Paper Review: MODEL-ENSEMBLE TRUST-REGION POLICY OPTIMIZATION

Summary:

The paper discusses the limitations of model-based reinforcement learning when using deep neural networks to learn both the model and policy, highlighting issues related to instability during training due to exploitation of regions with insufficient data for model learning. To address this, the authors propose using an ensemble of models to maintain model uncertainty and regularize the learning process. They also introduce the use of likelihood ratio derivatives for more stable learning compared to backpropagation through time. The resulting approach, called Model-Ensemble Trust-Region Policy Optimization (ME-TRPO), is shown to significantly reduce sample complexity compared to model-free deep RL methods on challenging continuous control benchmark tasks.

Contributions:

The paper's main contribution is an approach that significantly reduces sample complexity compared to state-of-the-art methods while achieving similar performance levels. The authors analyze the limitations of vanilla model-based RL, suggesting issues related to model bias and numerical instability. They evaluate the importance of key components of their algorithm, including using TRPO (Trust Region Policy Optimization) and model ensembles, and confirm the effectiveness of using model uncertainty to reduce model bias.

Strengths and Weaknesses:

The main strength of this paper is that it presents a simple and robust model-based reinforcement learning algorithm that significantly reduces sample complexity while achieving similar performance as model-free methods in challenging domains. However, I believe the main weakness of this approach is that it may require careful tuning of hyperparameters, and further investigation is needed to explore how to effectively use the model ensemble to encourage policy exploration in areas where the different models disagree.

Experimental Validity:

The paper compares their proposed method with several state-of-the-art reinforcement learning algorithms, including TRPO, PPO, DDPG, and SVG, in terms of sample complexity and performance. They find that prior model-based methods perform worse compared to model-free methods, are difficult to train over long horizons, and exhibit instability. In contrast, their proposed method achieves similar performance as model-free approaches with significantly less data, making it the first purely model-based approach to optimize policies for high-dimensional motor-control tasks such as Humanoid.

How can this work be extended:

Future directions for research include exploring how to use the model ensemble to encourage policy exploration in regions where models disagree and applying the approach to real-world robotics systems.