CNN.

A. Training Data Generation

1) *Experience Ranking CNN*: We use the steps listed in Algorithm 2 to generate training images and then classify them to have labeled data for training. The two inputs are the distance from the robot's end-effector to the door handle and the hinge's angle (to decide whether the door is open or not). The first *while loop* will generate images from training log files and name those images with distance information. The second *while loop* loops over these images, and based on their names (associated distance information), classify images into 5 different groups. The final data set contains in total 1630 images classified in group 0-4. Group 0 images indicate that the manipulator has successfully opened the door at the end of those episodes. Images belonging to Group 4 means that in the episode associated with it, the manipulator did not even touch the door, and the distance from its end-effector to the door handle was the largest. Images in group 1-3 are episodes being in between.

2) *Door Handle Detection CNN*: In MuJoCo, a camera is integrated to move with the mobile robot's end-effector, and it captures images during the door handle training episodes

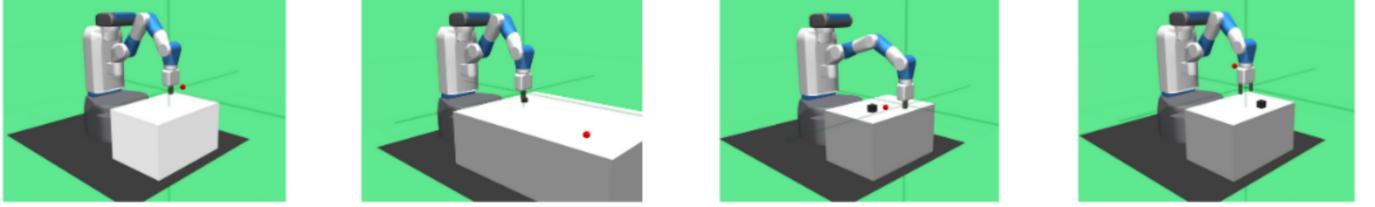


Fig. 4: Fetch Environment. From Left to Right: (1) *FetchReach-v1* Fetch has to move its end-effector to the desired goal position; (2) *FetchSlide-v1* Fetch has to hit a puck across a long table such that it slides and comes to rest on the desired goal; (3) *FetchPush-v1* Fetch has to move a box by pushing it until it reaches a desired goal position; and (4) *FetchPick&Place-v1* Fetch has to pick up a box from a table using its gripper and move it to a desired goal above the table.



Fig. 5: Auto-i5 robot working to open a door.

Algorithm 2 Generating labeled images for ER-CNN

```

1: while Replay Training Logs do      ▷ Generate images
2:   if hinge angle > 0.05 rad then
3:     File name = distance + 'O' + '.png'
4:   else
5:     File name = distance + '.png'
6:   end if
7: end while
1: while Not Last Image do ▷ Classify generated images
2:   Extract the name
3:   if Name contains 'O' then
4:     Group 0
5:   else
6:     Retrieve distance from the name
7:     Classify base on the distance
8:   end if
9: end while

```

creating the data set used for training and validating the detection network.

We also experiment with real images, where the training data are acquired from an original set of 180 images of door handles captured around the campus. After that, the data set is augmented by applying augmentation techniques such as horizontal flipping, random rotation, translation or color distortion. The final data set contains 1500 images of door handle and 1484 images of non-door handle images. Some samples of training images can be seen in Figure 7. The structure of the CNN shown in Figure 3, and it is used for detecting the door handle in both simulation and real images.

B. Training and Accuracy

Both CNNs were trained in an Alienware Aurora R7 computer with an NVIDIA GeForce GTX 1080. Training parameters are listed in Table II. The accuracy achieved for the ER-CNN was about 80.2% while the second CNN had the accuracy on the test set are 95.1% and 94.5% with simulated and outside images, respectively.

Name	Value
Input image	256x256
Batch size	16
Sample per epoch	1000
Epochs	500
Learning Rate	0.001
Loss	categorical crossentropy
Optimizer	SGD

TABLE II: CNN training parameters.

C. Improvement with Experience Ranking

1) *Door Push* training comparison: In this environment, we first try to measure the improvement when the classifier is 100% correct to rank the transition. We used entirely the two inputs: the distance from the robot's end-effector and the door handle, and the hinge's angle to correctly classify each episode before deciding to store them. We threw out any experience whose ranks are more than 4, and allow the remaining to enter the replay buffer. After that, we compared the training result with the performance of the original DDPG + HER as shown in Figure 8a. From the figure, we can see that with ER the performance (median success rate) increased more quickly than in case with original DDPG + HER. More interestingly after convergence, the success rate stayed at 100% while the original DDPG + HER still struggles to maintain the same accuracy. This phenomenon might indicate that supervised learning actually helps stabilize the learning in particular when the learning converges. We also implement the trained ER-CNN to replace the perfect classifier and compare the results. As expected, the performance decreases but it still converges faster and more robust than using original DDPG + HER.

