Memory-based Deep Reinforcement Learning for POMDP

Lingheng Meng¹, Rob Gorbet² and Dana Kulic³

Abstract—A promising characteristic of Deep Reinforcement Learning (DRL) is its capability to learn optimal policy in an end-to-end manner without relying on feature engineering. However, most approaches assume a fully observable state space, i.e. fully observable Markov Decision Process (MDP). In real-world robotics, this assumption is unpractical, because of the sensor issues such as sensors’ capacity limitation and sensor noise, and the lack of knowledge about if the observation design is complete or not. These scenarios lead to Partially Observable MDP (POMDP) and need special treatment. In this paper, we propose Long-Short-Term-Memory-based Twin Delayed Deep Deterministic Policy Gradient (LSTM-TD3) by introducing a memory component to TD3, and compare its performance with other DRL algorithms in both MDPs and POMDPs. Our results demonstrate the significant advantages of the memory component in addressing POMDPs, including the ability to handle missing and noisy observation data.

I. Introduction

Deep Reinforcement Learning (DRL) [1] [2] has been intensively studied in simulated environments, such as games [1] and simulated robots [2], as well as in real-world studies, such as robotics control [3], [4], [5] and human-robot interaction [6], [7], [8]. DRL enables end-to-end policy learning on tasks with high-dimensional state and action space, without relying on labour-consuming feature engineering. However, most works focus on developing algorithms [1], [9], [2], [10], [11], [12], [13] for Markov Decision Process (MDP) with fully observable state space [14], [15], i.e. the observation at each time step fully represents the state of the environment. Few works consider the more complex Partially Observable Markov Decision Process (POMDP) where the observation is just a partial representation of the underlying state. However, POMDP is ubiquitous in real robotics applications [16], [17], such as robot navigation [18], [19], robotic manipulation [20], autonomous driving [21], [22], [23], and planning under uncertainty [24], [25], [26], [27]. Partial observability may be due to limited sensing capability, or an incomplete system model resulting in uncertainty about full observability.

POMDP has been tackled with the concept of belief state [28], which represents the agent’s current belief about the possible physical states it might be in, given the sequence of actions and observations up to that point. These algorithms are designed to learn the belief state, then the value function and/or the policy are learned based on the belief state [29], [30] [31], [32]. However, these methods need to know the environment model and the state space and they only work on tasks with small state and action spaces.

POMDP has also been addressed with DRL, for both discrete [33], [9], [34] and continuous [35], [23] control problems. Recurrent Neural Networks (RNN) have been exploited in DRL to address the POMDP by considering both the current observation and action, and the history of the past observations and actions [36], [35], [37], [34], [9].

In this paper, we will present a memory-based DRL, called Long-Short-Term-Memory-based Twin Delayed Deep Deterministic Policy Gradient (LSTM-TD3), for continuous robot locomotion control, where both MDP and POMDP versions of the tasks are investigated to demonstrate the effectiveness of the proposed method for solving both POMDPs and MDPs. We will also provide an ablation study to show the contribution of each design component, and discuss the advantages and disadvantages of the proposed method.

II. Related Work

A number of previous works have investigated how memory can be incorporated into DRL. Deep Recurrent Q-Learning (DRQN) [33] adds recurrency to the Deep Q-Network (DQN) [38] by replacing the first post-convolutional fully-connected layer with a recurrent Long-Short-Term-Memory (LSTM). The results on Atari 2600 games show DRQN significantly outperforms DQN on POMDPs, which validates the effectiveness of memory extracted by the LSTM on solving POMDPs. [9] investigated a similar idea but
augmented the structure with an auxiliary game feature, e.g. presence of enemies, learning, on 3D environments in first-person shooter games. The results show the proposed architecture substantially outperforms DRQN. These methods only consider past observation in the history. [34] proposed Action-specific Deep Recurrent Q-Network (ADRQN) to also consider past actions in the memory. However, these works are based on tasks with discrete action spaces, rather than on continuous control tasks.

[36] extended Deterministic Policy Gradient (DPG) [39] to Recurrent DPG (RDPG) by adding LSTM and investigated it on continuous control tasks with partial observations. Dramatic performance improvement was observed with memory. However, even though the observation space was large for some tasks, the action space had relatively few dimensions for the investigated tasks. [35] investigated RDPG on bipedal locomotion tasks with both visual and sensory input, but only one task was examined. Different from directly optimizing RNN, [37] proposed to augment the observations and actions with the continuous memory states and use guided policy search to optimize a linear policy. The method shows better performance than other policy search methods. However, the guided policy search is less powerful and generalizable than non-linear policy.

In our work, we consider continuous control tasks with large observation and action spaces and propose LSTM-TD3 within a recurrent actor-critic framework, which is a further improvement of RDPG by exploiting TD3 to reduce the overestimation problem.

III. Background

A. Decision Process

A Markov Decision Process (MDP) is a sequential decision process for a fully observable, stochastic environment with a Markovian transition model and additive rewards [40], [28]. Formally, MDP can be defined as a 4-tuple \((S, A, P, R)\), where \(S\) is the state space, \(A\) is the action space, \(P\) is the transition probability and \(R\) is the reward function. At each discrete time \(t\), an agent selects an action \(a_t \in A\) in state \(s_t \in S\), transitions to the next state \(s_{t+1}\) with probability \(P(s_{t+1} | s_t, a_t)\), and receives the immediate reward \(R(s_t, a_t, s_{t+1})\).

Partially Observable Markov Decision Process (POMDP) [41], [42], [28] is a generalization of a MDP, but does not assume that the state is fully observable, and is defined as a 6-tuple \((S, A, P, R, O, \Omega)\), where \(S\), \(A\), \(P\), and \(R\) are the same as that in MDP, with an additional observation space \(O\) and observation model \(\Omega\). Although the underlying state transition in a POMDP is the same as that in an MDP, the agent cannot observe the underlying state, instead it receives an observation \(o_{t+1} \in O\) when reaching the next state \(s_{t+1}\) with the probability \(\Omega(o_{t+1} | s_{t+1})\).

