Policy Guided Diffusion: from Oxford, notes

Summary

The paper comes up with the idea of *policy guided diffusion*. Basically, this means that the authors used a diffusion model to determine best next steps based on probabilitstic means. Imagine a regular distribution of decisions, with both high probability decisions and also synthetic decisions. This allows you to make decisions and alternatives to offline world models. This allows us to generate synthetic training data without necessarily leading to offline datasets or biased sets.

What is a key obstacle to real-world adoption of RL?.

Sample Inefficiency; if data collection is slow, or data is hard to collect, you can't make strong RL decisions!

This problem is amplified if the exploration step is dangerous! Think a maze solver algorithm with traps, but if you die you can't actually start again.

how does Offline RL work? Part 1.

If you have an offline data set, without access to the environment, you can optimize a policy just using that dataset.

Before you can even think about this, we need to ask what makes machine learning work?

What makes modern machine learning work?.

You need large data sets that are broadly representative of the real situation that the model is meant to handle and very large and high capacity deep neural network models that can squeeze out all the knowledge contained in that data set and make accurate predictions on new unseen inputs. From UC Berkeley.

Offline RL, again.

Big models matter a lot. The larger the model, the better they work. However, dataset size may matter more! Think MNIST, ImageNet.

Decision making is also harder than prediction, because a decision influences another decision! Decision making is not independent, and so you can't assume independence.

This gets to the crux of the issue: Learning from data works, so can we do that in RL?

What is the offline RL workflow?.

1. Collect a dataset using any policy or a mixture of policies. Think humans performing the task, existing systems, random behaviours, etc.
2. Run offline RL on this dataset to learn a policy. A policy is just a mapping from states to actions, or just a deployment plan.
3. Deploy the policy in the real world. If the policy goes haywire, modify the algorithm and go back to step 2, *reusing* the data! (Levine, Kumar, Tucker, Fu)

Here's another question to begin the new page..

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Summary

And another summary that will float to the bottom of the next page.