2) Other tasks: Because door pushing is quite a simple task for DDPG + HER, we continue exploring the results in other more complex existing tasks in the Gym Robotics environment including: *FetchPick&Place-v1*, *FetchPush-v1* and *FetchSlide-v1*. In these tasks, we only perform the comparison between the perfect classifier (DDPG + HER + ER) and the original, without ER (DDPG + HER). From Figure 8b,c,d we can clearly see the improved learning performance gained from using ER, the learning converges faster and more robust after convergence than the original DDPG + HER. Our method took only half the number of episodes needed for the original DDPG + HER for convergence in *FetchPush-v1* and a third in *FetchPick&Place-v1* environment.

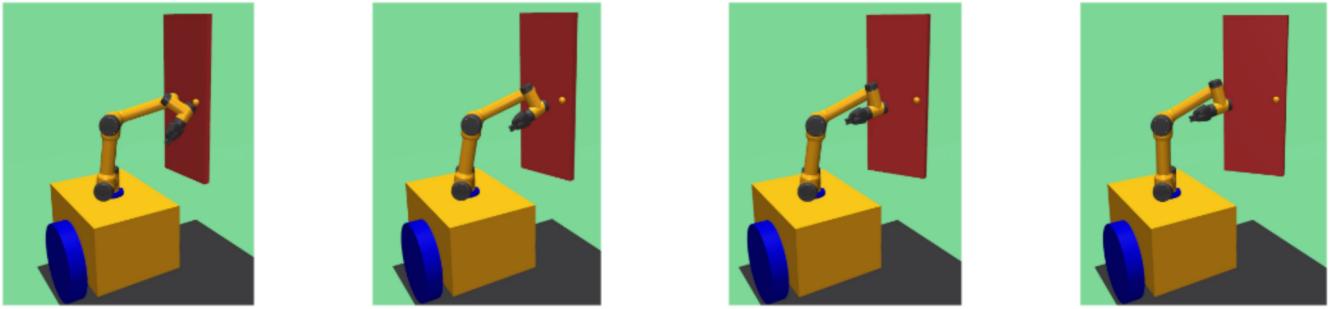


Fig. 6: Robot push open the door.

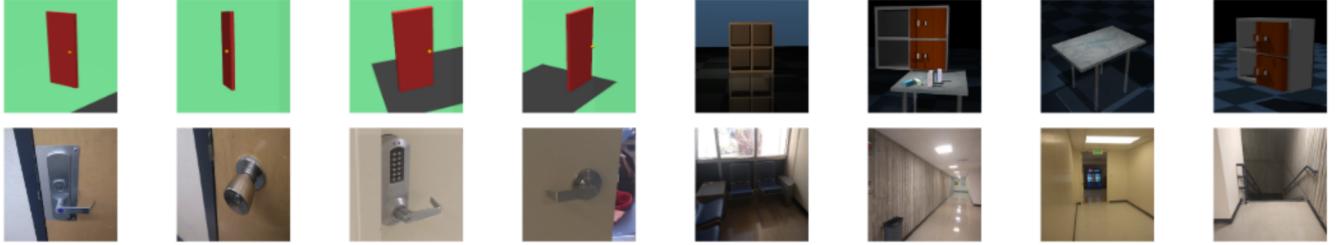
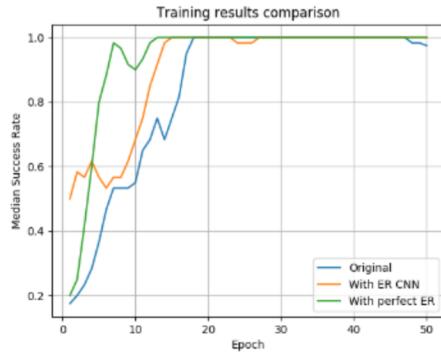


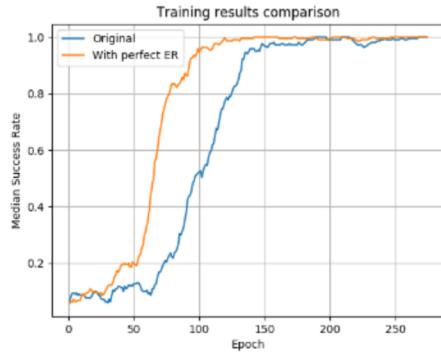
Fig. 7: Sample training images for door detection training (first row: simulation, second row: real.)



