University of Moratuwa

Bsc. in Comp. Science and Engineering

Mid-Evaluation Report: Enhanced Sample Efficiency and Generalization in Hindsight Experience Replay (HER) for Robotic Manipulation

Name: M.K.P.A. Mallawarachchi

Index No: 210362X

1. Introduction and Project Scope

1.1 The Challenge of Goal-Conditioned Reinforcement Learning

Goal-Conditioned Reinforcement Learning (GCRL) is a pivotal subdomain of Deep Reinforcement Learning (DRL) focused on training universal policies capable of achieving diverse tasks, moving the field towards generalizable intelligence in complex environments, such as robotics.[1] The primary goal of GCRL is to learn a policy

that dictates action selection based not only on the current state but also on a dynamically provided goal .[3]

A critical challenge in applying GCRL to real-world tasks, particularly robotics manipulation, is the **sparse and binary reward problem**.[3] When the reward

is only upon exact task completion and (or) otherwise, standard DRL algorithms, such as Deep Q-Networks (DQN), frequently fail if the state space is large, as the probability of accidental success is negligible. For instance, in a bit-flipping environment, DQN without effective mitigation fails for tasks with more than bits, illustrating the exploration limitation in high-dimensional state spaces.[3]

1.2 Baseline Solution: Hindsight Experience Replay (HER)

Hindsight Experience Replay (HER) [3] provides an elegant and domain-agnostic solution to the sparse reward challenge by leveraging the off-policy nature of underlying RL algorithms, such as Deep Deterministic Policy Gradients (DDPG). HER allows the agent to learn from failed attempts by retrospectively relabeling the goal, utilizing the Universal Value Function Approximator (UVFA) framework.[3] By considering what goal

was achieved () rather than what goal *was intended* (), HER dynamically densifies the reward signal. Crucially, this mechanism avoids the need for complex, hand-engineered **shaped reward functions**, which often necessitate domain-specific expertise and may introduce local optima that compromise the metric the user truly cares about (binary success/failure).[3]

The current project establishes the DDPG+HER framework, applied to the challenging Fetch Robotics MuJoCo manipulation suite (Pushing, Sliding, and Pick-and-Place), as the State-of-the-Art (SOTA) baseline.[3]

1.3 Project Objectives and Enhancement Methodology

Although DDPG+HER is highly effective, as demonstrated by its ability to solve the Fetch tasks where DDPG alone fails [3], analysis reveals systematic opportunities for performance enhancement, specifically targeting sample efficiency and generalization capacity.

The systematic enhancement methodologies selected for this project, aligning with the "Training Strategy Enhancements" and "Architecture Modifications" stipulated for the Continuous Assessment [3], are:

1. **Prioritized Hindsight Experience Replay (PHES):** A training strategy enhancement to maximize the utility of sampled transitions and significantly boost sample efficiency.
2. **Goal-Conditioned Cross-Attention Policy Architecture:** An architectural modification to improve the structural robustness and generalization power of the policy network by explicitly fusing state and goal features.

This report serves as the mid-evaluation deliverable, detailing the theoretical grounding and preliminary experimental design for validating these two performance enhancement methods against the established DDPG+HER baseline.

2. Baseline Model Review and Performance Gaps

2.1 Formalism: UVFA and DDPG Integration

The HER algorithm relies on the principle of training universal policies and goal-conditioned value functions .[3] In the baseline configuration, this is achieved by concatenating the state

and the goal before feeding them into the network: .[3]

The chosen off-policy algorithm is DDPG [3], which is well-suited for the continuous action space required by the 7-DOF Fetch Robotics arm.[3] The off-policy nature of DDPG is fundamental, as it allows the agent to train the Critic (Q-function approximator) and Actor (policy

) using experiences that were generated under a different (behavioral) policy, thereby permitting the retrospective modification of the goal associated with an experience tuple stored in the replay buffer.

