VCRLPD: Visual Cortex Reinforcement Learning with Prior Data

David Ihim
Stanford University
dihim@

Eric Werner
Stanford University
ewern@

Abstract

This project aims to develop a model-free framework, Visual Cortex Reinforcement Learning With Prior Data (VCRLPD), incorporating state-of-the-art approaches for retrieving visual representations, learning from offline demonstrations, and online reinforcement learning (RL) fine-tuning. The framework aims to achieve state-of-the-art results in challenging robot control domains, specifically focusing on the Franka Kitchen (Gupta et al., 2019) environment using image inputs and joint positions of a 9-DOF robot arm.

Currently, no model-free solutions exist for this problem. The objective is to outperform the VRL3 (Wang et al., 2023) framework, the current state-of-the-art, which utilizes inefficient RL methods and suboptimal pre-training. Our proposed framework replaces each of the 3 stages of VRL3 with a superior method. The encoder (which was ResNet18) is replaced with Aritical Visual Certex (VC-1) (Majumdar et al., 2023), leveraging a task-agnostic visual representation model tailored to robotics. We replaced the offline pre-training that used imitation learning with behavioral cloning . We also replaced the online fine-tuning with a SOTA online RL method (RLPD) (Ball et al., 2023).

Within this project, we accomplished evaluating the VRL3 framework within the Franka kitchen environment, a SOTA baseline for model-free frameworks. Next, we implemented our proposed framework VCRLPD to outperform VRL3 which uses visual foundation model to learn richer representations, for offline pre-training, we used behavior clonning, and for online fine tuning method, VCRLPD utilizes Reinforcement Learning Using Prior Data (RLPD).

Unfortunately, we weren't able to thoroughly test our proposed framework in the Franka kitchen environment due to technical issue with AWS Spot instances being lossed. This also resulted in us losing some graphs of our experiments of VRL3 learning from the different task of datasets (Complete, Partial). Nevertheless, we were able to retrieve the plot for the VRL3 on "mixed" dataset in the kitchen which resulted in on average 2 tasks being completed. After a million steps, VRL3 starts to acheive reward of 3. This is inefficient compare to model-based methods that solves the Franka kitchen.

1 Introduction

In recent years, the field of artificial intelligence has witnessed remarkable advancements in the area of reinforcement learning (RL), (Wang et al., 2023) a paradigm that enables intelligent agents to learn optimal decision-making strategies through interaction with their environment. Reinforcement learning has shown exceptional promise in solving complex tasks across various domains, including robotics, gaming, and autonomous systems.

Visual representation learning has emerged as a critical component of reinforcement learning systems, enabling agents to efficiently process and extract useful information from visual input sources. Traditionally, the focus of visual representation learning has been on designing handcrafted features or using unsupervised learning techniques such as autoencoders or generative adversarial networks. However, recent advancements have demonstrated the efficacy of learning visual representations directly within the RL framework. (Majumdar et al., 2023)

By leveraging large-scale offline datasets, agents can learn from a diverse range of experiences, providing a head start before engaging in online reinforcement learning. Moreover, imitation learning, a technique where agents learn from expert demonstrations, by mimicking the actions and behaviors of domain experts, agents can effectively transfer their knowledge and learn from their expertise. We examine the integration of imitation learning with visual representation learning and its impact on subsequent reinforcement learning performance.

Finally, we explore the crucial role of online reinforcement learning in fine-tuning and refining. Through iterative interaction with the environment, agents learn to adapt their visual representations to the specific task at hand. By combining these techniques, we aim to develop a comprehensive understanding of how model-free reinforcement learning systems can learn rich and effective training framework from online and offline learning.

2 Prior work / Motivation

2.1 Visual Reinforcement Learning (VRL3)

This paper introduces VRL3 (Wang et al., 2023), a data-driven framework for solving challenging visual deep reinforcement learning (DRL) tasks. The framework consists of three stages: task-agnostic visual representation learning using non-RL datasets, pretraining with offline RL data using this encoder, and fine-tuning through online RL. Compared to state-of-the-art methods, VRL3 achieves remarkable improvements in sample efficiency, with an average of 7.8x better performance on challenging hand manipulation tasks with sparse rewards and realistic visual inputs. Notably, VRL3 is 12.2x more sample efficient on the hardest task, solving it with only 10% of the computation required compared to previous approaches, demonstrating the great potential of data-driven deep reinforcement learning.

