VARIATIONAL STATE ENCODING AS INTRINSIC MOTIVATION IN REINFORCEMENT LEARNING

Martin Klissarov*, Riashat Islam*, Khimya Khetarpal & Doina Precup
Department of Computer Science
Mila/McGill University
{mklissa,dprecup}@cs.mcgill.ca
{riashat.islam,khimya.khetarpal}@mail.mcgill.ca

ABSTRACT

Discovering efficient exploration strategies is a central challenge in reinforcement learning (RL), especially in the context of sparse rewards environments. We postulate that to discover such strategies, an RL agent should be able to identify surprising, and potentially useful, states where the agent encounters meaningful information that deviates from its prior beliefs of the environment. Intuitively, this approach could be understood as leveraging a measure of an agent's surprise to guide exploration. To this end, we provide a straightforward mechanism by training a variational auto-encoder to extract the latent structure of the task. Importantly, variational auto-encoders maintain a posterior distribution over this latent structure. By measuring the difference between this distribution and the agent's prior beliefs, we are able to identify states which can hold meaningful information. Leveraging this as a measure of intrinsic motivation, we empirically demonstrate that an agent can solve a series of challenging sparse reward, highly stochastic and partially observable maze tasks. We also perform experiments on continuous control tasks with dense rewards and show improved performance in most cases.

1 INTRODUCTION

Reinforcement learning (RL) algorithms have achieved several recent accomplishments, especially by using non-linear function approximators to solve high dimensional complex tasks. However, most RL algorithms rely on well designed reward functions to guide the behaviour of the agent. Hand-crafting such reward functions is complex and can sometimes lead to unexpected behaviour. In order to be deployed in real-world settings, RL agents will have to be able to learn from sparse rewards environments. A key step towards scaling RL algorithms for unknown reward functions is for the agent to naturally adapt its behaviour by learning a good exploration strategy.

Exploiting task structure is way to learn efficient exploration strategies in RL. Recent approaches include the discovery of bottleneck states (Goyal et al., 2019) or learning a feature space (François-Lavet et al., 2018). Exploration can also be formulated as an agent's internal drive towards learning more about the environment. This is often defined as *intrinsic motivation*, or curiosity of the agent (Schmidhuber, 1991a; Oudeyer et al., 2016). Intrinsic motivation is also an important concept in developmental psychology, where it is defined as the desire to pursue an activity for its inherent satisfaction rather than for some external pressure or reward (Oudeyer & Kaplan, 2009). Curiosity or intrinsic motivation can therefore be thought of as a task agnostic exploration heuristic towards the goal of learning in an online fashion based on the agent's interactions with the environment.

In this work, we propose a formulation of intrinsic motivation based on the definition of Bayesian surprise as expressed in Itti & Baldi (2009). The intuition behind this approach is that experiences which deviate from the agent's prior beliefs about the world are surprising, and potentially useful for learning. In other words, the agent should be able to identify the states which create important changes to its prior knowledge by measuring the difference between posterior and prior distribution after visiting such states. We propose a framework to identify surprising or useful states in the environment via latent representation learning, which we use as intrinsic motivation for solving sparse rewards and partially observable maze tasks as well as continuous control with dense rewards.

Our Contributions: We use a Variational Auto-Encoder (VAE) to project the state space into a probabilistic latent representation that would represent the inherent structure of the environment. By using a VAE we naturally obtain a measure of the agent's surprise defined by how much the posterior distribution over the latent representation deviates from its prior belief. This is measured in the form of a KL divergence $KL(p(Z|S)||p(Z))$ where $p(Z)$ is the agent's prior distribution over the latent structure of the environment and $p(Z|S)$ the posterior. We incentivize the agent to visit surprising (and potentially useful) regions of the state space by providing this KL divergence as intrinsic motivation.