2.2 Mechanism and Evaluation of Hindsight Experience Replay

HER operates by generating and storing additional transitions in the experience replay buffer . After an episode produces a trajectory with initial goal , HER replays each transition not only with but also with a set of alternative goals .[3] The new goal

 is typically a state that was actually achieved during the trajectory. When a transition is replayed with , a corresponding non-zero reward is generated, ensuring that the episode, even if globally unsuccessful, provides locally positive learning signals.[3]

The selection of goals is governed by the sampling strategy . The original research compared several strategies, including final (using only the final achieved state), episode (random states from the episode), random (random states from the entire buffer), and future (random states achieved *after* the current timestep in the same episode).[3]

Evaluation across the MuJoCo Fetch environments demonstrated that the **future strategy, particularly with or additional replays per transition, performs optimally**.[3] This empirical result establishes that the learning signal is maximized when the hindsight goal

 is temporally close and incrementally reachable from the state being replayed. This temporal coherence ensures that HER functions effectively as an implicit curriculum, where mastering achievable, short-term goals rapidly bootstraps the policy towards mastering the more distant, original objectives. Conversely, strategies like random or simple final often yield learning signals that are too distant or noisy, leading to poorer performance.[3]

2.3 Identified Baseline Limitations

While HER dramatically improves learning possibility, two limitations inherent to the standard DDPG+HER configuration are addressed by the proposed enhancements:

2.3.1 Limitation 1: Inefficient Sample Utilization (Uniform Sampling)

Standard DDPG, and thus the DDPG+HER baseline, relies on sampling transitions uniformly from the replay buffer .[3] Although HER ensures that non-zero rewards are generated, the uniform sampling mechanism fails to capitalize on the fact that certain transitions—specifically those with high Temporal Difference (TD) error—provide significantly more information for improving the Q-function estimation than others.[7] High TD error indicates a substantial mismatch between the predicted Q-value

and the target value . Maximizing the reuse of these informative, rare transitions is essential for accelerating the learning rate and improving overall sample efficiency, especially in resource-constrained environments or when dealing with limited interaction time.[9]

2.3.2 Limitation 2: Suboptimal State-Goal Fusion for Generalization

The baseline policy and Q-function use simple feature concatenation .[3] This assumes that the subsequent dense Multi-Layer Perceptron (MLP) layers can implicitly learn the complex, non-linear dependencies between specific elements of the state (e.g., gripper

position, angular velocity) and the components of the goal (e.g., target object coordinates). For tasks requiring high precision and complex relative geometric reasoning, such as 3D Pick-and-Place, simple concatenation is architecturally inefficient and limits the agent's ability to generalize to novel goal configurations.[10] A more explicit mechanism is required to allow the policy to dynamically weigh which state features are most relevant to achieving the current, specific goal.

Table 1 summarizes the recognized limitations and the corresponding enhancement targets.

Table 1: HER Baseline Performance Profile and Enhancement Targets

Limitation Category	**Observed Evidence**	**Enhancement Target**
Sample Efficiency	Uniform replay sampling fails to prioritize informative hindsight transitions.[3]	Maximize reuse of high TD-error transitions using **Prioritization**.

3. Proposed Enhancement I: Prioritized Hindsight Experience Replay (PHES)

3.1 Synergy of HER and Prioritized Experience Replay

Prioritized Experience Replay (PER) is a training strategy enhancement that addresses the inefficiency of uniform sampling by prioritizing experiences based on their significance, typically measured by the magnitude of the TD error.[7] The TD error quantifies how much the agent learned from a particular transition.

PHES combines the reward-densifying capability of HER with the efficiency-boosting mechanism of PER. HER ensures the existence of valuable non-zero reward transitions (hindsight goals), while PER ensures that these highly informative transitions—both original successful transitions and critical hindsight transitions—are sampled more frequently, thus accelerating the convergence of the Q-function approximation towards the optimal .[3] The original HER work confirmed that PER is an orthogonal technique that can be readily combined with HER.[3]

3.2 Mathematical Formulation of Priority Calculation

For the DDPG critic, the objective is to minimize the loss , where is the target value.[3] The priority score

for a transition is derived from the magnitude of the TD error , calculated using the goal-conditioned Q-function:

Where is the target critic network (used for stability) and is the actor network.[3]

The Proportional Prioritization scheme will be employed, where the sampling probability is determined by:

Here, is a small offset to guarantee that no transition has zero probability of being selected, and $\alpha \in $ controls the degree of prioritization. Setting recovers uniform sampling.[8]

The entire replay buffer

is sampled non-uniformly based on the normalized probability .

3.3 Implementation Strategy and Bias Correction

The priority scores must be calculated immediately after the transition is collected and stored in the replay buffer. Crucially, in the PHES implementation, the TD error calculation must occur *after* the HER goal relabeling strategy (future with or) has been executed. This ensures that the prioritization reflects the informativeness of the hindsight transitions generated by HER, maximizing the value derived from those relabeled successes.