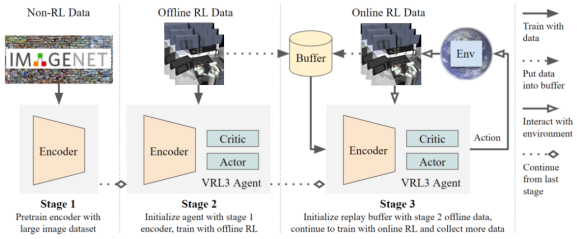

Figure 1: VRL3 stages

2.2 Efficient Online Reinforcement Learning with Offline Data (RLPD)

This paper addresses the challenges of sample efficiency and exploration in online reinforcement learning (RL). It investigates the use of existing off-policy RL algorithms with minimal modifications to leverage offline data for online learning. By applying these modifications, the proposed approach achieves reliable performance, leading to a substantial 2.5× improvement over existing methods across diverse competitive benchmarks, without incurring additional computational overhead.

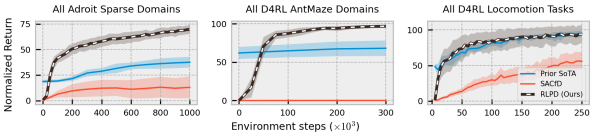

Figure 2: VRL3 stages

Specifically, RLPD (**?**) makes three simple but effective design choices. RLPD firstly uses "symmetric-sampling" to leverage offline data while training online, where 50% of the batch data is sampled from the replay buffer and 50% is from the online data buffer. Additionally, RLPD normalizes the critic update through a LayerNorm on the what?. Finally, RLPD uses large ensembles to improve sample efficiency how?.

2.3 Behavior Cloning (BC)

This paper introduces a two-phase autonomous imitation learning technique called behavioral cloning from observation (BCO) (**?**) to address these aspects what aspects? and improve overall performance. In the first phase, the agent acquires experience in a self-supervised manner, using it to develop a model. This model is then utilized in the second phase to learn specific tasks by observing an expert's demonstrations without access to explicit action knowledge. Experimental comparisons between BCO and existing imitation learning methods, including the state-of-the-art generative adversarial imitation learning (GAIL) technique, demonstrate comparable task performance across multiple simulation domains, with BCO exhibiting enhanced learning speed upon availability of expert trajectories.

2.4 Artificial Visual Cortex (VC-1)

VC-1 (Majumdar et al., 2023) introduces a novel artificial visual cortex, denoted as VC-1, which represents a significant advancement in perception modeling. Unlike previous models, VC-1 demonstrates remarkable versatility by supporting a wide range of sensorimotor skills, environments, and embodiments. The training of VC-1 is conducted using a pioneering dataset called Ego4D, curated by Meta AI in collaboration with academic partners, which consists of videos capturing individuals engaged in everyday tasks. Notably, VC-1 achieves comparable or superior performance to the state-

of-the-art results across 17 distinct sensorimotor tasks within virtual environments. Additionally, the paper proposes an innovative approach termed adaptive (sensorimotor) skill coordination (ASC). This method exhibits exceptional efficacy, achieving a near-perfect success rate of 98 percent, in tackling the demanding challenge of robotic mobile manipulation. The task involves navigating to a designated object, grasping it, moving to a different location, placing the object, and repeating the process. Importantly, ASC achieves this high level of performance in physical environments, further emphasizing its practical applicability.

3 Methods

We implement a similar strucutre to VRL3 and replace each components of VRL3 with different methods to test and try and improve the sample efficiency.

3.1 Stage 1: Encoder, VC-1

For the first part of our framework, we utilize the foundation model VC-1 for visual representation learning. The foundation model takes as input the 224 x 224 pixel image of environment and outputs 1x768 visual representation of the image. Encoders play a crucial role in reinforcement learning (RL) by transforming high-dimensional observations into compact representations, facilitating state abstraction and feature extraction. This enables RL agents to effectively handle complex and continuous state spaces, leading to improved decision-making and generalization. Additionally, encoders support transfer learning and domain adaptation, allowing RL agents to leverage prior knowledge and accelerate learning in new environments. Moreover, encoders enhance the scalability and flexibility of RL algorithms by reducing computational complexity and providing a modular component that can be easily integrated into the learning pipeline.

3.2 Stage 2: Pre-training, Behavior Cloning

Next, we pre-train the agent by completing behavior cloning from the demostations from the D4RL paper. We use the complete demostrations which concatenate the visual presentations with 9-DOF Franka Robot joint positions. We used the "complete" demostration data which is the comprehensive dataset consists of expert demonstrations covering all four target subtasks, completed sequentially in a predefined order.

