6.8200 Final Project Abstract: Diffusion Policy Evaluations **(Option C)**

Ram Goel, Franklin Wang

1 Overview

Diffusion Policy [1] is a recent method for sensorimotor learning in the context of offline reinforcement learning, where the model learns a policy based on demonstration data only. Diffusion Policies leverage DDPMs (Denoising Diffusion Probabilistic Models)—with key techinical differences—to more effectively model the many possible trajectories that the policy can take.

Behavior Cloning (BC) in RL is much more challenging than supervised learning since there is not necessarily a single correct ground truth solution. Instead, different trajectories might give equally valid solutions, requiring these models to have multimodal modeling capabilities. This is especially true when the action space is continuous, and we will be focusing on such environments in this work.

Early methods for BC relied on explicit policies, such as directly predicting a scalar value in a regressive manner or allowing for the modeling of multimodal distributions through binning or Gaussian mixture models. More recent work such as Behavior Transformers (BET) [3] have continued to improve upon this to allow for better multimodality modeling.

Another approach uses implicit policies, which evaluates the an energy function $E_\theta(o, a)$ for a given observation and action pair. It then finds the action which minimizes this energy function for a given observation, potentially allowing for more complex multimodal distributions to be modeled. This approach is implemented in Implicit Behavior Cloning (IBC) [2].

The Diffusion Policy builds upon all of this work by suggesting another approach that has even better multimodal modeling capibilities. DDPMs are diffusion models which generate data from noise by training a denoising process on an existing dataset. Diffusion Policy uses the basic idea of a DDPM but with key technical differences, and adapted to an RL context. First, the states and actions are encoded into a latent space. When the inputs are images, this is done by training a ResNet on the concatenation of the latent embeddings of various camera angles. Let $\mathbf{O}_t$ be a clipped set of previous observations, let k denote the number of denoising iterations, and let $\mathbf{A}_t^k$ denote the action taken given the initial set of observations. We have a noise prediction network $\varepsilon_\theta(\mathbf{O}_t, \mathbf{A}_t^k, k)$ to predict the noise from an arbitrary action prediction given this set of previous observations. We train this neural net on expert data by minimizing the difference between noise ε^k and the noise prediction of data added to ε^k. Finally, the conditional action distribution can be found through recursively applying the Langevin method on the score function [4], which in turn can be found as the negative of the noise prediction network.

2 Hypothesis

We will be comparing the performance of the Diffusion Policy to the following baselines:

- More recent approaches: Implicit Behavior Cloning [2] and Behavior Transformers [3]
- Simpler baselines: Vanilla behavior cloning (with Gaussian and Gaussian mixture policies), Nearest neighbors

Based on past research, we predict that Diffusion Policy has higher (1) overall performance as measured by the reward, (2) training stability, and (3) improved ability to model multiple valid trajectories when compared to IBC and BET [1].

However, we theorize that since the Diffusion Policy method was intended for environments where the observation is an image capturing the scene, if we use a simpler environment (such as where the observation is a low dimensional-vector instead of an image) then it will work about the same or worse than simpler baselines. This is because simpler baselines are already effective enough to model the environment, and the additional complexity of the Diffusion Policy will make it take much longer to train or train more poorly.

3 Experiments

We will implement the Diffusion Policy method and the simpler baselines.

For the experiments, we will be using the Push-T simulation environment for which the Diffusion Policy has been shown to work well [1]. This environment involves controlling a robot hand to push a T-shaped object to a specific position and orientation.

There are two variations of this environment: state-based and vision-based. The state-based version has observations of the form $(x_{agent}, y_{agent}, x_{block}, y_{block}, \theta_{block})$ while the vision-based version has observations consisting of a 96 by 96 pixel RGB image of the scene from a 2D top-down view. In both cases, the action space consists of all (x, y) coordinates, and each such action represents the desired (x, y) coordinate for the hand to move to.

We will then evaluate the Diffusion Policy and all the baselines on both variations of the environment. For the state-based environments we will adapt the baselines to work with flattened vector observations instead of images. We will compare the baselines based on the success rate (success is when the block within a set distance and angle to the correct location and orientation) and the average Intersection over Union (IoU) between the target block outline and the final outline of the block. We will also analyze the training stability and multimodality modeling effectiveness by viewing the training curves and trajectory visualizations, respectively. If the Diffusion Policy is better, we would expect it to have a smooth training curve and execute all the different valid trajectories for solving a given Push-T instance. On the other hand, other baselines may have large jumps in the training curve and miss out on certain solution modes.

4 Why does it matter

Diffusion Policy is able to model multimodal distributions much better than previous methods. Thus, when there are various ways to achieve a goal, the Diffusion Policy is able to better construct a multimodal action distribution which can find several such solutions, which is an extremely important problem within offline RL. This also allows for other possible benefits such as more stable training, making hyperparameter tuning and checkpoint selection easier and more accurate. Therefore, the Diffusion Policy approach is a promising step in multimodal modeling, and we would like to replicate their results to provide further evidence of its effectiveness or seek out limitations compared to other baselines.

References/Relevant Literature

[1] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2023.

[2] Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning, 2021.

[3] Nur Muhammad Mahi Shafiullah, Zichen Jeff Cui, Ariuntuya Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone, 2022.

[4] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020.