The goal for an agent in either MDP or POMDP is to choose actions at each time step that maximize its expected future discounted return \(\mathbb{E}[\sum_{i=0}^{\infty} \gamma^i r_i]\), where \(r_t\) is the immediate reward received at time \(t\) and \(\gamma \in [0, 1]\) is the discount factor that describes the preference of the agent for current rewards over future rewards.

B. Reinforcement Learning (RL)

RL [43] solves decision problems such as MDP and POMDP based on a learning paradigm where an agent learns to act by trial-and-error without knowing the underlying transition and reward model. Specifically, at a discrete time \(t\) an agent interacts with the external environment by taking action \(a_t\) according to either a stochastic policy \(\pi(a_t | s_t)\) or a deterministic policy \(a_t = \mu(s_t)\) when observing the current state \(s_t\). By continuously interacting with the environment and receiving new states and rewards, an agent learns an optimal policy to maximize the expected future return. Accompanying the policy \(\pi\) or \(\mu\), it is common to learn a state-value function \(V(s)\) and/or an action-value function \(Q(s, a)\). For MDP, The state-value function \(V(\pi)(s)\) of a state \(s\) under a policy \(\pi\) represents the expected return starting from \(s\) and following \(\pi\) thereafter, and the action-value function \(Q^*(s, a)\) of taking action \(a\) in state \(s\) under a policy \(\pi\) is the expected return starting from \(s\), taking \(a\), and following \(\pi\) thereafter. For POMDP, since the state \(s\) is not observable, the observation \(o\) is used in learning these value functions or policy. Traditional RL methods employ either tabular representation or simple function approximation, with hand-crafted feature construction, to represent the value function, while DRL [1] [2] learns the value function and policy in an end-to-end manner without feature engineering by recruiting deep neural networks.

C. Long-Short-Term-Memory (LSTM)

LSTM [44] is a type of Recurrent Neural Network (RNN) [45] that has an outer recurrence from the outputs to the inputs of the hidden layer and also an internal recurrence between LSTM-Cells. Within a LSTM-Cell, a system of gating units control the flow of information, and enable the remembering and forgetting of information given a sequence of inputs.

D. Twin Delayed Deep Deterministic Policy Gradient (TD3)

TD3 [12] is a variant of Deep Deterministic Policy Gradient (DDPG) [2] to address the overestimation problem [46] in Actor-Critic methods. Specifically, TD3 employs two critics \(Q_1\) and \(Q_2\), and uses the minimum of the predicted optimal future return in state \(s_{t+1}\) to bootstrap the Q-value of the current state \(s_t\) and action \(a_t\).

IV. Proposed Approach

In this paper, we propose a memory-based DRL algorithm named LSTM-TD3 within a recurrent actor-critic framework, where both the actor and the critic employ recurrent neural networks, as illustrated in Fig. 1. In this section, we will first introduce the proposed recurrent actor-critic framework, then present the optimization method for the actor-critic.

In the proposed approach, a mini-batch of \(N\) experiences \(\{(h^i_t, o_t, a_t, r_t, o_{t+1}, d_t)\}_{i=1}^{N}\) is sampled from the replay buffer \(D\) of experiences \((o_t, a_t, r_t, o_{t+1}, d_t)\), where for each sample the past history \(h^i_t\) with length \(l\) until observation \(o_t\)
at time $t$ is defined as:

$$h^l_t = \begin{cases} a_{l-t}, a_{l-t-1}, \cdots, a_{l-t-1} & \text{if } l \geq 1, \\ a^0, a^0 & \text{Otherwise.} \end{cases} \quad (1)$$

where $a^0$ and $a^0$ are the zero-valued dummy observation and action with the same dimension as that of the normal observation and action. As defined in Eq. 1, if history length $l \geq 1$ and time step $t \geq 1$, the history $h^l_t$ at time $t$ is defined as the past $l$ (observation, action) pairs, otherwise no history is used and zero-valued dummy (observation, action) are used as input to the memory component.

A. Recurrent Actor-Critic Framework

The structure of the proposed recurrent actor-critic framework is illustrated in Fig. 1, where Long-Short-Term-Memory is introduced in the memory component to extract information beneficial to the actor and critic from past history. The proposed framework can handle history of any length.

Formally, given a mini-batch sample of experiences, the memory-based critic $Q$, as illustrated in Fig. 1, can be seen as a compound function of the memory extraction $Q^{me}$, the current feature extraction $Q^{cf}$, and the perception integration $Q^{pi}$ components, following Eq. 2:

$$Q(o_t, a_t, h^l_t) = Q^{me} \circ Q^{cf} \circ Q^{pi} = Q^{pi}(Q^{me}(h^l_t) \times Q^{cf}(o_t, a_t)) \quad (2)$$

where $\times$ indicates the concatenation operation, $Q^{me}$ is the extracted memory based on history $h^l_t$, and $Q^{cf}$ is the extracted current feature based on current observation $o_t$ and action $a_t$.

Similarly, the memory-based actor $\mu$ is also a compound function of the memory extraction $\mu^{me}$, the current feature extraction $\mu^{cf}$, and the perception integration $\mu^{pi}$ components, defined as follows:

$$\mu(o_t, h^l_t) = \mu^{me} \circ \mu^{cf} \circ \mu^{pi} = \mu^{pi}(\mu^{me}(h^l_t) \times \mu^{cf}(o_t)) \quad (3)$$

where $\mu^{me}$ is the extracted memory based on history $h^l_t$, and $\mu^{cf}$ is the extracted current feature based on current observation $o_t$.

