Jenish Maharjan **Week 6**

Paper: *World Models*

Summary:

The paper by David Ha and Jurgen Schmidhuber entitled *"World Models"* demonstrates that their Reinforcement Learning agent is able to learn by training in its own simulated environment. The basic idea is to first train a model that understands the world and then either use features learned by the model to train an agent to learn policies that can solve the required task or train an agent entirely inside the model and transfer the policy back into the actual environment. Based on how humans use predictions using a mental model of the environment to make decisions, they propose a model with three major components – the vision module, the memory RNN and a small controller. Vision model encodes the high-dimensional observation into a low-dimensional latent vector, the memory RNN integrates the historical codes to create a representation that can predict future states and the controller uses the representations from both vision and memory model to select good actions.

The Vision module consists of a basic Variational Auto Encoder (VAE) which maps images from a 64x64x3 space down to a 64-dimensional vector. The VAE finds a lower dimensional representation of the observation by minimizing reconstruction loss with a penalty for moving far from beliefs. The memory model is a Mixture Density Recurrent Neural Network that serves as a predictive model of the future z vectors that VAE is expected to produce. The controller is responsible for deciding what actions to take, where in the experiments the actions take on continuous values. The simple controller is a single layer linear model which uses an evolutionary algorithm to find its parameters.

The training of their system follows a four-step process after initializing all three components. First the trajectories are collected from the controller after which the raw observations are use to train the vision model. The raw observations are encoded to the latent space with the vision model and then the memory model is trained. The outputs from the memory model are then used to train the controller. In the third step, the memory model can either be trained on the actual environment or inside its own world referred to as "dreams". The authors mention that training inside synthetic environments was not feasible in recurrent environment simulators because the controller was able to take advantage of imperfections in the model. In this case, however, a temperature parameter is introduced which creates more randomness in sample from the memory model. This forces the controller to be more robust.

The authors have demonstrated the performance of their model for the car racing experiment. In their environment, the tracks are randomly generated for each trial and the agent is rewarded to visiting as many tiles as possible in the least amount of time. The agent is responsible for three continuous actions – steering, acceleration and brake. They trained the V model using a dataset of 10,000 random rollouts of the environment. Initially, only the vision model is used to handicap the controller which performed in line with other agents on OpenAI Gym's leaderboard and traditional Deep RL methods. When the system with both the vision and the memory models were used, the controller is able to learn a good representation of both the current observation and what to expect in the future. The driving in this case is more stable and the agent is able to handle sharp corners more effectively. The authors implement their model for VizDoom, in which they experiment by training solely using sampled z vectors as a "dreamt up" interface and then transferring to the actual environment and show that training within the hallucinated environment with high temperature and that uncertainty leads to a better-performing and low variance

policy when transferred to the actual environment. The authors discuss how the controller having access to the memory model's hidden state can lead to exploitation of some edge cases or adversarial behavior in which case the agent learns that when training within predicted z, certain actions can lead to magically extinguishing incoming bullets which is interesting and not too unexpected.

Strengths:

- The paper is well written and presents a really novel idea using language that is suitable for anyone interested in learning about new findings in the field of artificial intelligence.
- The images (and animations in the online version) used in the paper are really helpful in understanding their proposed system.
- They've also put the pseudocode in the paper which helps in understanding the process.

Weaknesses:

- The paper could be made more concise by omitting some information. The paper is very elaborate and has sections that focus on previous related works.

Confusions:

- How does the Covariance-Matrix Adaptation Evolution Strategy (CMA-ES) work?
- What is intuition behind Mixture density models?

Discussion Questions:

- What are some real-world scenarios where a system like this could be used and how well do such systems translate to those scenarios?
- Why do they use the term "dream"? Is it simply used to refer to a world created by the model?
- How would increasing the complexity of the controller affect the system?