Perception-Driven Curiosity with Bayesian Surprise

Bernadette Bucher, Anton Arapin, Ramanan Sekar, Feifei Duan,
Marc Badger, Kostas Daniilidis, Oleh Rybkin
University of Pennsylvania

Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

Motivation and Model

- Methods encouraging agents to explore their environment by rewarding actions that yield unexpected results are commonly referred to as *curiosity*.
- In scenarios in which extrinsic rewards are sparse, combining extrinsic and intrinsic curiosity rewards gives a framework for agents to discover how to gain extrinsic rewards.
- When agents explore, they can build more robust policies for their environment even if extrinsic rewards are readily available.

Our Contributions

1. Perception-driven approach to curiosity. We develop a perception model which gives a measurement of how much information an agent has retained about the visual characteristics of their environment using a conditional variational autoencoder (CVAE).

2. Bayesian metric for surprise. We derived an conditional probability expression from sampling the latent space of our perception model. We anticipate that this formulation will be less brittle to environmental noise than existing approaches .

Perception Model Losses

Mean-squared error between the embedded state and reconstructed embedded state. Approximates conditional probability used in computation of Bayesian surprise.

$$L_{MSE} = \left\| \phi_{t+1} - \widehat{\phi_{t+1}} \right\|_2^2 \approx -\frac{1}{N} \sum_{i=1}^{N} \log p(\phi_{t+1} | z_i, \phi_t, a_t)$$

Kullback-Leibler divergence between the realized and target (standard normal) latent space distributions.

$$L_{KL} = D_{KL}\left(q(z|\phi_t, a_t, \phi_{t+1}) || p(z)\right)$$

Total model loss with a hyper-parameter to weight loss contributions.

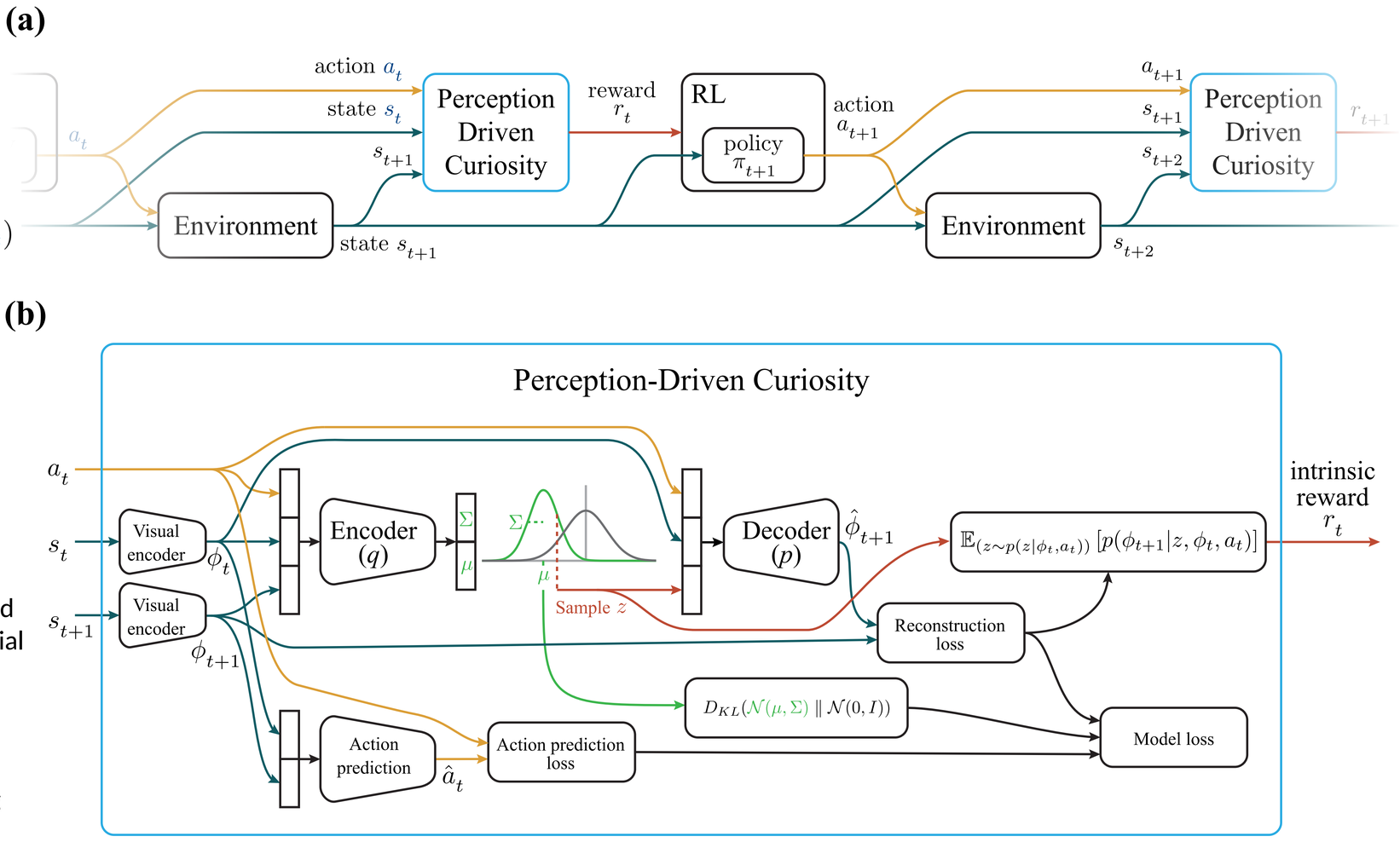
$$L_{total} = \lambda\left(L_{KL} + L_{MSE}\right) + (1-\lambda) L_A$$