B. Optimization of the Recurrent Actor-Critic

The optimization of the proposed recurrent actor-critic framework follows that of TD3. Specifically, each critic $Q \in \{1, 2\}$ is optimized to minimize the mean-square-error between the predicted $Q$ and the estimated target $\hat{Q}$ with respect the the parameters $\theta^{Q_j}$ of the critic $Q_j$, as follows:

$$\min_{\theta^{Q_j}} \mathbb{E}_{\{(h^l_t, o_t, a_t, r_t, o_{t+1}, d_t)\}_{i=1}^N} (Q_j - \hat{Q})^2 \quad (4)$$

where given the definition of memory-based critic (Eq. 2) and actor (Eq. 3), the target Q-value $\hat{Q}$ based on the target actor $\mu^-$ and critic $Q^-$ is defined as Eq. 5

$$\hat{Q} = r_t + \gamma * (1 - d_t) * \min_{j=1,2} Q^-_{j} (o_{t+1}, a^{-}, h^l_{t+1}) \quad (5)$$

where $a^- = \mu^-(o_{t+1}, h^l_{t+1}) + \epsilon$ with $\epsilon \sim \text{clip}(N(0, \sigma), -c, c)$ and $c$ is the boundary of target action noise, $h^l_{t+1} = (h^l_t - (o_{t-1}, a_{t-1})) \cup (o_t, a_t)$ is the $l$ observation and action pairs before $o_{t+1}$, and the minimum of the estimated optimal Q-values of the two target critics in $(o_{t+1}, h^l_{t+1})$ is taken to bootstrap the target Q-value of $(o_t, a_t, h^l_t)$.

For the actor, its parameters $\theta^\mu$ are optimized to maximize the approximated Q-value in observation $(o_{t+1}, h^l_{t+1})$ and the corresponding estimated optimal action $\mu(o_t, h^l_t)$ with respect to the parameters of the actor, as follows:

$$\max_{\theta^\mu} \mathbb{E}_{\{(h^l_t, o_t)\}_{i=1}^N} Q(1, \mu(o_t, h^l_t), h^l_t) \quad (6)$$

where the $Q$ could be either of the two critics $Q_1$ and $Q_2$, as in TD3. The pseudo-code for optimizing the recurrent actor-critic can be found in Alg. 1.

V. Experiment Settings

The tasks (Fig. 3) tested in this work come from PyBullet-Gym 1, an open-source implementation of the OpenAI Gym MuJoCo environment based on BulletPhysics 2. In this work, an MDP-version and 4 POMDP-versions of each task are investigated, described in Table I. The MDP-version is the original task, as it has a fully observable state-space, while the 4 POMDP-versions simulate different scenarios that potentially cause partial observability in real applications. Specifically, the POMDP-RemoveVelocity (POMDP-RV) is designed to simulate the scenario where the observation space is not well-designed, which is applicable to a novel control task that is not well-understood by researchers and therefore the designed observation may not fully capture the underlying state of the robot. The POMDP-Flickering (POMDP-FLK) models the case where remote sensor data is lost during long-distance data transmission. The case when a subset of the sensors are lost is simulated in POMDP-RemoveRandomSensorMissing (POMDP-RSM). Sensor noise is simulated in POMDP-RandomNoise (POMDP-RN).

The baselines used to compare with the proposed LSTM-TD3 are the DDPG [2], SAC [13], TD3 [12], TD3 with

1https://github.com/bulletphysics/bullet3
2https://github.com/bulletphysics/bullet3
Algorithm 1: Pseudo-code for LSTM-TD3

**Input**: History length \( L \)

1. Initialize critics \( Q_{\theta^{Q_1}}, Q_{\theta^{Q_2}}, \) and actor \( \mu_{\theta^\mu} \) with random parameters \( \theta^{Q_1}, \theta^{Q_2} \) and \( \theta^\mu \)
2. Initialize target networks \( \theta^{Q_1}_t \leftarrow \theta^{Q_1}, \theta^{Q_2}_t \leftarrow \theta^{Q_2} \) and \( \theta^\mu_t \leftarrow \theta^\mu \)
3. Initialize environment \( o_t = \text{env.reset()} \), past history \( h_t^1 \leftarrow 0 \), and replay buffer \( D \)
4. for \( t = 1 \) to \( T \) do
   /* Interacting */
   5. Select action with exploration noise \( a_t \sim \mu_{\theta^\mu}(o_t, h_t^1) + \epsilon, \epsilon \sim N(0, \sigma) \)
   6. Interact and observe new observation, reward, and done flag: \( o_{t+1}, r_t, d_t = \text{env.step}(a_t) \)
   7. Store experience tuple \( (o_t, a_t, r_t, o_{t+1}, d_t) \) in \( D \)
   8. if \( d \) then
      9. Reset environment \( o_{t+1} = \text{env.reset()} \) and history \( h_{t+1}^1 \leftarrow 0 \)
   else
      10. /* Update \( h_{t+1}^1 \) */
      11. \( h_{t+1}^1 = \{(h_t^1 - (o_{t-1}, a_{t-1})) \cup (o_t, a_t)\} \)
   end
   /* Learning */
   12. Sample mini-batch of \( N \) experiences with their corresponding histories \( \{(h_t^1, o_t, a_t, r_t, o_{t+1}, d_t)\}_{i=1}^N \) from \( D \)
   13. Optimize \( Q_j \) according to Eq. 4
   14. Optimize \( \mu \) according to Eq. 6
   15. Update target actor-critic

---

**Table I: MDP- and POMDP-version of Tasks**

| Name          | Description              | Hyperparameter |
|---------------|--------------------------|----------------|
| MDP           | Original task            | -              |
| POMDP-RV      | Remove all velocity-related entries in the observation space | - |
| POMDP-FLK     | Reset the whole observation to 0 with probability \( p_{flk} \). | \( p_{flk} \) |
| POMDP-RN      | Add random noise \( \epsilon \sim N(0, \sigma_{rn}) \) to each entry of the observation. | \( \sigma_{rn} \) |
| POMDP-RSM     | Reset an entry of the observation to 0 with probability \( p_{rsm} \). | \( p_{rsm} \) |

are averaged over 10 evaluation episodes based on 4 different random seeds. All hyperparameters and additional results are moved to the Supplementary Material (available at [http://arxiv.org/abs/2102.12344](http://arxiv.org/abs/2102.12344)) due to space limitation.