Non-uniform sampling, while beneficial for focus, introduces a bias to the estimate of the expected value of the loss function. To ensure that the asymptotic convergence properties of DDPG are maintained, Importance Sampling (IS) weights must be applied when calculating the gradient updates for the DDPG critic loss .[7]

Where is the total number of transitions in the buffer, is the sampling probability, and $\beta \in $ is the annealing exponent.[12]

is often annealed from an initial low value (e.g., 0.5) towards 1 over the course of training to mitigate initial instability, as prioritizing transitions early on can sometimes favor noisy or erroneous samples.[13]

3.4 Expected Impact and Theoretical Justification

PHES is expected to yield substantial gains in sample efficiency. By focusing computation on experiences that represent the largest discrepancy in current value estimates, the Q-function convergence is accelerated. This technique should significantly reduce the **Epochs to Convergence** required to reach a high success threshold (e.g., 90% Success Rate) and improve the metric of **Samples Per Second (SPS) effectiveness** when comparing learning curves based on total environment steps.[14] This focuses the computational effort on transitions offering the highest potential information gain regarding the optimal control policy.

4. Proposed Enhancement II: Goal-Conditioned

Cross-Attention Policy Architecture

4.1 Rationale for Explicit State-Goal Fusion

In the standard DDPG+HER model, the Universal Value Function Approximator (UVFA) relies on the concatenated vector . When dealing with high-dimensional states and complex spatial goals, as in the Fetch Pick-and-Place task (which requires intricate 3D positioning and gripper control), the simple concatenation mechanism forces the network to implicitly learn feature importance.[3] The architectural limitation becomes evident in generalization: the policy struggles to dynamically discern which specific state variables (e.g., object rotation, gripper velocity) are critical for achieving the current goal, especially when the goal changes.[10]

To overcome this, an explicit mechanism for dynamic contextual fusion between the state and the goal is required. Attention mechanisms, particularly cross-attention, are structurally designed to query the relevance of one set of input features (the state) against another (the goal), leading to a highly contextualized representation.[17]

4.2 Architectural Design: Integrating Cross-Attention

The proposed modification involves replacing the simple concatenation layer with a Goal-Conditioned Cross-Attention layer in both the DDPG Actor () and Critic () networks (Architecture Modification).

The process involves three sequential steps following the raw input:

1. **Feature Embedding:** The raw state vector is processed by an initial embedding MLP layer to generate a high-dimensional State Feature Vector, . Similarly, the goal vector is processed by an identical or shared MLP to produce a Goal Feature Vector, .
2. **Cross-Attention Fusion:** The core fusion occurs via a multi-head cross-attention layer. The mechanism uses the State Feature Vector as the **Query (Q)** and the Goal Feature Vector as both the **Key (K)** and **Value (V)**.
   This operation produces a Goal-Attended State Vector, . This vector is an enhanced representation of the state where each feature has been weighted by its importance relative to the input goal. This structural change embeds goal-contextual information directly into the state representation.
3. **Final Network Processing:** The resultant is typically combined (e.g., concatenated or added via a residual connection) with the original State Feature Vector . This enriched

vector is then passed through the remaining dense layers of the actor to output the action , or passed into the critic network along with the sampled action to output the Q-value .

4.3 Computational Implications and Generalization Gain

The primary advantage of the cross-attention fusion architecture is the resultant increase in generalization capability. The network learns a robust mapping that allows it to effectively transfer skills learned in one goal region to an unseen goal region because it explicitly learns feature relevance (e.g., "target is far away, so maximize force along the -axis") rather than relying on complex implicit feature interactions within dense MLPs.[11] This explicit modeling is expected to drastically improve the policy's robustness to randomized goal placements and complex geometry in tasks such as Pick-and-Place.

For the DDPG Critic network, the action input is introduced subsequent to the state-goal fusion via cross-attention. This structure ensures that the Q-function evaluates the value of an action taken from a state that has already been highly contextualized by the target goal .

