Introduction

- Reinforcement Learning (RL) algorithms typically deal with only monotonic actions (discrete or continuous).
- However, in several pivotal real-world applications like drug discovery, the agent needs to reason over different types of actions.
- To accommodate these needs, we propose a novel framework: TAC (towered actor critic) architecture that aims to handle multiple action types.

Contributions

- Introduce a novel mechanism, towered actor critic to train agents on MDPs with multiple action types.
- TAC is applied to the task of reaction-based de novo drug design (where there are different action types at every time step) and show significant improvement over the existing state-of-the-art results.

Method

The TAC formulation is developed on top of our previous work for de novo drug design called Policy Gradient for Forward Synthesis (PGFS). PGFS is an off-policy algorithm having an actor module that consists of \( f \) and \( \pi \) networks. The \( f \) network takes in the current state \( s \) (reactant \( R(1) \)) as input and outputs reaction template \( T \). The \( \pi \) network uses \( R(1) \) and \( T \) to compute action \( a \) (feature representation for second reactant). Inside the environment, a KNN uses \( a \) and computes the \( k \) closest valid second reactants among which the maximum rewarding reactant is selected as \( R(2) \). \( R(1) \), \( R(2) \), and \( T \) are used to simulate a reaction and transition to the next state. PGFS uses TD3 for updating the actor \( (f, \pi) \) and critic \( (Q) \) networks.

procedure BACKWARD(buffer minibatch)

\[ T_{i+1}, a_{i+1}, R_{mask, i+1} \leftarrow \text{Actor-target}(R(1)_{i}) \]
\[ a_{i+1} \leftarrow a_{i+1} + \epsilon, \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \delta), -c, c) \]
\[ A_{i+1} \leftarrow \{T_{i+1}, R_{mask, i+1}, a_{i+1}\} \]
\[ P_i, Q_i, Q_{i+1} \leftarrow \text{min}_{i=1,2} \text{Critic-target}_i(R(1)_{i}, A_{i+1}) \]
\[ y_i \leftarrow r_i + \gamma Q_{i+1} \]
\[ P_i, Q_i \leftarrow \text{Critic}(R(1)_{i}, \{T_i, R_{mask}, a_i\}) \]
\[ L_{value} \leftarrow \sum_i [y_i - Q_i]^2 \]
\[ L_{critic} \leftarrow L_{value} + \alpha L_{auxil} \]
\[ L_{policy} \leftarrow -\sum_i \text{Critic}(R(1)_{i}, \text{Actor}(R(1)_{i})) \]
\[ L_{auxil-actor} \leftarrow -\sum_i (T_{i}^{(1)}, \log f(R(1)_{i})) \]
\[ L_{actor} \leftarrow L_{policy} + \beta L_{auxil-actor} \]

Conclusion

We propose a novel framework for handling multiple action types and demonstrated significant improvements over SOTA in de novo drug design. We further extended the framework to all gym environments and demonstrated better or comparable performance to TD3.