The action prediction loss is formulated as a maximum likelihood estimation of the parameters of our network under a multinomial distribution.

Bayesian Surprise

Sampling from the latent space of our CVAE yields the following expression for conditional probability.

$$r_t = -\mathbb{E}_{(z\sim p(z|\phi_t,a_t))}\left[p(\phi_{t+1}|z,\phi_t,a_t)\right] = -\mathbb{E}_{(z\sim q(z|\phi_{t+1},\phi_t,a_t))}\left[\frac{p(\phi_{t+1}|z,\phi_t,a_t)p(z)}{q(z|\phi_{t+1},\phi_t,a_t)}\right]$$

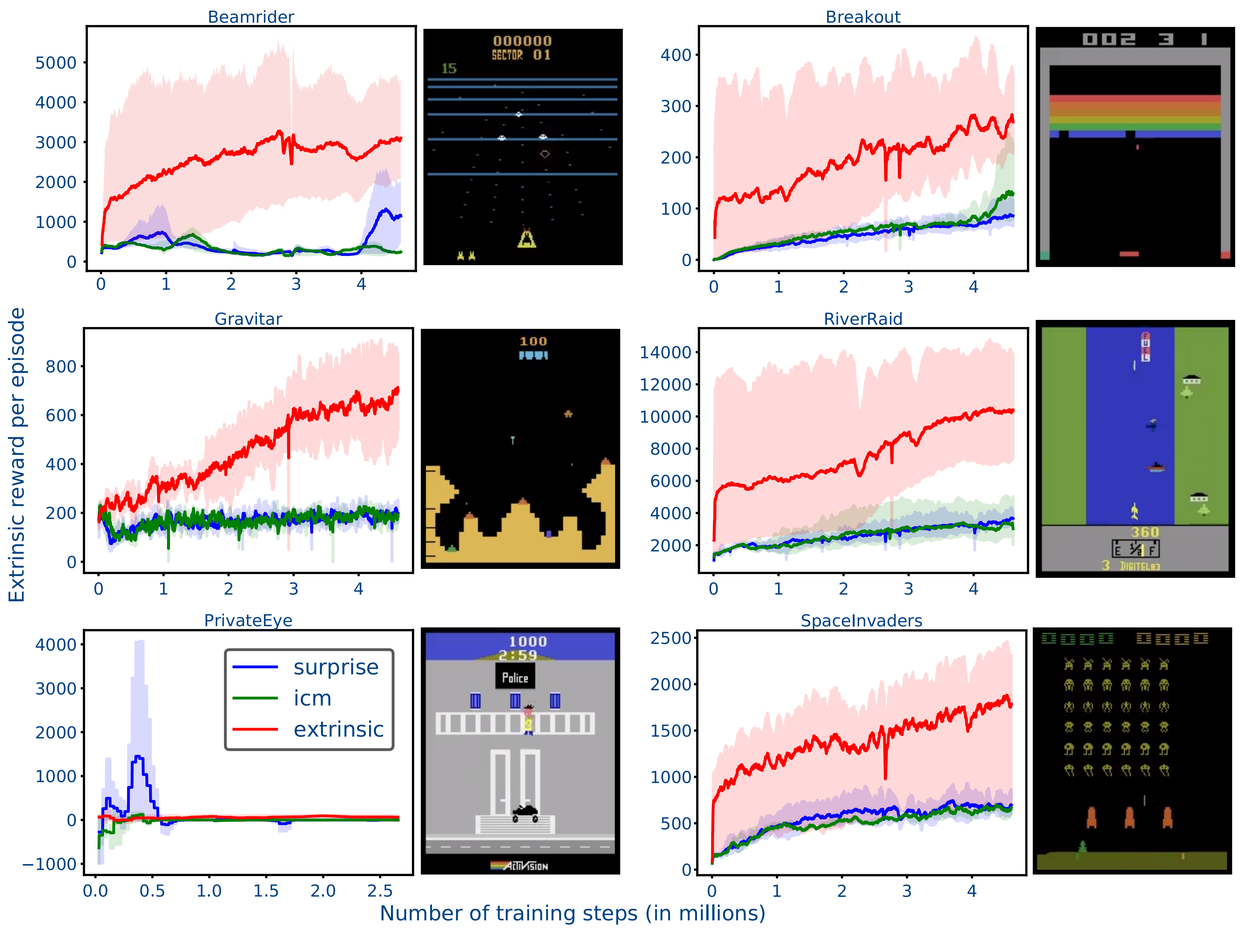


Our model easily integrates with reinforcement learning algorithms (a) and captures both perception- and prediction-driven surprise (b).

Results

We compare our perception-driven curiosity model with an intrinsic reward given by our Bayesian surprise metric to ICM (Intrinsic Curiosity Module), the leading prediction-driven curiosity approach.

We use the PPO (Proximal Policy Optimization) algorithm for reinforcement learning in all of our experiments which receives extrinsic rewards from the game environments only in our extrinsic reward baselines.



Extrinsic reward over time steps in training for the following Atari games: Beamrider, Breakout, Gravitar, River Raid, Private Eye, Space Invaders. The extrinsic reward achieved by the policies trained with ICM, our Bayesian surprise, and extrinsic reward are all shown. Pictures of the game scenes are next to each plot.

Table 1: Mean extrinsic reward over 4.6 million time steps and across 3 independent runs in Atari games achieved by policies trained with extrinsic rewards, the ICM, and our Bayesian surprise.

	Reward Strategies		
Atari Game	Extrinsic	ICM	Ours
Gravitar	458.08	**168.50**	167.58
Private Eye	59.39	-14.08	**102.21**
Space Invaders	1392.22	501.03	**550.59**
Beamrider	2570.83	325.77	**421.78**
Breakout	192.68	**56.54**	47.67
River Raid	7736.56	2595.18	**2612.51**

Conclusions and Future Work

- Our method performs comparably or outperforms ICM on each Atari game in our experiments except for Breakout which is the only entirely deterministic game.
- Initial results indicated that our model deals with stochasticity in the environment well.
- To confirm causality between the formulation of our model and our successful exploration in the presence of random actions, we intend to follow up this work with a large scale study of curiosity in environments with stochasticity.

Acknowledgments

This work was funded by the Honda Research Institute through the Curious Minded Machines project.