VI. Results

A. Performance Comparison

Fig. 4 shows the learning curves of the sampled tasks, where the first column shows the performance on MDP, while the following 4 columns show results on POMDPs. The results on MDP show that the proposed method has competitive performance to the baselines. The results on the POMDPs highlight the advantage of having memory when solving partially observable tasks. On all types of POMDP, LSTM-TD3 outperforms all baselines, except on POMDP-RV of HalfCheetahPyBulletEnv-v0, where LSTM-TD3(5) shows slightly worse performance than TD3. Although TD3-OW shows better performance than DDPG, TD3, and SAC on POMDPs for most tasks, it still fails for some POMDPs, such as the POMDP-FLK version of most tasks. This reveals that simply concatenating observations is not a good choice, compared to having a LSTM-based memory extraction component as that in LSTM-TD3.

Particularly, a significant performance gap can be observed on POMDP-FLK for all tasks (the 3rd column in Fig. 4), where the baselines basically fail while LSTM-TD3 achieves comparative performance to that on MDP. This is especially promising for tasks where whole sensor data may be lost, either caused by hardware failure or by temporary occlusion, etc. Similar, dramatic performance improvement can be seen on POMDP-RN and POMDP-RSM.

Surprisingly, comparing LSTM-TD3 and TD3, memory does not always help for POMDP-RV (the 2nd column in Fig. 4) of HalfCheetahPyBulletEnv-v0 and AntPyBulletEnv-v0. Intuitively, if the velocity is important to learn a task, there is no way to infer such information without past observations i.e. memory. However, if previous observations are available, the velocity can be inferred by differences in position between consecutive steps. This intuition can be clearly observed on the POMDP-RV of HopperPyBulletEnv-v0 and InvertedDoublePendulumPyBulletEnv-v0, where the performance of LSTM-TD3 is significantly better than that of TD3. For the HalfCheetahPyBulletEnv-v0 and AntPyBulletEnv-v0, if we compare the performance on MDP and POMDP-

---

Observation-Window (TD3-OW) where the \( o_t \) is simply concatenated with the observations within the history window \( h_t^1 \) to form an observation as input, and TD3 with Observation-Window-AddPastAct (TD3-OW-AddPastAct) where \( o_t \) is concatenated with the observations and the actions within the history window \( h_t^1 \). The hyperparameters for the baseline algorithms were always the defaults provided in OpenAI SpinningUp\(^3\). For the proposed algorithm, hyperparameters were empirically set to that for TD3, and the network structures of the LSTM-TD3 were chosen to have the similar number of parameters of the networks in TD3. All reported results

---

\(^3\)https://spinningup.openai.com
Interestingly, by comparing the results of TD3-OW and TD3-OW-AddPastAct, we found that adding past actions consistently harms the performance compared to TD3-OW, which does not have past actions in its observation window. Even though TD3-OW-AddPastAct still outperforms TD3 on POMDP, it performs worse than TD3 on MDP, which is undesirable if we have no prior knowledge about if the current design of the observation space is partially or fully observable. Ideally, even if the past action-related information does not provide anything new besides the past observation, it can be safely ignored and should not harm the performance. We think this is related to the simple construction method where actions are concatenated with observations to form a single observation that includes history information. In this way, the observation dimension is expanded, which makes the learning harder. In addition, this simple construction method does not differentiate what is the most recent observation, which is normally more valuable than far away observations for making a decision, and what are far away observations, but treats everything equally important no matter when the observations are observed. This observation based on TD3-OW and TD3-OW-AddPastAct in fact supports our idea to structurally separate the memory extraction and current feature extraction in the recurrent actor-critic framework (Fig. 1) designed for LSTM-TD3, then combine them together to further learn a presentation of the critic and the actor. In section VII-C, we will further investigate if adding past actions is beneficial for LSTM-TD3.

B. Policy Generalization

To better understand the generalization of the learned policy using LSTM-TD3, we evaluated the learned policy on a different version of a given task, e.g. if the policy is learned on the MDP-version of a task, and evaluated on the POMDP-versions of the task. This is valuable for real applications where the environment may be non-stationary. Fig. 5 shows the cross evaluation results on AntPyBulletEnv-v0, where POMDP-RV is not examined here as it has different observation dimension which corresponds to a different input shape for the neural networks. From the first panel, i.e. policies trained on MDP, TD3, TD3-OW, and LSTM-TD3 significantly outperform DDPG, SAC, and TD3-OW-AddPastPact. When evaluating on POMDPs, there is always a decrease for TD3, TD3-OW, and LSTM-TD3, but LSTM-TD3 is the most robust and achieves better performance.
POMDPs, whereas TPS seems not so important to the performance compared to Full on all MDPs and most MDP-version tasks. Recall that one

VII. Ablation Study

To further understand the effect of each component of the proposed LSTM-TD3, in this section we perform an ablation study. Specifically, we examine the effects of the following components: (1) double critics (DC), (2) target policy smoothing (TPS), (3) current feature extraction (CFE), and (4) including past actions (PA) in the history. Fig. 7 shows the learning curves of ablated versions of LSTM-TD3, each removing a different component.

A. Effect of Double Critics and Target Policy Smoothing

As shown in Fig. 7, Full−DC shows a significant decrease in performance compared to Full on all MDPs and most POMDPs, whereas TPS seems not so important to the best performance of Full. When simultaneously removing DC and TPS, the performance significantly decreases. Note that without DC and the TPS, the LSTM-TD3 is in fact reduced to LSTM-DDPG, similar to RDPG proposed in [36]. To ease the comparison, we summarized the results of DDPG and LSTM-DDPG, i.e. Full−DC−TPS, in Table II. When compared on the same version of a task, LSTM-DDPG always outperforms DDPG, which can be observed by comparing the results in each row. Remarkably, LSTM-DDPG even achieves significantly better performance on POMDPs than DDPG on MDP.