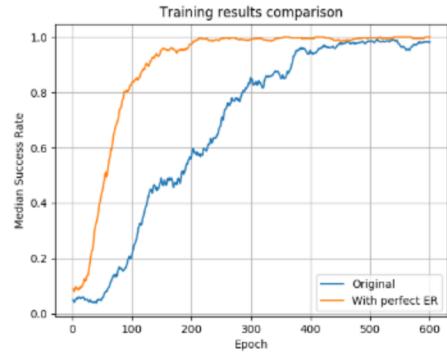
(a) Door Pushing Task.



(b) Fetch Slide v1 Task.



(c) Fetch Push v1 Task.



(d) Fetch Pick and Place v1 Task.

Fig. 8: Learning performance comparison in four simulated tasks.

VI. CONCLUSION AND FUTURE WORK

We have proposed a new method to improve performance of existing DDPG + HER combination by introducing experience ranking using CNN for each episode to decide whether it should be stored and replayed. We also validated the proposed methodology with four different robotics simulated tasks. By just throwing out the lowest rank experience, the learning performance outperformed the original DDPG +

HER in all tasks. Therefore, experience ranking using an external evaluation element can speed up learning, and it is easily extended to other more complex RL problems as well. We believe that by embedding task-relevant knowledge within robots (in our cases in forms of pre-trained CNNs), they will learn faster just like the way human learn a new task. Finally, perhaps to learn faster, we should focus more on the quality of training data rather than the quantity.

REFERENCES

- [1] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 3, April 2004, pp. 2619–2624 Vol.3.
- [2] E. Theodorou, J. Buchli, and S. Schaal, "Reinforcement learning of motor skills in high dimensions: A path integral approach," in *2010 IEEE International Conference on Robotics and Automation*, May 2010, pp. 2397–2403.
- [3] M. Kalakrishnan, L. Righetti, P. Pastor, and S. Schaal, "Learning force control policies for compliant manipulation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2011, pp. 4639–4644.
- [4] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba, "Hindsight experience replay," in *Advances in Neural Information Processing Systems*, 2017, pp. 5048–5058.
- [5] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *CoRR*, vol. abs/1509.02971, 2015. [Online]. Available: <http://arxiv.org/abs/1509.02971>
- [6] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous deep q-learning with model-based acceleration," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 2829–2838. [Online]. Available: <http://proceedings.mlr.press/v48/gu16.html>
- [7] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, V. Kumar, and W. Zaremba, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *CoRR*, vol. abs/1802.09464, 2018. [Online]. Available: <http://arxiv.org/abs/1802.09464>
- [8] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.
- [9] G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsáki, and M. B. Zugaro, "Selective suppression of hippocampal ripples impairs spatial memory," *Nature Neuroscience*, vol. 12, pp. 1222 EP –, Sep 2009. [Online]. Available: <http://dx.doi.org/10.1038/nn.2384>
- [10] V. Ego-Stengel and M. A. Wilson, "Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat," *Hippocampus*, vol. 20, no. 1, pp. 1–10, Jan 2010, 19816984[pmid]. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2801761/>
- [11] M. F. Carr, S. P. Jadhav, and L. M. Frank, "Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval," *Nature Neuroscience*, vol. 14, pp. 147 EP –, Jan 2011, review Article. [Online]. Available: <http://dx.doi.org/10.1038/nn.2732>
- [12] L. J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine Learning*, vol. 8, pp. 293–321, 1992.
- [13] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," in *NIPS Deep Learning Workshop*, 2013.
- [14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015. [Online]. Available: <http://dx.doi.org/10.1038/nature14236>
- [15] S. Cheng and L. M. Frank, "New experiences enhance coordinated neural activity in the hippocampus," *Neuron*, vol. 57, no. 2, pp. 303 – 313, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S089627307010306>
- [16] L. A. Atherton, D. Dupret, and J. R. Mellor, "Memory trace replay: the shaping of memory consolidation by neuromodulation," *Trends in Neurosciences*, vol. 38, no. 9, pp. 560 – 570, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0166223615001551>
- [17] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *International Conference on Learning Representations*, Puerto Rico, 2016.
- [18] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," *CoRR*, vol. abs/1611.01224, 2016. [Online]. Available: <http://arxiv.org/abs/1611.01224>
- [19] T. Le, S. Gibb, N. Pham, H. M. La, L. Falk, and T. Berendsen, "Autonomous robotic system using non-destructive evaluation methods for bridge deck inspection," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3672–3677.
- [20] S. Gibb, T. Le, H. M. La, R. Schmid, and T. Berendsen, "A multi-functional inspection robot for civil infrastructure evaluation and maintenance," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 2672–2677.
- [21] S. Gibb, H. M. La, T. Le, L. Nguyen, R. Schmid, and H. Pham, "Nondestructive evaluation sensor fusion with autonomous robotic system for civil infrastructure inspection," *Journal of Field Robotics*, vol. 0, no. 0, 2018. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21791>
- [22] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 5026–5033.
- [23] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Openai baselines," <https://github.com/openai/baselines>, 2017.