2 Preliminaries and Background

In this work we consider the standard reinforcement learning setting which considers the environment as a Markov Decision Process $\mathcal{M}$, which is defined as a tuple $\doteq(\mathcal{S}, \mathcal{A}, \gamma, r, P)$. Here $\mathcal{S}$ is the state set, $\mathcal{A}$ the action set, $\gamma \in [0, 1)$ the discount factor, $r : S \times A \to Dist(\mathbb{R})$ the reward function and $P : S \times A \to Dist(S)$ the transition probability distribution. A policy $\pi : S \to Dist(A)$ specifies a way of behaving, and its value function is the expected return obtained by following π: $V_\pi(s) \doteq \mathbb{E}_\pi \left[\sum_{t=0} \gamma^t r(S_t, A_t) | S_0 = s\right]$. V_π satisfies the following Bellman equations: $V_\pi(s) = \sum_a \pi(a|s) \left(r(s, a) + \gamma \sum_{s'} P(s'|s, a) V_\pi(s')\right)$.

Curiosity as a form of intrinsic motivation has been argued to be a fundamental component for efficient learning (Friston et al., 2006). One of the ways to implement curiosity is by maintaining a forward dynamics model of the environment and using its prediction error (Pathak et al., 2017; Schmidhuber, 1991b) or prediction uncertainty (Houthooft et al., 2016) as intrinsic reward. These approaches encourage the agent to visit regions of the state space where the dynamics of the environment are less well understood, therefore guiding exploration. However, their performance tends to suffer in stochastic environments as it becomes harder to predict the consequences of the agent's actions. Another direction (Ostrovski et al., 2017; Bellemare et al., 2016) formulated an exploration bonus as a measure of novelty in terms of unseen states. Such approaches have shown great potential but contain some strict requirements on the density model of the states, such as it should be learning-positive. Other ways to improve exploration include the optimal rewards framework (Singh et al., 2010) where the authors propose that the optimal intrinsic reward is the one that would maximize the extrinsic reward. However, defining such optimal reward function is an open question. Our approach on the other hand aligns with the Bayesian perspective on surprise (Itti & Baldi, 2009) which has been shown to be applicable across different spatio-temporal scales and levels of abstractions.

3 Leveraging State Encoding for Intrinsic Motivation

3.1 Intrinsic motivation

In this work, we assume that the experiences S of an agent are generated by a random latent process defined through the variable Z. This latent process can encode some structure or pattern present in the observed data. The goal of the agent would be to extract and learn this structure in order to have a better understanding of the world it is interacting with. However, to faithfully represent the latent factors of variation, an agent has to successfully explore the environment. Therefore, the objective of extracting structure from the environment is deeply interlaced with the objective of exploration. One way to attend to this challenge is by adding an intrinsic reward that would depend on the quality of the model of the environment. On one hand, this intrinsic motivation would encourage the agent to gather unseen data which would improve the model, while on the other hand guiding the agent to fully explore its environment.

We propose a measure of intrinsic motivation formulated as the distance between the posterior distribution over the latent variable $p(Z|S)$ after seeing new data S and the prior $p(Z)$. A natural way to measure this distance is through the KL divergence. Therefore, we can define the intrinsic reward at a state S as

$$r_{intrinsic}(S) = KL\big((p(Z|S)||p(Z)\big)$$

Our measure of intrinsic reward is closely related to the definition of Bayesian surprise proposed by (Itti & Baldi, 2009). In this work, the authors argue that the only rigorous definition of surprise is by

measuring how data affects the beliefs of an observer about the world. This measure of surprise is computed as the difference between the prior distribution $p(M)$ of the observer, where M represents the possible models of its environment, and posterior distribution $p(M|D)$ after observing data D. Our definition of intrinsic motivation can then be seen as an approximation to Bayesian surprise, with the slight conceptual difference that the variable Z represents a latent encoding of the structure of the environment. This difference will have a key impact in our work as it directly guides our implementation.