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(\pi_\theta(a_i|s_i), a_i), \tag{1}$$

where N is the total number of demonstration samples, θ represents the parameters of the policy π_θ, s_i denotes the state at time step i, a_i is the action taken at time step i, and $\ell(\cdot)$ is the chosen loss function that measures the discrepancy between the predicted actions and the expert actions.

3.3 Stage 3: Online Finetuning, Reinforcement Learning with Offline Data (RLPD)

After the pretraining is completed, we finetune using online reinforcement learning. The method we utilize for this step is RLPD which makes important utilization of the data integrated during the offline pretraining. RLPD is built of SAC but includes 3 primary key design decisions which are attributed for its success.

3.3.1 Symmetric-sampling

The first design choice is sysmmetric sampling which takes 50% of the data in the batch comes from the offline directory and 50% of the data comes from the online fine tuning.

3.3.2 LayerNorm

The next design choice RLPD makes it to use layer norm within the critic to hinder overbearing estimations for that actions that aren't experience within the pretraining.

3.3.3 Ensembling

The last main design decision that RLPD makes is to use an ensemble of critics networks. This helps regularization the network from overestimation and provide a grounded estimation

4 Results

Our experimental findings based on running the baseline VRL3 framework on the Franka Kitchen environment revealed significant improvements in online sample efficiency through actor pretraining. During online training, the eval reward reach reached approximately 1.75 out of the 4 sub-tasks after 5x sooner when pretraing had been done versus not, or 20k versus 100k steps respectively.

We had multiple instances crash that were training models, so I will explicate my observations. We ran VRL3 on the complete and mixed datasets,

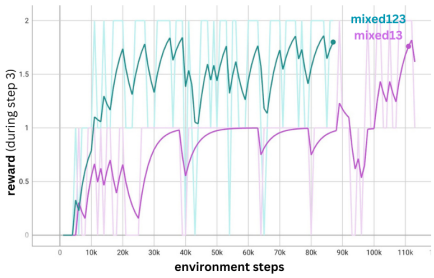

Figure 3: Mixed with and without stage 2

and additional ablated stage 2 pretraining in another train run with the mixed dataset. After 500k steps, the acgent trained on the complete dataset had achieved an eval reward of 2.12. The mixed dataset had achieved an eval reward of 2.0 with stage 2 and 1.75 without stage 2. After training further on mixed and complete to 1.5M steps, the agent trained from complete demonstrations was still stuck at 2.12 eval reward, whereas the agent trained on mixed demonstrations displayed an emergent jump to 3.0 eval reward by moving the tea pot, turning the knob, and opening the sliding cabinet. This emergent jumpt to an eval reward of 3.0 when trained on mixed data likely resulted from the mixed data covering more of the state and action space.

While the data was lost from training on a mixed dataset to 1.5M steps, we were still able to save a video. The Franka Robot was successfully able to solve 3 of the 4 tasks, achieving and eval reward of 3.0, indicating the potential for model-free solving of the kitchen environment. Notably, the reward had not yet converged at this stage, suggesting the possibility of further improvements through continued training and evaluation.

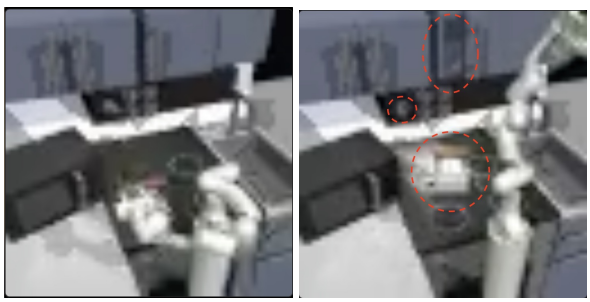

Figure 4: VRL3 completing 3 tasks after 1.5M steps

5 Conclusion

While we weren't able to complete the training of the full VCRLPD framework, we got see a current SOTA baseline for RL framework. We improved upon ResNet18 by using VC-1 to learn state representations. We also used RLPD and the improvements it offered to sample efficiency by augmenting online training with offline data.

References

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. 2023. Efficient online reinforcement learning with offline data.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. 2019. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. 2023. Where are we in the search for an artificial visual cortex for embodied intelligence?

Che Wang, Xufang Luo, Keith Ross, and Dongsheng Li. 2023. Vrl3: A data-driven framework for visual deep reinforcement learning.