B. Effect of Current Feature Extraction

In this paper, we intentionally separate the memory extraction and the current feature extraction, then combine them together (Fig. 1), in order to differentiate the current and the past and to reduce the interference from useless information in the memory. Alternatively, we can directly combine the current observation for the action (or the current observation and action for the critic) with the extracted memory, i.e. removing the CFE (Full−CFE) (Fig. 6a). As shown in Fig. 7, Full−CFE performs much worse, compared to Full, especially for MDP-version tasks.
VIII. Discussion and Conclusion

In this paper, we proposed a memory-based DRL algorithm called LSTM-TD3 by combining a recurrent actor-critic framework with the TD3. The proposed LSTM-TD3 was compared to standard DRL algorithms on both the MDP- and POMDP-versions of continuous control tasks. Our results show that LSTM-TD3 not only achieves significantly better performance on POMDPs than the baselines, but also retains the state-of-art performance on MDP. Our ablation study shows that all components are essential to the success of the LSTM-TD3 where DC and TPS help in stabilizing learning, CFE is especially important to retain the good performance in MDP, and PA is beneficial for tasks where past actions provide information about the current state of the agent.

The proposed approach is particularly useful when the engineers do not have enough knowledge about the environment model and the appropriate design of the observation space to capture the underlying state. Memory can be useful in such a scenario to help infer the underlying state. However, the interpretation of the extracted memory is a challenge. If there is a way to properly interpret the extracted memory, such information, e.g., if the current task is a POMDP or a MDP, can be exploited to improve the observation space design and advance the understanding of the task. Unfortunately, without adding a specific constraint terms in the cost functions Eq. 4 and 6 to facilitate the interpretation, there is no way to properly interpret the extracted memory. Future research should give attention to this direction.

In this paper, for each run of LSTM-TD3 we treat the history length $l$ as a hyper-parameter and fixed it for each run. While LSTM-TD3 with a history length $l = 5$ achieves good performance on the devised POMDPs, this may not be achieved for other tasks where the underlying state depends on less recent memory. However, a long history length increases computation resources and time, during both the training and the inferring, i.e., decision making, phases. In the future, an approach for dynamic adaptation of history length $l$ that achieves the best performance while minimising training and decision making time should be investigated.

ACKNOWLEDGMENT

This research was enabled in part by support provided by Compute Canada (www.computecanada.ca). This work is...
Memory-based Deep Reinforcement Learning for POMDP:
Supplementary Material

Lingheng Meng\textsuperscript{1}, Rob Gorbet\textsuperscript{2} and Dana Kulic\textsuperscript{3}

Abstract—This manuscript is the supplementary material for the paper "Memory-based Deep Reinforcement Learning for POMDP" by Lingheng Meng, Rob Gorbet and Dana Kulic. It provides additional implementation details, and complementary results.

I. Algorithms Implementation

The implementation of the algorithms is based on OpenAI Spinningup\textsuperscript{1}. The code for this work can be found in https://github.com/LinghengMeng/spinningup_new. Table I details the hyperparameters used in this work, where \( - \) indicates the parameter does not apply to the corresponding algorithm. For the actor and critic neural network structure of LSTM-TD3, the first row corresponds to the structure of the memory component, the second row corresponds to the structure of the current feature extraction, and the third row corresponds to the structure of perception integration after combining the extracted memory and the extracted current feature.

II. Supplementary Results

A. Performance Comparison

Table II summarizes the maximum average return of each algorithm on different tasks, where the POMDP-FLK, POMDP-RN, and POMDP-RSM are examined with \( p_{flk} = 0.2 \), \( \sigma_{rn} = 0.1 \), and \( p_{rsm} = 0.1 \). From Table II, we can see that LSTM-TD3 either outperforms or achieves comparative performance to other baselines on MDP, and significantly outperforms other baselines on POMDP-versions of each task. These results provide evidence of the promising advantages of memory based DRL on both MDP and POMDP. On one hand, LSTM-TD3 can be used out-of-box without caring too much of the design of the observation as memory component can partially compensate the lost information. On the other hand, by comparing the performance difference of adding and removing the memory component, LSTM-TD3 provides a way to detect if the current design of the observation space is improvable or not in terms of capturing the underlying state.

B. Robustness to Partial Observability

Fig. 1 compares the proposed LSTM-TD3 with the baselines in terms of the robustness to different partial observabilities, where the higher the \( \sigma_{rn} \) and the \( p_{rsm} \), the lower the observability. In general, LSTM-TD3 has better robustness than TD3-OW which has better robustness than TD3. A small reduction of observability can be handled by LSTM-TD3, but for POMDP with severe reduction of observability LSTM-TD3 performance also degrades.

C. Effect of History Length

The history length \( l \) is a hyperparameter of LSTM-TD3, and it determines the maximum history length in the observation window. As illustrated in Fig. 3 and Table II, LSTM-TD3 with a relatively short history length \( l = 5 \) produces significantly better performance than other baselines without a memory component. Fig. 2 shows the performance of LSTM-TD3 with different history lengths on POMDP-FLK AntPyBulletEnt-v0 with various flickering probabilities.

\begin{table}[h!]
\centering
\caption{Hyperparameters for Algorithms}
\label{tab:hyperparameters}
\begin{tabular}{|c|c|c|c|c|}
\hline
Hyperparameter & DDPG & TD3 & SAC & LSTM-TD3 \\
\hline
discount factor: \( \gamma \) & 0.99 & & & \\
batch size: \( N_{batch} \) & & 100 & & \\
replay buffer size: \([D]\) & & 10\textsuperscript{6} & & \\
random start step: \( N_{start, step} \) & & 10000 & & \\
update after \( N_{update, after} \) & & 1000 & & \\
target NN update rate \( \tau \) & & 0.005 & & \\
optimizer & & Adam [1] & & \\
actor learning rate \( \beta_{actor} \) & & 10\textsuperscript{-3} & & \\
critic learning rate \( \beta_{critic} \) & & 10\textsuperscript{-3} & & \\
actor NN structure: & [256, 256] & [128] + [128] & [128] & [128, 128] \\
critic NN structure: & [256, 256] & [128] + [128] & [128] & [128, 128] \\
actor exploration noise \( \sigma_{act} \) & 0.1 & - & 0.1 & \\
target actor noise \( \sigma_{targ, act} \) & - & 0.2 & - & 0.2 \\
target actor noise clip boundary \( c_{targ, act} \) & - & 0.5 & - & 0.5 \\
policy update delay & - & 2 & - & 2 \\
entropy regulation coefficient \( \alpha \) & - & 0.2 & - & \\
history length \( l \) & - & - & - & \{0, 1, 3, 5\} \\
\hline
\end{tabular}
\end{table}

