Unsupervised Data Collection for Multi-task Robot Learning

George Hu[1], Jared Watrous[1], Jianhao Zheng[2]
[1]Deparment of Computer Science, Stanford University [2]Department of Civil and Environmental Engineering, Stanford University

Problem Definition

Goal – In the zero/few-shot expert demonstration setting, we leverage additional data from unsupervised reinforcement learning (RL) to more efficiently train multi-task offline RL algorithms.

- Given a small or empty set of expert demonstrations, what types of data on related tasks best augment the training dataset for learning a multi-task model?
- What unsupervised data collection methods are most effective for offline multi-task learning in various robot arm environments?

Key Contributions

Compare the performance of unsupervised autonomous data collection methods

Mix autonomously collected data with few-shot expert demonstrations for downstream offline RL.

Background and Related Work

Offline Reinforcement Learning (RL) Challenges

- Offline RL has historically relied upon task-specific expert demonstrations to fully decouple the **exploration/exploitation tradeoff** [5].
- However, this approach **fails to benefit from data scaling** of large diverse datasets and is **expensive in the multi-task setting** [5].

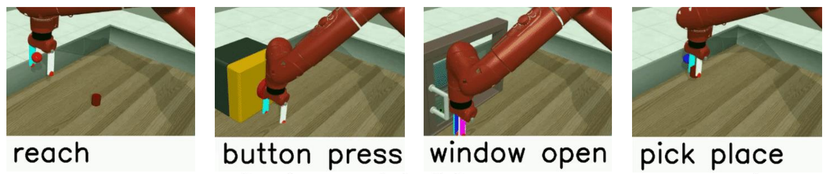
Unsupervised Data Collection

To collect data in a task-agnostic setting, various unsupervised approaches relying only on intrinsic reward have been proposed [5].

- Knowledge-based: **RND** [1] incentivizes exploring states with high novelty by visiting states with encodings unlearned by a student model.
- Data-based: **Hindsight GCRL** [2] samples goals and learns a goal-conditioned policy to visit those goals, ensuring the policy explores the entire state space.
- Skill-based: **APS** [4] and **CIC** [3] maximize the mutual information between explored states and latent skill vectors so that the data collection contains a diverse set of skills.

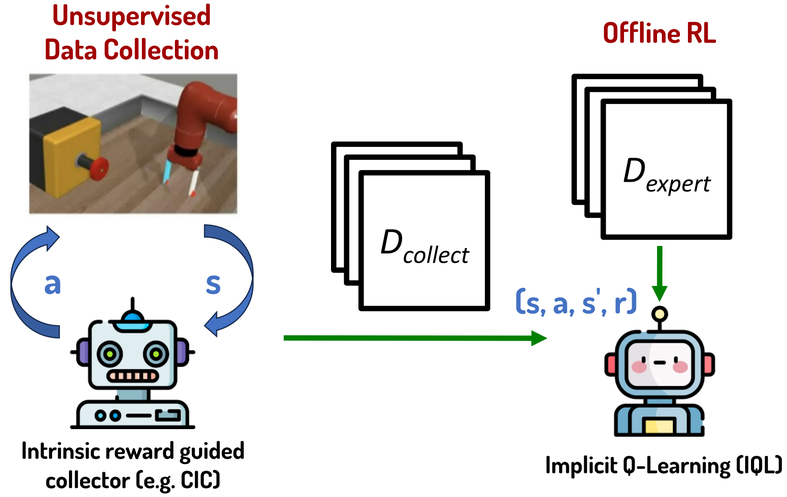
Meta-World

We use Meta-World as our simulator, which is an open-source simulated benchmark for meta-reinforcement learning and multi-task learning consisting of 50 distinct robotic manipulation environments [6]. We conduct experiments on the following 4 tasks:



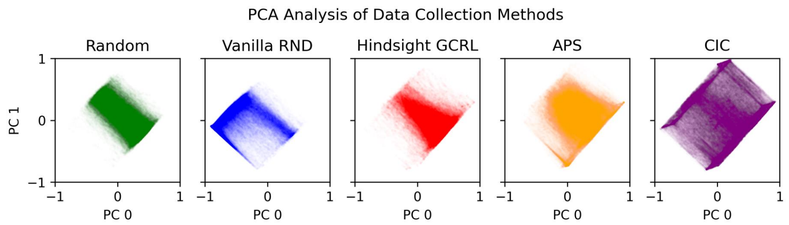
Observation: $\boldsymbol{o}^t = [\boldsymbol{s}_e^t, \boldsymbol{s}_{o1}^t, \boldsymbol{s}_{o2}^t, \boldsymbol{s}_e^{t-1}, \boldsymbol{s}_{o1}^{t-1}, \boldsymbol{s}_{o2}^{t-1}, \boldsymbol{x}_{goal}] \in \mathbb{R}^{39}$, where $\boldsymbol{s}_e^t$ and $\boldsymbol{s}_{\boldsymbol{oi}}^t$ is the state of the end effector and the **i**th object at time step t, and $\boldsymbol{x}_{goal}$ is the goal position.
Action: $\boldsymbol{a} = [\nabla \boldsymbol{x}_e, \ \tau] \in \mathbb{R}^4$, where $\nabla \boldsymbol{x}_e$ is the position gradient of end effector and τ is the torque that the gripper applies. **Reward**: $\mathrm{r} \in [0,10]$ is uniquely defined for different tasks.

