Paper Review - "*VIME: Variational Information Maximizing Exploration, Houthooft et al, 2016*"

Summary: The authors try to tackle the problem of exploration of RL algorithms with a curiosity-based approach, making use of Information Gain. The goal is to plan for maximum expected information gain i.e. reduce the uncertainty in the model dynamics or plan to be surprised. VIME does this by adapting the reward function. They achieve this by adding Information Gain (IG) to the original reward i.e. the difference between the distribution over the model parameters given the current trajectory and the next state-action pair.

Main Contributions: The authors mentioned that the practical issue with maximizing IG for exploration is the computational complexity of realized exploitation-exploration trade-off. They present a practical implementation of measuring IG using variational inference. Here, the agent's current understanding of the environment dynamics is represented by a Bayesian Neural Network which maximizes the variational lower bound. They then use stochastic gradient variational Bayes (SGVB) for optimizing it. IG of new transition samples is measured by measuring the KL-divergence between the original & updated parameter distributions. This KL-divergence is interpreted as Information Gain.

Strengths: VIME modifies the MDP reward function, and is not just limited to TRPO as described in the paper but can be applied with several different RL algorithms. Another strength is that it achieves significantly better performance compared to heuristic exploration methods across a variety of continuous control tasks and algorithms, including tasks with very sparse rewards.

Weakness: It is more computationally expensive which is also mentioned in the paper. This may limit its applicability in scenarios where computational resources are limited. It is slower as it attempts to learn a single policy that visits many states. It fails in scenarios where no extrinsic rewards are required and the model can only be trained on intrinsic rewards.

Experiments: The authors present broadly four experiments. First, they show that VIME performs better than naive exploration techniques like only Gaussian noise and l^2 BNN prediction as intrinsic rewards on problems like MountatinCar, HalfCheetah, etc. Second, they show that VIME is capable of achieving promising results on a highly difficult hierarchical task like SwimmerGather where none of the RL methods with naive exploration performed well. Third, they show that VIME is compatible with different RL methods also like TRPO, REINFORCE, and ERWR on different environments and helps them to converge faster and also avoid converging to the local suboptimal minima. Lastly, they investigate how η behaves for exploration-vs-exploitation trade-off. Higher η values cause more exploration and lower cause exploitation, closer to the baseline algorithm.

Extensions: VIME was only made to explore the continuous state and action space but it wasn't compared to discrete action spaces. It will be interesting to explore if it helps the model converge faster in discrete state and action spaces too or just increase the computation complexity. Another work can be around investigating and measuring the surprise in the value function and using the learned dynamics model for planning.