\textsuperscript{1}Lingheng Meng is with Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada lingheng.meng@uwaterloo.ca
\textsuperscript{2}Rob Gorbet is with Departments of Knowledge Integration and Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada rob.gorbet@uwaterloo.ca
\textsuperscript{3}Dana Kulic is with the Faculty of Engineering, Monash University, 14 Alliance Lane, Melbourne, Victoria, Australia dana.kulic@monash.edu

\textsuperscript{1}https://spinningup.openai.com
Fig. 1: Performance Comparison on POMDP-version of AntPyBulletEnv-v0 with different RSMs.

| Task             | Version       | Algorithms       | Task             | Version       | Algorithms       |
|------------------|---------------|------------------|------------------|---------------|------------------|
| MDP              | DDPG          | 487.6 ± 6.1      | MDP              | DDPG          | 850.5 ± 101.4    |
| HalfCheetahPB    | POMDP-RV      | 508.6 ± 23.9     | HalfCheetahPB    | POMDP-RV      | 683.5 ± 101.4    |
| Walker2DPB       | POMDP-RS      | 283.7 ± 27.0     | Walker2DPB       | POMDP-RS      | 465.2 ± 51.0     |
| AntPB            | POMDP-RN      | 268.7 ± 70.2     | AntPB            | POMDP-RN      | 449.6 ± 18.5     |
|                  | POMDP-RSM     | 1000.0 ± 0.0     |                  | POMDP-RSM     | 683.5 ± 101.4    |

For POMDP-RSM, $p_{fik} = 0.2$. For POMDP-RN, $\sigma_{rn} = 0.1$. For POMDP-RSM, $\rho_{prsm} = 0.1$.

Fig. 2: Relationship Between Partial-Observability and History Length.

$p_{fik} = \{0.05, 0.1, 0.2, 0.5, 0.8\}$ where the higher the $p_{fik}$, the less the observability. For $p_{fik} = 0.05$, LSTM-TD3(3) and LSTM-TD3(5) show similar performance and both significantly outperform LSTM-TD3(0) and LSTM-TD3(1). When $p_{fik}$ increases from 0.05 to 0.1, LSTM-TD3(5) maintains similar performance, but LSTM-TD3(3) experiences a dramatic decrease. This means the decreased observability can still be compensated with history length 5, but cannot be compensated with history length 3. When further reducing the observability to $p_{fik} = 0.2$, the performance of LSTM-TD3(5) is also degraded, as shown in the 3rd panel of Fig. 2. When $p_{fik}$ is increased to 0.5 and 0.8, all examined history lengths fail the task (the last two panels in Fig. 2).

D. Policy Generalization

The second extension is to evaluate the learned policy with different history length from that used during training. If
we use $l_{\text{train}}$ and $l_{\text{eval}}$ to represent the history length used during training and evaluation respectively, the generalization capability for both $l_{\text{train}} > l_{\text{eval}}$ and $l_{\text{train}} < l_{\text{eval}}$ can be useful in different scenarios. On one hand, for $l_{\text{train}} > l_{\text{eval}}$, if $l_{\text{train}}$ and $l_{\text{eval}}$ can achieve the same performance, using a shorter history length during evaluation can reduce the inferring time of an action, and this is valuable for tasks where real-time decision making is important and training can be run in parallel and is not time-sensitive. On the other hand, if $l_{\text{train}} < l_{\text{eval}}$, if $l_{\text{eval}}$ can achieve better performance than $l_{\text{train}}$ by increasing the history length, which means how to extract useful memory information can be learned with a shorter history length and longer history length during evaluation is only to extract more useful information, then using a shorter history length during training can speed up the training and reduce resource consumption, at the same time without compromising the performance. As for other DRL algorithms, the training is on a mini-batch while the evaluating is only on a single data, so if the $l_{\text{train}}$ is large, the training will take more time and computation resources than evaluation. Therefore, it a good choice to achieve the same performance by reducing the cost during training and increasing the cost during evaluation. Normally, training cost is unvalued by researchers, especially some research [2], [3] proposed to use large distributed computer cluster. However, there are cases, where the robot cannot communicate fllely with the remote computation center and onboard computation resources are limited, that the effort in saving computation is still required. This practical consideration inspires the second evaluation extension.

