An Analysis of the Impact of Dataset Characteristics on Data-Driven Reinforcement Learning for a Robotic Long-Horizon Task

Ingook Jang*, Samyeul Noh*, Seonghyun Kim*, and Donghun Lee*
*Autonomous IoT Research Section
Electronics and Telecommunications Research Institute
Daejeon, Korea
{ingook, samuel, kim-sh, donghun}@etri.re.kr

***Abstract*—Data-driven reinforcement learning (RL) is a cost-effective method for training agents without online interaction with the real-world environment. This approach involves collecting and storing data from various sources such as expert demonstrations or random policies, and learning from these datasets without further online interaction with the environment. However, learning an optimal behavioral model from offline data is challenging as it may not cover the entire state-action space. The paper discusses an experimental study analyzing how dataset characteristics impact the performance on a long-horizon robot manipulation task using a robotic arm. The goal of the paper is to provide guidance on strategically organizing datasets for training agents via data-driven RL.**

***Index Terms*—Data-driven reinforcement learning, robot manipulation, long-horizon task**

I. Introduction

In Reinforcement Learning (RL), agents are trained to interact with the environment and learn the skills needed to solve a given task. However, in terms of real-world application domains such as autonomous driving or industrial robotics, interacting with the real-world environment to train an agent is costly and likely to raise safety concerns. Instead of online methods that directly interact with the real world, data-driven RL methods can be more cost-effective and safer, where agents learn behavioral models from datasets collected from past experiences.

Data-driven RL (also call as offline RL or batch RL) [1] is a method that collects and stores data from traditional behavioral policies, expert demonstrations, etc. and learns from offline datasets without any further interaction with the environment. The dataset can be constructed in a variety of ways, including expert demonstrations, random policies in a given environment, or even the optimal policy for a task. However, even if the dataset is collected in a variety of ways, it is difficult to learn an optimal behavioral model from the data because it does not cover the entire state-action space. This means that the performance of a model learned through data-driven reinforcement learning can vary significantly depending on how the dataset is constructed.

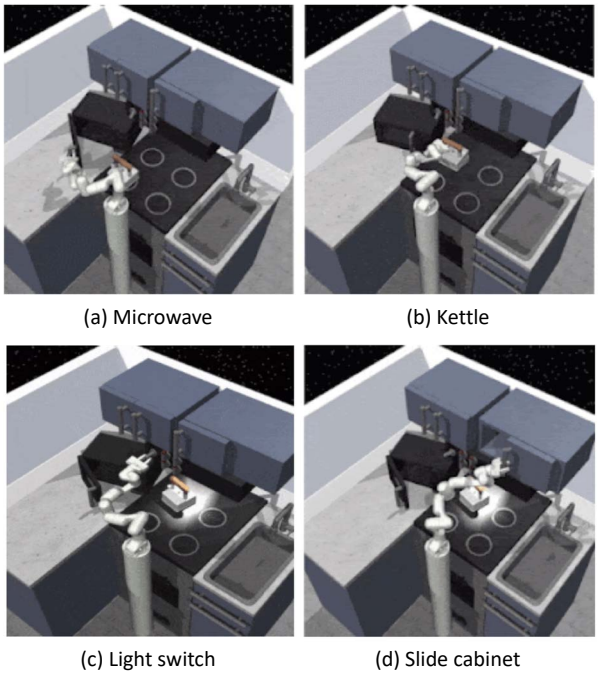

Fig. 1. Subtasks in Franka Kitchen

In this paper, we analyze how dataset characteristics (how the data is organized) affect performance in a robot manipulation task using a robotic arm through an experimental study. By comparing and analyzing the data-driven reinforcement learning performance of a robot in a long-horizon task consisting of a series of four smaller goal tasks, we aim to provide direction on how to strategically organize the dataset when training agents via data-driven RL.

I. Jang, S. Noh, S. Kim, and D. Lee are with the Autonomous IoT Research Section, Electronics and Telecommunications Research Institute, Daejeon, Korea, e-mail: {ingook, samuel, kim-sh, donghun}@etri.re.kr

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [23ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways]

II. Background

Reinforcement Learning. We utilize conventional RL framework, in which an agent engages with the environment to optimize the expected cumulative reward. In a more precise sense, during each time step t, the agent observes a state s_t, and based on its policy π, selects an action a_t. The environment provides the agent with a reward r_t and gives a transition to the subsequent state s_{t+1}. The primary goal of the agent is to maximize the expected return, denoted as $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0, 1)$ means the discount factor for learning.