3.2 APPROACH

It is usually impractical to infer exactly the posterior distribution $p(Z|S)$ as it involves intractable integrals. We will therefore choose to approximate this posterior by a variational distribution $q_\phi(Z|S)$. A natural candidate to represent this distribution is through a Variational Auto-Encoder (VAE). VAEs take the inputs S and project them into latent space Z, which is usually of smaller dimensionality. This latent space is meant to capture factors of variation (patterns) within the data. Importantly, a VAE minimizes the following loss:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(Z|S)}\big[\log p_\theta(S|Z)\big] - KL\big(q_\phi(Z|S)||p(Z)\big)$$

where the first term is the reconstruction loss while the second term encourages the approximate posterior $q_\phi(Z|S)$ to stay close to the prior $p(Z)$. In practice, the prior is chosen as a unit Gaussian to simplify the implementation. This also our choice for the prior.

Algorithm 1: Training loop with intrinsic motivation for A2C.

for *Episode=0,1,2,....* **do**
  Initialize dataset $\mathcal{D}$ and insert s_0 in $\mathcal{D}$.
  for *t=0,1,2...T* **do**
    Take action a_t and observe next state s_{t+1} and extrinsic reward $r_{extrinsic}(s_{t+1})$
    Compute intrinsic reward: $r_{intrinsic}(s_{t+1}) = KL(q_\phi(z|s_{t+1})||p(z))$
    Store tuple $(s_{t+1}, a_t, r_{intrinsic}(s_{t+1}), r_{extrinsic}(s_{t+1}))$ in $\mathcal{D}$
    if *mod(t,N)* **then**
      Train the actor and critic on return $\mathrm{G}_t = \sum_t r_{extrinsic}(s_t) + \beta KL(q_\phi(z|s_t)||p(z))$
      Train the VAE on the collected states s in $\mathcal{D}$.
      Initialize dataset $\mathcal{D}$ and insert s_t in $\mathcal{D}$.
  end
end

The overall loss function is a lower-bound to the likelihood of the data. This lower-bound is appealing as it explicitly evaluates the KL divergence between posterior and prior distributions. It is therefore straightforward to leverage VAEs for intrinsic motivation. To do so, we need to separately train a VAE on the stream of data an RL agent experiences. We can then define the useful states, or states which contain a high degree of surprise, in places where the KL is high between the posterior and the prior. This KL between the posterior and prior, whenever high, would encourage the agent to visit that region of the state space when it is provided as intrinsic motivation. By doing so, the agent would efficiently explore its environment and improve the quality of the VAE for encoding the hidden structure in the data.

We define the intrinsic motivation reward as $r_{intrinsic}(s_t) = KL(q_\phi(z|s_t)||p(z))$ such that at every step, the agent gets a total reward of $r_{total}(s_t) = r_{extrinsic}(s_t) + \beta r_{intrinsic}(s_t)$. We can therefore define policy gradient objectives based on the cumulative discounted total return, which includes both the extrinsic and intrinsic task rewards. In our implementation, we use be using actor-critic to solve the task at hand. However, our definition of intrinsic motivation could be readily used with any other policy gradient algorithm, as well as value-based algorithms. We provide a description of the overall process in Algorithm 1.

4 EXPERIMENTAL RESULTS

4.1 MAZE ENVIRONMENTS

We first perform experiments on the multi-room maze tasks which are partially observable and sparse reward tasks, as part of the MiniGrid environment (Chevalier-Boisvert & Willems, 2018). In MiniGrid, the agent has to navigate a number of rooms, by opening *doors* or by using a *key*, in order to get to the goal situated at the other end of the maze. Due to the sparsity in rewards, these maze tasks are often hard to solve, hence requiring efficient exploration strategies. The goal of our experiments is to show that our definition of intrinsic motivation can achieve efficient exploration. To do so, we compare our implementation (VAE) against two baselines: a standard A2C agent and an A2C agent using the prediction error of a model of the transition dynamics as intrinsic motivation (ICM) (Pathak et al., 2017). In Figure 1 we show empirical results on three domains: Multi-Room-N3S4, Multi-Room-N4S4 (where N represents the number of rooms and S the size of the rooms) and Door-Key-8x8. We see that our approach, VAE, and the approach based on the prediction error, ICM, both outperform significantly the A2C baseline. In the Multi-Room-N4S4 and the Door-Key-8x8 environments, we notice that our approach outperforms ICM. Upon investigating the behaviour of the agent, we noticed that the KL divergence was highest at key states such as hallways, in the sight of the door and near the goal. Therefore, we believe one reason explaining the better empirical performance is due to a possible correlation between surprising states and useful states in these particular environments.

