Curiosity Driven Exploration for MineRL Navigation Challenge

Kartik Patath, Sapan Agrawal, Vamshi K. Uppununthala, Johnathan A'Vant

INTRODUCTION

- For most of the real-world problems the rewards are extremely sparse.
- For extremely sparse environments, curiosity serves as an intrinsic reward signal to explore the environment and learn the skills better.
- Curiosity[2] can be formulated as the error in agent's ability to predict the consequence of its own actions in a visual feature space.
  - Visual feature space is learned by self-supervised learning.
  - This feature space can be translated to high dimensional spaces as well.
- In this project, we have compared two state of the art methods for curiosity driven exploration - ICM and RND

RELATED WORK

Many approaches measure information gain or exploration bonus:

- [1] uses an exploration strategy that maximizes information gain about the agent's belief of the environment's dynamics.
- [2] shows that training a forward dynamics model in a random feature space typically works as well as any other feature space when used to create an exploration bonus
- [3] calculates curiosity but that is not attracted by the stochastic elements of an environment improving upon ICM.

MineRL ENVIRONMENT

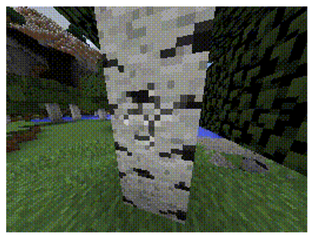


Figure 1: MineRL Environment [4]

- Nine different challenges featuring Navigation, searching and collecting tools, Chopping trees, finding diamond.
- Observation Space: Dist
  - Compass Angle: Floating value
  - Inventory: Dict
  - Pov: Image (64x64x3)
- Action Space: Dict

INTRINSIC CURIOSITY MODULE (ICM)

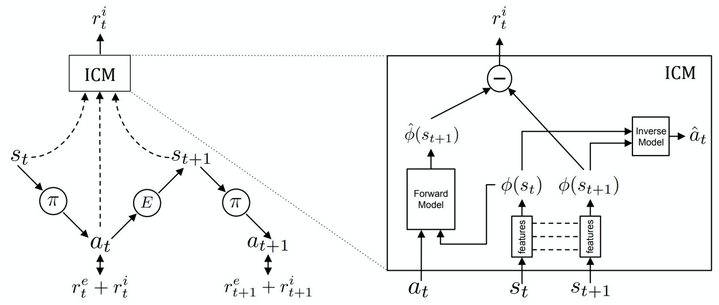


Figure 2: ICM Module [2]

- The ICM generates the intrinsic reward based on how well it predicts the consequences of its own actions.
- Instead of predicting raw pixel values, the agent learns to predict only those features in environment which affects the agent's actions.
- This feature space is learned using self-supervision - training a neural network on inverse dynamics task of predicting agent's action given its current and next states.
- The policy subsystem is trained to maximize the sum of extrinsic and intrinsic reward.

RANDOM NETWORK DISTILLATION (RND)

- Because of the way ICM calculates the intrinsic reward, our agent can fall into what we call the "Noisy TV problem".
- Noisy TV problems show how next-state prediction agents can be attracted by stochastic or noisy elements in the environment.
- RND exploration is a method that calculates curiosity but that is not attracted by the stochastic elements of an environment.
- A target network, f, with fixed, randomized weights, which is never trained. That generates a **feature** representation for every state.
- A prediction network, f_hat, that tries to predict the target network's output.
- An calculates the intrinsic reward as an L2 norm of the predicted next states from feature and predictor networks

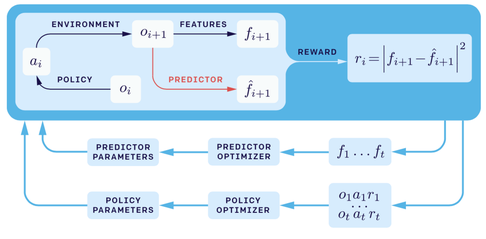
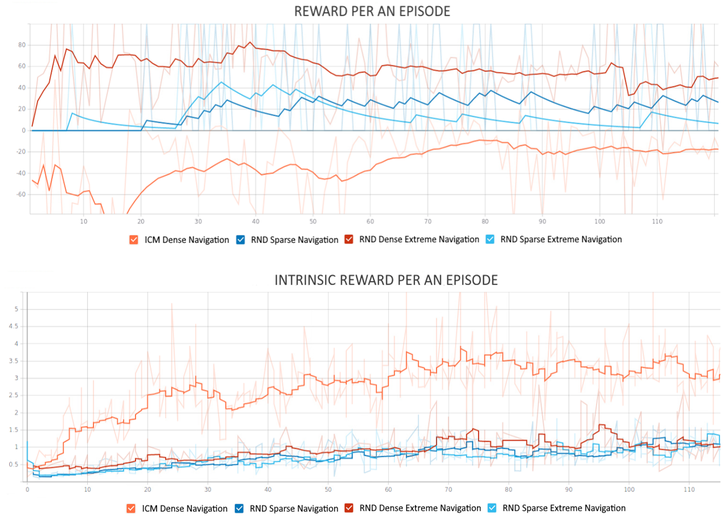


Figure 2: RND Intrinsic reward process Module [5]

RESULTS



ANALYSIS

- ICM generates curiosity by calculating the error of predicting the next state given the current state, and this leads to a big problem: procrastinating agents i.e. the agent stays at a place predicting the noisy environment.
  - looking at rendering scenes (unpredicatable), and punching corners to view particles (as shown below).

- Such problems were efficiently tackled by RND as seen in the results section.

REFERENCES

[1] Houthooft, Rein, et al. "Variational information maximizing exploration." (2016).

[2] Pathak, Deepak, et al. "Curiosity-driven exploration by self-supervised prediction." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.

[3] Burda, Yuri, et al. "Exploration by random network distillation." *arXiv preprint arXiv:1810.12894* (2018).

[4] Mine RL : http://minerl.io/competition/

[5] RND :https://openai.com/blog/reinforcement-learning-with-prediction-based-rewards/