Conervative Q-learning. CQL [2] is a recent data-driven RL algorithm that directly addresses the problem of overestimation of Q-values. CQL adds a Q-value regularizer to the policy evaluation objective of the Q function to ensure that the Q-value estimated by the policy $\pi_\theta(s, a)$ does not overestimate the Q-value of the data distribution $\mu(a|s)$, providing stable data-driven learning performance.

III. Method

We utilize the recently proposed D4RL [3] (Datasets for Deep Data-Driven Reinforcement Learning) for learning long-horizon tasks using a robotic arm. D4RL is a virtual environment benchmark that provides datasets in various environments for data-driven reinforcement learning, and has been actively utilized to validate the latest data-driven reinforcement learning technologies from various research groups.

In D4RL, Franka Kitchen is a learning environment that utilizes a robotic arm to interact with various objects to reach a desired goal state in a kitchen environment. Object interactions that the arm can perform include moving a kettle, turning on a light switch, opening and closing a microwave and cabinet door, and pushing a sliding cabinet door. In the Kitchen environment, the long-horizon task has four subtasks: open the microwave, move the kettle, turn on the light switch, and push the cabinet door. The reward function is designed to give a reward of 1 for each successful subtask, so the arm can earn a maximum of 4 rewards.

There are three types of datasets for training.

- The *"complete"* dataset contains human demonstrations that complete all four subtasks in order. Therefore, this dataset can be used effectively primarily for imitation learning.
- The *"partial"* dataset contains not only data that completes the four subtasks sequentially, but also data that performs subtasks other than the four subtasks. Therefore, when training with this dataset, the imitative learning agent can learn by selecting performance data that corresponds to the target four subtasks.
- The *"mixed"* dataset contains data that performs a variety of subtasks, but does not contain data that successfully performs all four target subtasks sequentially. This means that the dataset consists of suboptimal behavioral trajectory data for a given long-horizon task. Therefore, training with this dataset requires stitching (often referred to as "stitching") the robot arm's behavioral trajectory for each subtask. Training an agent on this dataset requires a high degree of generalization.

TABLE I
Major hyper-parameters for training

Hyper-parameter	Value
Epochs	800
Seed	8
Max trajectory length	280
Learning rate	3e-5
Discount factor	0.99
Batch size	256
Hidden layers	3
Layer width	256

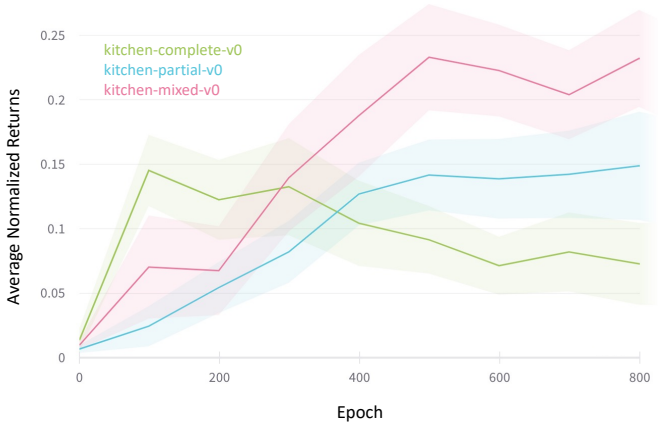

Fig. 2. Average normalized return

IV. Experimental Results

A. Implementation

In data-driven reinforcement learning, we utilized CQL to experiment with performance on the long-horizon task based on the characteristics of different dataset types. Table I shows the hyper-parameters used in our experiments. We train CQL agents for 800 epochs with different 8 seeds for each variant dataset.

B. Results

Under the same conditions, we evaluate and compare the performance of policy models based on data characteristics. Figure 2 shows the average normalized return results by dataset type. From the graph, we can see that the best learning performance is achieved using *"mixed"* datasets. This can be seen as a learning performance advantage of stitching in dynamic programming-based algorithms such as CQL when learning robot manipulation tasks with data-driven reinforcement learning algorithms.