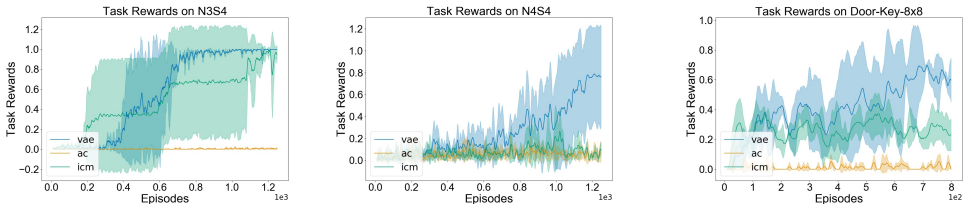
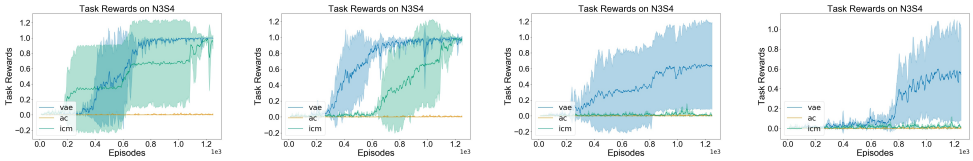


Figure 1: **Task Rewards** on partially observable and sparse reward tasks from the MiniGrid environment. We see that our approach, *VAE, significantly outperforms* both the approach based on the prediction error i.e. ICM, and the A2C baseline.

It is widely-known that intrinsic motivation based on the prediction error of a transition model is sensitive to the inherent stochasticity of the environment (Burda et al., 2018). As such, we performed a series of experiments on the same task but with different degrees of randomness and we show our results in Figure 2. We notice that as the stochasticity in the environment is increased (from left to right), the prediction error of ICM becomes an unreliable source of intrinsic rewards which in turn degrades the performance of the agent on the task. This highlights an important difference between ICM and our approach: the agent is not trying to predict the consequences of its actions, as sometimes they can be very complex, but instead tries to encode the structure present in the stream of observations. This provides a efficient intrinsic signal that can guide the agent even when the environment becomes less predictable.



(a) (b) (c) (d)

Figure 2: **Task Rewards for different degrees of environmental stochasticity**: As the stochasticity in the environment is increased from (a) to (d), the prediction error of ICM becomes an unreliable source of intrinsic rewards which in turn degrades the performance of the agent on the task. On the contrary, our approach VAE shows consistency and is robust to stochasticity in the environment.

4.2 CONTINUOUS CONTROL

Next, we performed experiments on more complex continuous control tasks in the MuJoCo environment (Todorov et al., 2012). We present results in Figure 3. In these experiments we opted for the DDPG algorithm (Lillicrap et al., 2016) as our baseline as it is a competitive algorithm in these set of tasks. We compare the baseline agent (DDPG) to our approach (VAE) on three locomotion tasks: HalfCheetah-v1, Hopper-v1 and Walker2d-v1.

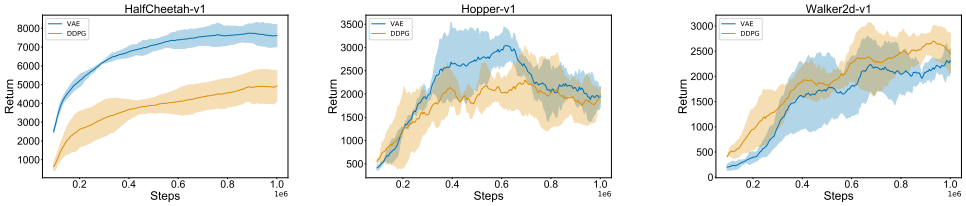


Figure 3: **Task Rewards** on Mujoco environment. We see that our approach is able to outperform the DDPG baseline in two out of the three environments explored.

