CS7011 : Topics in Reinforcement Learning
Summary of World Models : David Ha and Jürgen Schmidhuber

Sahana Ramnath : EE15B109

Abstract

World Models (Ha & Schmidhuber, 2018) proposes building generative neural network models of environments for reinforcement learning. The aim of this model is to learn a compressed spatial and temporal representation of the environment which is then used by the controller to take actions in the environment. The environment is modelled using a large network(the 'world model') which can capture its complexities, and the controller is modelled using a much smaller network which will allow it to focus on the task at hand. Experiments on the Car-Racing domain have clearly outperformed the previous scores. Experiments were also conducted in Vizdoom, where the agent was trained inside its own hallucinated dream generated by its world model.

1. Agent Model

The Agent model has 3 components :

1. Visual Sensory Component(**V**) : A variational auto encoder(VAE) is used to learn an abstract,compressed *spatial* representation of observed input frame z.
2. Memory Component(**M**) : This component gives a compressed *temporal* representation of the environment. It predicts the future z vectors that **V** is expected to produce. Accounting for the stochasticity of the environment, **M** gives a probability density function $p(z)$ instead of a deterministic z. In this paper, **M** is implemented as a Mixture Density Network on top of a Recurrent Neural Network(MDN-RNN).It models $P(z_{t+1}|a_t, z_t, h_t)$ where a_t is the action taken at time t and h_t is the hidden state of the RNN at time t.Model uncertainty is controlled using a temperature parameter τ.
3. Controller (**C**) : **C** determines the policy to be followed in order to maximize the expected cumulative reward in an episode. **C** is made as small and simple as possible, and is modelled as a single layer linear model that predicts action to be taken as $a_t = W_c[z_t h_t] + b_c$.

2. Training

V and **M** are trained in an unsupervised manner whereas **C** is trained in a supervised manner using rewards from the environment.

- $10,000$ random rollouts collected from the environment are used to train **V**. Difference between each frame and its reconstructed version is minimized.
- The trained **V** predicts z_t which is used along with a_t from the rollouts to train M to model $P(z_{t+1}|a_t, z_t, h_t)$ as a mixture of gaussians.
- The parameters of **C** are optimized using Covariance Matrix Adaptation Evolution Strategy (CMA-ES) since number of parameters is very less.

3. Experiments

1. Car Racing Environment : A trained **V** and **M** are used to predict z_t and h_t which the controller uses to take actions in the actual environment. The agent achieved a score of 906 ± 21, which surpasses all previous entries on the leadeboard.
2. Vizdoom : For Vizdoom, the agent was trained inside its own dream generated by its world model, after which the policy was transferred back to the original environment. Since the world model cannot generate the original rewards, the cumulative reward is defined as the number of steps the agent stays alive during a rollout.Here **M** predicts whether the agent will die at $t + 1(\ d_{t+1})$ in addition to z_{t+1}.
   - Since the world model is only an approximate probabilistic model of the environment, occasionally it generates trajectories which are not possible in the actual environment.
   - Since **C** has access to the hidden states of the model and not just the game observations, the agent can further manipulate in the dream to increase cumulative reward in ways that won't work on the actual environment.
3. For more complex games, the paper proposes that the same model can be used with a slight change to **M**. **M** should now model $P(s_{t+1}, r_{t+1}, a_{t+1}, d_{t+1}|s_t, a_t, h_t)$. For more difficult tasks, **C** needs to actively explore parts of the environment that can help improve the world model.

References

Ha, David and Schmidhuber, Jürgen. World models. *arXiv preprint arXiv:1803.10122*, 2018.