Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Curriculum Offline Imitating Learning]
Date: [30-10-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

The paper addresses the *quantity-quality dilemma* in offline imitation learning, where behavior cloning (BC) faces challenges in learning effective policies from datasets with mixed-quality trajectories. The main issues are:

- **Quantity Requirement**: Large amounts of data are necessary for stable BC performance.
- **Quality Requirement**: High-quality trajectories are sparse, making it inefficient for BC to learn from an entire mixed dataset.

Goal: To develop an offline imitation learning approach, *Curriculum Offline Imitation Learning (COIL)*, that adaptively selects data to imitate progressively better policies from mixed datasets, thus achieving high performance without requiring online evaluation.

2 Key contributions of the paper

- Proposed a curriculum-based offline imitation learning method, *Curriculum Offline Imitation Learning (COIL)*, that adaptively selects and imitates progressively better trajectories from mixed datasets.
- Developed a neighboring policy experience-picking strategy that enables the policy to imitate close-to-optimal trajectories at each curriculum stage.
- Introduced a return filter mechanism to ensure that only trajectories with returns above a threshold are used, improving stability and efficiency.
- Demonstrated competitive performance on continuous control benchmarks, where COIL outperforms traditional behavior cloning and rivals state-of-the-art offline reinforcement learning methods.

3 Proposed algorithm/framework

Algorithm 1 Curriculum Offline Imitation Learning (COIL)

Require: Offline dataset $\mathcal{D}$, number of trajectories picked at each curriculum N, moving window of the return filter α, number of training iterations L, batch size B, number of pre-train times T, learning rate η.

Initialize policy π with random parameter θ.
Initialize the return filter $V = 0$.
if $\mathcal{D}$ is collected by a single policy **then**
    Do pre-training for T times using BC.
end if
while $\mathcal{D} \neq \emptyset$ **do**
    for all $\tau_i \in \mathcal{D}$ **do**
        Calculate $\pi(\tau) = \{\pi(a_i|s_i), \pi(a_i|s_i), \ldots, \pi(a_i|s_i)\}$.
        Sort $\pi(\tau)$ into $\{\pi(a_0|s_0), \pi(a_1|s_1), \ldots, \pi(a_h|s_h)\}$ in ascending order, such that

$$\pi(a_i|s_i) \leq \pi(a_{i+1}|s_{i+1}), \quad j \in [0, h-1]$$

        Choose $s(\tau) = \pi(a_{\lfloor \beta h \rfloor}|s_{\lfloor \beta h \rfloor})$ as the criterion of τ_i.
    end for
    Select $N = \min\{N, |\mathcal{D}|\}$ trajectories $\{\tau\}_1^N$ with the highest $s(\tau)$ as a new curriculum.
    Initialize a new replay buffer B with $\{\tau\}_1^N$.
    $\mathcal{D} = \mathcal{D} \setminus \{\tau\}_1^N$.
    for $n = 1$ to $L \times N$ **do**
        Draw a random batch $\{(s, a)\}_1^B$ from B.
        Update π_θ using behavior cloning:

$$\theta \leftarrow \theta - \eta \nabla_\theta \frac{1}{B} \sum_{j=1}^{B} -\log \pi_\theta(a_j|s_j)$$

    end for
    Update the return filter $V \leftarrow (1 - \alpha)V + \alpha \cdot \min(R(\tau))^N$.
    Filter $\mathcal{D}$ by $\mathcal{D} = \{\tau \in \mathcal{D} | R(\tau) \geq V\}$.
end while

4 How the proposed algorithm addressed the problem

- **Adaptive Curriculum Learning**: COIL creates a curriculum by progressively selecting higher-quality trajectories from a mixed-quality dataset. At each stage, the policy is trained on a subset of trajectories with high cumulative rewards, which gradually increases the quality of data used for imitation.
- **Stable Behavior Cloning Updates**: By using behavior cloning on filtered, high-quality data, COIL maintains stability in training, which would otherwise be compromised by low-quality data.
- **Minimizing Quantity-Quality Trade-off**: COIL balances the need for a large dataset with the requirement for high-quality trajectories by gradually shifting the focus to higher-quality data as the policy improves. This allows COIL to leverage the entire dataset effectively over time without sacrificing policy performance.