1) Evaluation with Different History Lengths: Fig. 4 illustrates the evaluation results of LSTM-TD3 with various history lengths that may be different from the history length used for training. From this figure, it can be seen that when trained with a specific history length $l_{\text{train}}$ but evaluated with a different history length $l_{\text{eval}} > l_{\text{train}}$, the performance remains at the same level. However, when evaluated with a history $l_{\text{eval}} < l_{\text{train}}$, the performance cannot be guaranteed, as for some cases a shorter evaluation history length will cause dramatic decrease in performance, e.g. LSTM-TD3(3) on POMDP-RN (the blue line in the 4th panel of Fig. 4), while for others a shorter evaluation history length can still achieve similar performance, e.g. LSTM-TD3(5) on POMDP-RN and on POMDP-RSM (the red line in the 4th and 5th panels of Fig. 4). Based on this observation, it seems the performance can be generalized to a longer evaluation history length rather than a shorter one. It is worth to note that the longest training history length investigate here is only 5, so more valuable insights may be found with more results with longer training history length.
The results that when evaluating with history length longer than 0, the performance does not change (the cyan lines in Fig. 4) of LSTM-TD3(0), are very interesting, because in the training phase LSTM-TD3(0) takes zero-valued dummy observation and action as history as defined in Eq. ?? and this dummy history cannot provide any useful information about how to extracting memory that is useful to current task. Therefore, when replacing this dummy history with real history in the replay buffer, the extracted memory can be anything and will disturb the decision making, which makes us to expect that when evaluating the learned policy with history length longer than 0, the performance will be decreased. However, the results is surprisingly remained at the same level as that for LSTM-TD3(0). This reminds us that LSTM-TD3(0) may have learned a policy that intentionally ignores the history. To validate this conjecture, we plotted the average extracted memory of the actor in Fig. 5c. As shown in Fig. 5c, the average extracted memory of the actor of LSTM-TD3(0) initially starts with a non-zero value, but after a few thousands steps its value remains at a value around $0.0034 \pm 0.0024$, which is very close to 0, whereas for other LSTM-TD3s with longer history length the average extracted memories are relatively far away from 0 and have relatively large standard deviation. And this observation is consistent for both MDP and POMDPS. Even though we cannot claim the 0 in the extracted memory can be interpreted as neglect of past history, but at least we can say the history is uniformly mapped to a roughly fixed value rather than a random value for each history. In this way, even replacing the zero-value dummy history with a real history of experiences, the performance will not bad than that for a dummy history.

E. A Glance of The Relationship Among the Return, the Predicted Q-value, and the Extracted Memory of the Actor-Critic

The proposed LSTM-TD3 has been experimentally proved to be useful according to the results presented in Section II-A, but the understanding of the LSTM-TD3, especially the interpretation of the extracted memory, is still a big challenge. In Fig. 5, we presented the average test return, the average predicted Q-value, the average extracted memory of actor and critic to have a glance of the relationship among them.

By comparing Fig. 5a and 5b, we found that the average test return and the average predicted Q-value match approximately perfect, which means there is no overestimation problem as studied in [4], [5], [6] and is desirable.

Our special interest is in the relationship between the test return and the extracted memory of the actor-critic. It is worth to note that both the actor and the critic have a memory component and there is no sharing of the memory component, which means they may learn different coding of the memory that is helpful for learning a Q-value function and a policy, respectively. As shown in Fig. 5c and 5d, no matter if we comparing these plots horizontally or vertically, there is no consistent trend can be found. Horizontally, for both the actor and the critic, the average extracted memories for different versions, i.e. the MDP and the various POMDPS, of the task are showing different trends, where an exception is the extracted memory of the actor of LSTM-TD3(0) as discussed in the previous section II-D.1. Vertically, for different versions of the task, the extracted memories of the actor and the critic show different trends too. Especially, the average extracted memory of the critic of the LSTM-TD3(0) is not similar with that of the actor whose value is consistently close to 0. Contrarily, the average extracted memory of the critic of LSTM-TD3(0) is even further from 0 than that of LSTM-TD3(5). This may indicate the different roles of the memory component playing in the actor and the critic. Again, the interpretation of the extracted memory is not so straight-forward, and special constraints may be forced to improve the interpretability of the memory component, which is out of the scope of this paper and will be left for the future study.

F. Supplementary results for the Ablation Study

Fig. 6 shows the learning curves of various ablated LSTM-TD3 by removing different components, namely (1) using double critics (DC), (2) using target policy smoothing (TPS), (3) having current feature extraction (CFE) component, and (4) including past actions (PA) in the history. Table III reports the maximum average return of the investigated ablated algorithms.
Fig. 5: Relationship Among the Return, Predicted Q-value, Extracted Memory of Actor-Critic, where (a), (b), (c), and (d) shows the average test return, the average predicted Q-value, the average extracted memory of actor, and the average extracted memory of critic. The average extracted memory of actor and critic is an average over all output neurons of the memory component.

References

[1] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[2] I. Adamski, R. Adamski, T. Grel, A. Jędrych, K. Kaczmarek, and H. Michalewski, “Distributed deep reinforcement learning: Learn how to play atari games in 21 minutes,” in International Conference on High Performance Computing. Springer, 2018, pp. 370–388.

[3] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al., “Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation,” arXiv preprint arXiv:1806.10293, 2018.

[4] S. Thrun and A. Schwartz, “Issues in using function approximation for reinforcement learning,” in Proceedings of the Fourth Connectionist Models Summer School. Hillsdale, NJ, 1993, pp. 255–263.

[5] S. Fujimoto, H. Van Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” arXiv preprint arXiv:1802.09477, 2018.

[6] L. Meng, R. Gorbet, and D. Kulić, “The effect of multi-step methods on overestimation in deep reinforcement learning,” arXiv preprint arXiv:2006.12692, 2020.
Fig. 6: Learning curves of ablation study, where to ease the comparison only average values are plotted. In the legend, Full, Full−CFE, Full−PA, Full−DC, Full−TPS, and Full−DC−TPS correspond to LSTM-TD3 with full components, removing current feature extraction, excluding past action, not using double critics, not using target policy smoothing, and simultaneously not using double critics and target policy smoothing.
TABLE III: Maximum Average Return for Ablation Study. Maximum value of evaluated algorithms for each task is bolded, and ± indicates a single standard deviation.

