Week 4 Paper Review

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

This is a brief review of [1], [2], [3], [4].

1 Plan To Predict: Learning an Uncertainty-Foreseeing Model For Model-Based Reinforcement Learning

The authors present Plan to Predict (P2P), a novel approach to Model-Based Reinforcement Learning (MBRL) that claims to increase model performance and sample efficiency by anticipating and actively quantifying the uncertainty of future predictions. This is achieved during the P2P model training step where the current policy is fixed and considered as the dynamics, while the learnt environment model acts as a sequential decision-maker which is updated to minimize the accumulative prediction error along the generated trajectories after interacting with the policy (dynamics). The idea considers the long-term effects of immediate predictions, thereby avoiding uncertain regions in the future.

1.1 Practical Implementation

The authors also show 2 implementations of P2P learning:

- P2P-MPC which uses the current policy to plan and optimize for an action sequence and a neural network trained on the environment dataset for the reward function.
- P2P-RL which trains the model on the environment dataset due to high-sample complexity and treats the model learning process as an offline RL problem.

1.2 Experiments

P2P-MC outperforms several baselines (SAC, MBPO) followed by P2P-RL (due to unstable parameter learning) on the most challenging MuJoCo tasks. Moreover, ablation studies provide empirical proof for increased model performance due to the minimization of the multi-step P2P loss compared to simply summing up the multi-step errors on trajectories collected by any other policies. This emphasizes the importance of active interactions between the model and current policy to mitigate uncertainty. Overall, P2P improves the quality of model-generated trajectories by reducing the accumulative model error along these trajectories.

Preprint. Under review.

2 Imagination-Augmented Agents for Deep Reinforcement Learning

The authors propose a novel deep RL architecture - Imagination-Augmented Agents (I2A) that enhances the learning of model-free (MF) algorithms.

2.1 Architecture

I2A architecture consists of 3 modules:

- Imagination Core (IC) is used to produce N imagined trajectories which is a sequence of features (predicted output along with it its corresponding reward) by using an environment model (EM) conditioned on an action sampled from a rollout policy. EMs are generally pretrained recurrent networks learnt with an unsupervised approach from agent trajectories.
- A Rollout Encoder is used to encode each trajectory separately as a whole and extract useful information for the agent's decision or ignore any noise. This information is used in addition to the predicted reward sequence in order to subdue erroneous predictions.
- A Policy module is a network which combines information from its model-free (a network with only real observations as inputs) and model-based (imagination-augmented) paths to output the imagination-augmented policy and its estimated value. The model-based path is an aggregation of several rollout encodings.

2.2 Experiments

I2A outperforms various baselines such as a standard MF agent, an agent with no EM, and an agent with no reward on the Sokoban task. I2A also performs well with a flawed EM by ignoring the latter part of the rollouts where the error accumulates because of the rollout encoder. Furthermore, I2A also showed generalizing capabilities for several tasks in the same environment.

The trade-off in all cases, however, was better performance for higher computation power.

3 When to Trust Your Model: Model-Based Policy Optimization

The paper starts with analysing a model-based (MB) RL algorithm with a monotonic improvement at each step which suggests no model usage at all. However, an empirical estimate of model generalization is incorporated which advocates for short model-generated rollouts branched from real data.

3.1 MBPO

To this effect, the authors propose Model-Based Policy Optimization (MBPO) framework that uses the predictive model for policy optimization only for short rollouts and otherwise relies on model-free (MF) policies, thereby limiting accumulated error. The framework has several design choices:

- The predictive model is a bootstrapped ensemble of dynamics. Bootstrapping accounts for regions with uncertainty in the case of low-quality samples and avoids overfitting due to model exploitation.
- SAC is used for policy optimization.
- For the predictive model usage, a branching strategy is employed that replaces a few long rollouts from the initial state distribution with many short rollouts starting from replay buffer states. This helps in mitigating the problem of induced compounding model errors for long rollouts and the policy's dependence on them during its updation.

3.2 Experiments

MBPO was run on several MuJoCo continuous tasks and it outperformed the state-of-the-art at the time namely SAC, PPO, STEVE, and PETS. It has asymptotic performance comparable to the best MF algorithms, a faster learning rate, and scales well to long horizons due to single-step rollouts as opposed to prior works.

4 Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning

The authors show how neural networks as a dynamics model achieve significant sample efficiency when used in MBRL and in combination with MF approaches for various contact-rich simulated locomotion tasks.

4.1 Model-Based Deep RL

The idea is to use a deep neural network as a dynamics model that predicts the change in state over a time step along with a reward function that encodes specific tasks. A model-based controller (MPC) is used to generate an optimal action sequence from this dynamics. Reinforcement Learning is further applied to improve its performance by preventing a mismatch between the data's state-action distribution and the model-based controller's distribution. This combination of a neural dynamics model along with a controller is trained only once, however, by changing the rewards function, several tasks can be accomplished without task-specific retraining.

However, due to high model bias, its performance was still lesser than purely MF algorithms.

4.2 Hybrid MB-MF

In order to augment the high task-specific performance of MF approaches by incorporating high sample efficiency from MB methods, the same MB algorithm (deep neural network dynamics model + MPC) is used to initialize a model-free learner (TRPO). MPC provides good rollouts, which enable supervised initialization of a policy that is fine-tuned with model-free algorithms.

4.3 Experiments

In the case of MBRL, aggregated on-policy rollouts from reinforcement learning improve sample efficiency. However, MPC for very short or large horizons is detrimental to task performance and the final rewards are subpar compared to pure SOTA MF approaches. However, hybrid MB-MF augments the high performance of pure MF methods due to increased sample efficiency because of its MB counterpart. Therefore, the paper showcases a potential method that alleviates the problem of achieving high success on complex tasks solely due to a large amount of data (which is difficult to collect in real time).

References

[1] Zifan W., Chao Y., Chen C., Jianye H., & Hankz H.Z. (2022). Plan To Predict: Learning an Uncertainty-Foreseeing Model For Model-Based Reinforcement Learning. Advances in Neural Information Processing Systems. https://openreview.net/forum?id=L9YayWPcHA_

[2] Weber, T., Racanière, S., Reichert, D. P., Buesing, L., Guez, A., Rezende, D. J., Badia, A. P., Vinyals, O., Heess, N., Li, Y., Pascanu, R., Battaglia, P., Hassabis, D., Silver, D., & Wierstra, D. (2017). Imagination-Augmented Agents for Deep Reinforcement Learning. ArXiv. https://doi.org/10.48550/arXiv.1707.06203

[3] Janner, M., Fu, J., Zhang, M., & Levine, S. (2019). When to Trust Your Model: Model-Based Policy Optimization. ArXiv. https://doi.org/10.48550/arXiv.1906.08253

[4] Nagabandi, A., Kahn, G., Fearing, R. S., & Levine, S. (2017). Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning. ArXiv. https://doi.org/10.48550/arXiv.1708.02596