Methods

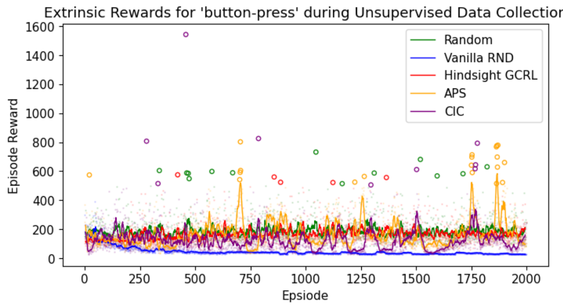


During data collection, we use an online algorithm that interacts with the environment but is **only guided by intrinsic rewards**. For example, in APS, the agent **maximizes the mutual information** $I(s;z) = H(s) - H(s|z)$ for states s and latent skill vectors z. The collected data is then **combined with few-shot expert demonstrations** to form the dataset used for IQL offline learning.

Data Collection



(Above) For 'button-press', CIC captures the whole environment while still densely exploring the center button location.



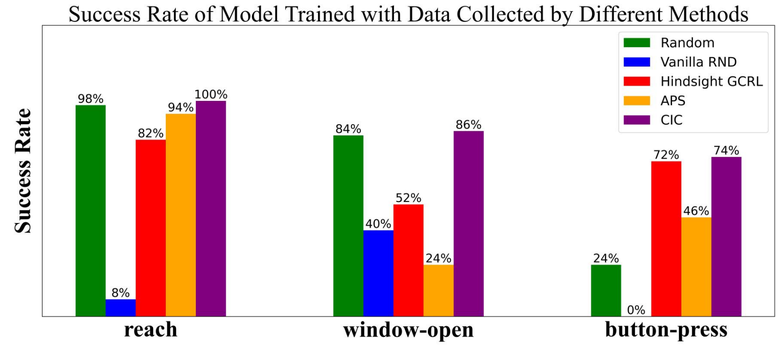
(Left) All algorithms except RND can discover extrinsic episode reward in 'button-press' over 500, but only CIC has higher than 1000.

The skills-based methods (APS, CIC) tend to provide more diverse data.

Experiments/Analysis

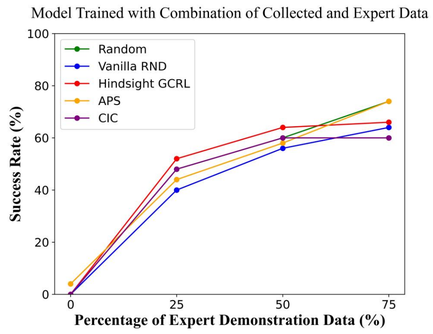
We use five methods **(Random, RND** [1], **Hindsight GCRL** [2], **APS** [4], **and CIC** [3]**)** to collect data and train an offline RL agent by **IQL** [7]. The trained models are evaluated on 50 subtasks with different goal positions per each task and the success rate are reported.

(Below) We compare the performance of the agent trained on only data collected by different methods (**no expert data**). CIC achieves the best performance on all tasks.



(Right) We report the results on the **pick and place** task where models are trained with a mixture of expert demonstration data and unsupervised method-generated data in varying proportions.

Combined with few expert data (25% and 50%), Hindsight GCRL and CIC outperform others. However, with more of the expert data (75%), their performances worsen, while APS and Random achieve the best performance.



Conclusion and Future Work

- Under a zero/few-shot expert demonstration setting, data collected by CIC can make offline RL training more efficient. With sufficient expert data, APS would be a better choice.
- We believe that investigating other tasks in Meta-World MT-50 with a more exhaustive set of algorithms can bring further insights in unsupervised RL and offline RL.
- In the future, we plan to analyze the collected data and the expert data, comparing their distributions and properties.

[1] Burda et. al., Exploration by Random Network Distillation, 2018.
[2] Endrawis et. al., Efficient Self-Supervised Data Collection for Offline Robot Learning, 2021.
[3] Laskin et. al., CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery, 2022.
[4] Liu el. al., APS: Active Pretraining with Successor Features, 2021.
[5] Yarats et al., Don't Change the Algorithm, Change the Data: Exploratory Data for Offline Reinforcement Learning, 2022.
[6] Yu et. al., Meta-World: A Benchmark and Evaluation for Multi-task and Meta Reinforcement Learning, 2019.
[7] Kostrikov et al, Offline Reinforcement Learning with Implicit Q-Learning, 2021.

gehu@stanford.edu jwatrous2002@stanford.edu jianhao@stanford.edu

CS330: Deep Multi-Task and Meta Learning