5. Preliminary Experimental Design and Validation Plan

The proposed enhancements must be validated through a rigorous empirical framework that employs a controlled ablation study design, adhering to established RL research standards.[3]

5.1 Environment Selection and Baseline Configuration

- **Environment Suite:** The evaluation will use the three canonical Fetch Robotics manipulation tasks simulated in the MuJoCo physics engine: Pushing, Sliding, and Pick-and-Place.[3] These tasks involve continuous state and action spaces.
- **Reward Function:** Consistent with the baseline, the reward will be sparse and binary, , where if the object is within the goal tolerance .[3]
- **Baseline (E0) Configuration:** DDPG implemented with HER using the empirically validated future sampling strategy with additional replays per transition, as this configuration proved superior in the original ablation studies.[3] Standard DDPG

hyperparameters (e.g., discount factor
, Polyak update rate) will be maintained.[20]

5.2 Performance Metrics

The evaluation focuses on quantifying both the effectiveness and the efficiency of the enhanced models. All metrics will be averaged over five random seeds to mitigate inherent variance in DRL training.

1. **Effectiveness: Average Success Rate (%)**
   - Definition: The percentage of test episodes resulting in the object being within the task-specific spatial tolerance of the goal position at the end of the episode (e.g., 7 cm for Pushing/Pick-and-Place; 20 cm for Sliding).[3] This metric reflects the robustness and final quality of the learned policy.
2. **Efficiency: Samples Per Second (SPS)** and **Epochs to 90% Success Rate**
   - SPS measures the throughput efficiency of the training loop, reflecting computational overhead.[14]
   - Epochs (or total environment steps) to reach a threshold success rate (e.g., 90%) quantifies the true sample efficiency, demonstrating how quickly the agent learns compared to the baseline.

Note: Full implementation isn’t ready yet.

5.3 Hypothesized Ablation Study Design

A four-track ablation study is designed to isolate the contributions of PHES (E1) and Cross-Attention (E2) before testing their combined synergistic effect (E3).

Table 2: Ablation Study Design for Enhancement Validation

Experiment ID	**Replay Mechanism**	**Architectural Fusion**	**Primary Focus**	**Hypothesized Outcome**
E0 (Baseline)	Uniform HER (Future,)	Concatenation	-	0.90
E1 (PHES Only)	**Prioritized HER (PHES)**	Concatenation	-	0.90
E2 (Attention Only)	Uniform HER (Future,)	**Cross-Attention Fusion**	Generalization/ Robustness	Higher final success rate and stability, especially on Pick-and-Place.
E3 (Combined PHES-Attention)	**Prioritized HER (PHES)**	**Cross-Attention Fusion**	Combined Maximization	Fastest convergence and highest overall terminal success rate.

This systematic approach will provide empirical evidence detailing how each enhancement contributes to overall performance improvement, satisfying the requirement for quantifiable gain validation.[3] New hyperparameters introduced by PHES (

, ,) and Cross-Attention (, number of heads) will be tuned based on initial hyperparameter sweeps to ensure training stability and optimal performance.

6. Conclusion and Current Modified Research Trajectory

The DDPG+HER framework provides a robust baseline for solving complex GCRL tasks characterized by sparse rewards.[3] However, to achieve superior performance relevant to cutting-edge research, systematic enhancements targeting fundamental limitations in efficiency and generalization are necessary.

This project proposes two systematic and methodologically distinct enhancements: Prioritized Hindsight Experience Replay (PHES) and a Goal-Conditioned Cross-Attention Policy Architecture. PHES addresses the sample utilization gap by applying TD-error prioritization to the high-value hindsight transitions, maximizing the information gain derived from agent experience. The Cross-Attention architecture addresses the generalization gap by structurally improving the Universal Value Function Approximator (UVFA) to dynamically contextualize state information based on the current goal, moving beyond simple concatenation.

The synergy between these two enhancements is significant: PHES ensures the agent learns from its mistakes and successes as quickly as possible, while the Cross-Attention architecture ensures the learned policy is structurally robust and generalizes effectively across the entire goal space. The detailed four-track ablation study defined herein provides the necessary methodology for empirical validation, ensuring that the final research outcome demonstrates measurable performance gains, thereby meeting the stringent quality standards required for conference-ready submission.[3]

The immediate research trajectory involves the careful implementation of the PHES priority buffer and the Cross-Attention network components, followed by rigorous hyperparameter tuning and execution of the defined ablation study across the Fetch Robotics environment suite. The successful completion of this validation phase will lead to the authoring of a complete research paper detailing quantifiable gains in both sample efficiency and terminal success rate, achieving the stated project objectives.