Our approach improves the performance on most of the tasks. Particularly, in HalfCheetah-v1 we notice that our exploration strategy doubles the final score. In Hopper-v1 we also notice improvements, however in Walker2d-v1 our approach performs worse than the baseline. A possible reason explaining why we do not observe improvement in Walker2d-v1 as compared to the boost in performance in the other two environments is due to the fact that in Walker2d-v1 the agent has to learn a gait while avoiding to fall. This is something that is simply not present in HalfCheetah-v1 and more easily avoidable in Hopper-v1.

5 DISCUSSION AND FUTURE WORK

In this work, we presented an interestingly simple approach towards learning with intrinsic motivation inspired by the definition of Bayesian surprise. We emphasize that our approach is readily extendable towards existing RL frameworks as it requires little overhead. In contrast to several existing works which use prediction error of a transition dynamics model as intrinsic motivation, our approach does not suffer much from an increase in the environment's stochasticity.

A possible improvement to the current framework would be to use a model of the environment that better reflect its latent structure. As it has been noted in Ha & Schmidhuber (2018), variational auto-encoders tend to encode details about the observations that are not always meaningful. To overcome this issue, we could use auxiliary losses to refine the latent representation (François-Lavet et al., 2018). Another possible direction would be to consider explicitly the temporal aspect in reinforcement learning and use a generative model that would account for it (Gregor & Besse, 2018) while still providing a posterior distribution useful for intrinsic motivation.

The key to our approach is to leverage an agent's intrinsic motivation that is task agnostic and solely dependant on the latent structure of the environment. As a consequence, irrespective of a dense or sparse reward function, we can provide an exploration bonus defined by the agent's measure of Bayesian surprise, without requiring the agent to know the task-dependent goal information. This is an interesting step towards transfer learning, where even if the task reward changes or the transition dynamics change, the agent can use the learnt state encoding representation as intrinsic motivation in new tasks. In future work, we aim to evaluate the usefulness of our proposed method on transfer learning tasks where we would provide an exploration bonus in new tasks with the previously learnt variational encoder. The variational state encoding could in fact also be leveraged without an external task specific reward in the context of transfer learning.

REFERENCES

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. *CoRR*, abs/1606.01868, 2016. URL `http://arxiv.org/abs/1606.01868`.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Maxime Chevalier-Boisvert and Lucas Willems. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Vincent François-Lavet, Yoshua Bengio, Doina Precup, and Joelle Pineau. Combined reinforcement learning via abstract representations. *CoRR*, abs/1809.04506, 2018. URL http://arxiv.org/abs/1809.04506.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, 2006.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Sergey Levine, and Yoshua Bengio. Infobot: Transfer and exploration via the information bottleneck. *CoRR*, abs/1901.10902, 2019. URL http://arxiv.org/abs/1901.10902.

Karol Gregor and Frederic Besse. Temporal difference variational auto-encoder. *CoRR*, abs/1806.03107, 2018. URL http://arxiv.org/abs/1806.03107.

David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018. URL http://arxiv.org/abs/1803.10122.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pp. 1109–1117, 2016.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10): 1295–1306, 2009.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.02971.

Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2721–2730. JMLR. org, 2017.

P-Y Oudeyer, Jacqueline Gottlieb, and Manuel Lopes. Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies. In *Progress in brain research*, volume 229, pp. 257–284. Elsevier, 2016.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

Jürgen Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pp. 1458–1463. IEEE, 1991a.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991b.

Satinder Singh, Richard L Lewis, Andrew G Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, 2010.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, pp. 5026–5033. IEEE, 2012. ISBN 978-1-4673-1737-5. URL http://dblp.uni-trier.de/db/conf/iros/iros2012.html#TodorovET12.