Recurrent World Models Facilitate Policy Evolution

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

October 22, 2019

Ha and Schmidhuber [1] present a novel approach to modeling the environment through an internal representation also called *predictive model* which is inspired by the human cognitive system. Much like we humans are able to benefit from such a predictive model which is able to warn us of danger and allows for fast reflexive behaviors instead of planning everything we do far ahead. They believe reinforcement learning (RL) agents may benefit from having access to such a predictive model of the world state.

To implement their predictive model they use a RNN to predict the sequence of states based on previous observations and the actions taken by the agent. More precisely, they use visual sensory component V which is implemented as a VAE to obtain a compressed encoding of a 2D image, typically a frame from a video stream, and feed this latent vector z_t into the predictive model M (the RNN) which models the probability distribution $P(z_{t+1}|\alpha_t, z_t, h_t)$ where α_t is the action taken by the decision-making component C and h_t is the previous hidden state of M. As the representation of the world generated by M is not perfect the agent might be able to exploit these imperfections to achieve a higher reward. In order to prevent this, the authors additionally introduce a temperature parameter π to model uncertainty in the representation of the environment (this is also called a *mixture- density RNN*).

Ha and Schmidhuber test this model on two tasks. First they show that their model can be used to solve the car racing task while achieving a much higer average reward than previous models, allowing only very few driving mistakes. They compare the performance of their model without the predictive movel M and find that this is in line with the performance of other agents on the OpenAI Gym's leaderboard, while the model with full access to M performs much better. Interestingly, they observe that the model is also much better at sharp corners in particular which require faster and more reflexive driving decisions.

In a second experiment the authors find that it is possible to learn the policy inside the generated environment and then transfer it back into the actual environment. They use the VizDoom environment where the agent must learn to avoid fireballs shot toward her. By increasing the temperature parameter π they can increase the difficulty of the generated environment the agent learns in, even making it more difficult than the actual environment and preventing the agent from exploiting imperfections in the world model produced by M.

References

[1] Ha, David, and Jürgen Schmidhuber. "Recurrent world models facilitate policy evolution." *Advances in Neural Information Processing Systems*. 2018.