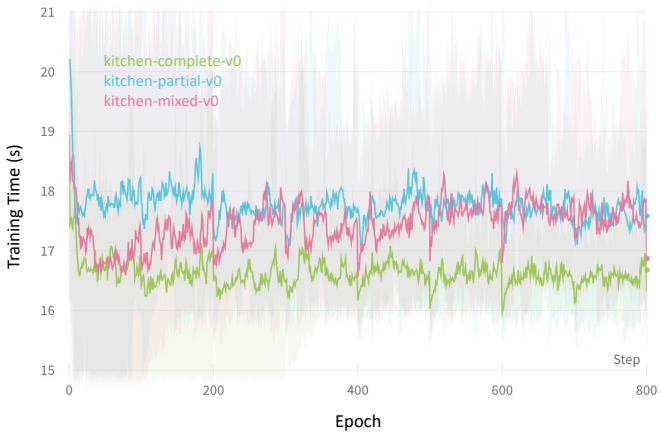

Fig. 3. Training time

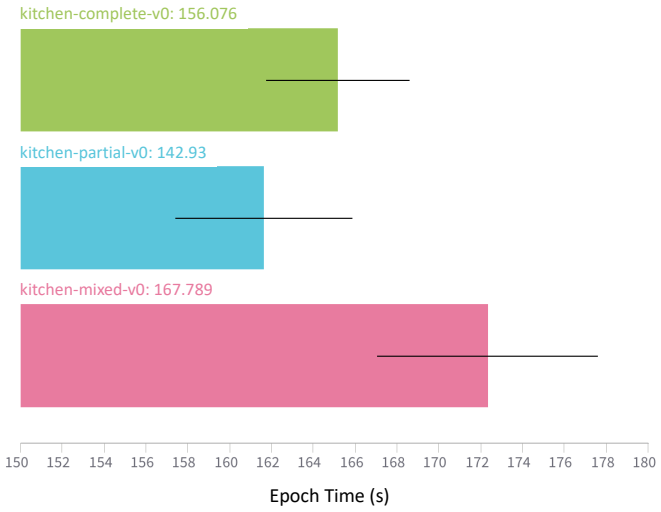

Fig. 4. Epoch time

On the other hand, we can see that the performance of data-driven reinforcement learning algorithms using a *"complete"* dataset consisting of data from human demonstrators drops significantly. The performance of data-driven reinforcement learning is somewhat worse when the dataset contains only human demonstrators who succeed in all robot manipulation tasks in a sequence. Human demonstrators do not follow a Markov decision process (MDP), which selects actions based solely on the current state. Also, when acquiring data from multiple humans, there are limitations such as inconsistent behavioral trajectories and significant variance in data quality and problem-solving strategies, which can increase the variance of the training dataset. For these reasons, when training with CQL on a *"complete"* dataset, the initially high performance of the training will decrease as the training progresses.

Figure 3 and 4 show the training time per epoch and epoch time, respectively. Each epoch time consists of the sum of training and evaluation time. The experiment on the *"complete"* dataset takes the least amount of time. In addition, we can see that the training time is longer for *"mixed"* and *"partial"* dataset than for *"complete"* data, even though the batch size is fixed at 256 for each type of dataset.

V. Conclusion

Reinforcement learning is a technique in which an agent learns a given task by interacting with its environment, but in real-world applications, data-driven reinforcement learning is emerging due to cost and safety concerns. In this paper, we experimentally analyze how the characteristics of the dataset affect the performance in a robot manipulation task. The experimental results show that utilizing mixed datasets is the best choice for learning to control robots in the Franka Kitchen environment.

References

[1] Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." arXiv preprint arXiv:2005.01643 (2020).

[2] Kumar, Aviral, et al. "Conservative q-learning for offline reinforcement learning." Advances in Neural Information Processing Systems 33 (2020): 1179-1191.

[3] Fu, Justin, et al. "D4rl: Datasets for deep data-driven reinforcement learning." arXiv preprint arXiv:2004.07219 (2020).