| Task   | Algorithms   | LSTM-TD3       | Full–CFE       | Full–PA       | Full–DC       | Full–TPS      | Full–DC–TPS  |
|--------|--------------|----------------|----------------|---------------|---------------|---------------|--------------|
|        |              | 1223.0 ± 582.3 | 1182.2 ± 522.0 | 1469.6 ± 637.5 | 569.1 ± 78.5  | 920.6 ± 48.8  | 517.4 ± 102.0 |
|        | MDP          | 918.4 ± 44.0    | 680.8 ± 129.9  | 871.0 ± 43.6  | 589.5 ± 73.8  | 945.6 ± 187.5 | 552.0 ± 1.4   |
|        | POMDP-RV     | 848.1 ± 60.2    | 8447.1 ± 211.3 | 946.9 ± 242.1 | 836.3 ± 346.7 | 809.9 ± 42.7  | 690.8 ± 0.0   |
|        | POMDP-RK     | 771.6 ± 18.2    | 641.5 ± 40.2   | 742.9 ± 79.2  | 1222.2 ± 313.4| 808.3 ± 468.1 | 731.1 ± 330.9 |
|        | POMDP-RS     | 954.0 ± 362.9   | 457.0 ± 99.1   | 784.0 ± 39.7  | 958.1 ± 229.1 | 693.9 ± 79.0  | 606.3 ± 63.0  |
|        |              | 2574.9 ± 79.0   | 1510.8 ± 421.4 | 2421.0 ± 305.4| 2121.1 ± 334.6| 2658.6 ± 1537.1| 1855.8 ± 494.2 |
| HalfCheatPB | POMDP-RV     | 1932.7 ± 199.6  | 1133.2 ± 421.6 | 1535.0 ± 385.1| 1711.9 ± 331.5| 1814.3 ± 90.3  | 1068.6 ± 363.0 |
|        | POMDP-RK     | 2036.7 ± 73.5   | 1817.2 ± 323.0 | 1578.1 ± 466.2| 2023.9 ± 348.0| 2083.8 ± 159.9 | 2145.1 ± 107.2 |
|        | POMDP-RS     | 1906.1 ± 171.4  | 1705.7 ± 105.8 | 1287.4 ± 360.0| 1588.0 ± 548.0| 1885.8 ± 92.7  | 879.3 ± 446.0 |
|        |              | 1324.9 ± 313.6  | 1737.4 ± 310.3 | 1817.6 ± 66.4 | 1728.8 ± 494.0| 1730.0 ± 474.4 | 1831.7 ± 339.9 |
| AntPB  | POMDP-RV     | 1970.5 ± 38.5   | 1320.9 ± 329.2 | 1719.6 ± 431.1| 1526.3 ± 96.5 | 2011.7 ± 61.6  | 1381.9 ± 801.0 |
|        | POMDP-RK     | 1479.9 ± 283.2  | 520.2 ± 85.4   | 1100.9 ± 243.6| 1214.2 ± 313.7| 1871.7 ± 138.3 | 1349.3 ± 101.1 |
|        | POMDP-RS     | 1264.9 ± 338.5  | 1132.6 ± 256.2 | 1027.3 ± 214.5| 1017.1 ± 122.6| 1320.2 ± 275.9 | 731.5 ± 160.0  |
|        |              | 984.1 ± 267.7   | 8449.2 ± 282.2 | 803.5 ± 22.9  | 662.7 ± 155.2 | 1240.7 ± 228.2 | 576.0 ± 403.8  |
| Walker2DPB | POMDP-RV     | 841.2 ± 91.6    | 762.9 ± 53.0   | 896.1 ± 46.8  | 1107.0 ± 226.4| 820.8 ± 45.4  | 1010.2 ± 324.9 |
|        | POMDP-RK     | 2405.0 ± 158.9  | 1517.0 ± 673.8 | 2178.2 ± 300.7| 2504.7 ± 180.5| 2613.1 ± 81.5  | 2015.0 ± 477.6 |
|        | POMDP-RS     | 2233.6 ± 176.6  | 1187.0 ± 479.1 | 2188.9 ± 240.2| 1775.5 ± 471.6| 2303.9 ± 156.6 | 1852.9 ± 327.4 |
|        |              | 2264.6 ± 72.3   | 1459.0 ± 513.2 | 2055.9 ± 48.1 | 2087.9 ± 194.0| 2117.9 ± 103.1| 2083.3 ± 168.7 |
|        |              | 1635.8 ± 180.7  | 1143.9 ± 298.7 | 1279.7 ± 444.7| 1337.7 ± 262.1| 1450.7 ± 357.3| 1524.0 ± 516.0 |
|        |              | 1349.1 ± 405.7  | 1343.1 ± 411.3 | 1008.6 ± 34.4 | 1634.3 ± 386.3| 1769.5 ± 210.0| 1938.2 ± 0.0  |
| Horse2DPB | POMDP-RV     | 1000.0 ± 0.0    | 874.3 ± 217.9  | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  |
|        | POMDP-RK     | 752.6 ± 428.5   | 804.4 ± 246.2  | 752.6 ± 428.6 | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  |
|        | POMDP-RS     | 1000.0 ± 0.0    | 1000.0 ± 0.0   | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  |
|        |              | 1000.0 ± 0.0    | 1000.0 ± 0.0   | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  | 1000.0 ± 0.0  |
| HorseFourPB | POMDP-RV    | 9359.7 ± 0.1    | 8677.9 ± 749.7 | 9359.0 ± 5403.4| 9343.9 ± 5394.7| 9358.8 ± 9345.8| 9352.7 ± 5999.7 |
|        | POMDP-RK     | 9358.9 ± 0.3    | 9358.6 ± 1.0   | 5568.0 ± 491.9| 4748.1 ± 6616.0| 9357.7 ± 5402.7| 8974.0 ± 377.3 |
|        | POMDP-RS     | 9358.4 ± 1.1    | 9358.6 ± 11.1  | 9357.6 ± 1.2  | 9357.7 ± 5402.3| 9359.4 ± 0.0  |
|        |              | 1952.7 ± 503.4  | 1534.0 ± 462.1 | 1308.5 ± 1220.5| 2453.9 ± 1576.2| 2360.5 ± 1403.6| 2544.9 ± 3599.0 |

For POMDP-FK, $p_{fk} = 0.2$, For POMDP-RN, $\sigma_r = 0.1$, For POMDP-RSM, $p_{rs} = 0.1$.

CFE: Current Feature Extraction; PA: Past Action; DC: Double Critics; TPS: Target Policy Smooth