LEARNING TO EXPLORE WITH DEEP GENERATIVE MODELS

William Whitney & Alexander Rives
Department of Computer Science
Courant Institute
New York University
{wwhitney, arives}@cs.nyu.edu

1 INTRODUCTION

In this project we explore learning to represent the world from experience. We investigate how an agent might use a generative model of its environment to learn how to act in the world. We are motivated by the problem of commonsense reasoning, and a specific aspect of it expressed by the following idea: an obligate grounding for commonsense is having an accurate model for the world, one that correctly represents the hidden variables that lie behind the states and events of experience. It is this condition for reasoning that our project begins to approach through a few small experiments.

We investigate these questions in the setting of an agent learning to play Atari. Despite the fact that Atari video games follow deterministic rules, have full observability of the environment, and offer only a small set of actions for an agent to choose from, the Atari environment models interesting problems that artificial agents must solve in the real world. These include learning how to behave, survive, and achieve a goal. The central algorithmic challenge is equivalent to that of acting in the real world: the agent must optimize a behavioral policy over an exponentially large sequence space of states and actions. The problem of how to assign credit to behaviors that lead to reward emerges from this.

Credit assignment can be cast as learning a value function that associates states and actions that have no explicit reward with their expected value under a behavioral policy. One approach to this problem is temporal difference (TD) learning. The temporal difference objective maintains a value function and propagates information about future rewards into the past by enforcing a regularity in the value function across time. Q-function learning and Actor-Critic, two of the predominant approaches to reinforcement learning, use the the temporal difference objective in their optimization procedure.

A challenge central to the reinforcement learning setting is that rewards in the environment are sparse. Sequences of successive states and actions leading to reward are unlikely under a random policy; an agent must learn how to act from an improbable signal in its environment. This means that starting from a random initialization an agent with a random policy for exploration repeatedly tries out sequences of actions until eventually a given sequence leads to a reward. In practice current state of the art algorithms must be trained on millions of frames to obtain a high level of performance in even simple Atari games.

There is a tradeoff between exploration of the space of states and actions and exploitation of rewarding regions. To learn an optimal value function, a complete exploration of the state-action space is required. This is possible only in a fully observable state space that is bounded in such a way that it is possible to visit every pair of states and actions. In this case temporal difference learning provably converges to the optimal state-action value function. In realistic settings it is not possible to visit all the states. The Atari environment is more interesting for this reason since the number of states in a game is exponentially large and reaching many of the states depends on being able to play the game well. The most challenging Atari games have the property that the most rewarding regions are the hardest to get to and require taking actions that appear to have negative expected value before it can be known that they lead to reward.

One way that unsupervised learning might be helpful in the reinforcement learning setting is by providing intrinsic rewards that could help agents to explore the state space. Our approach is to equip the agent with a generative model for the state space that it learns from experience, and assign

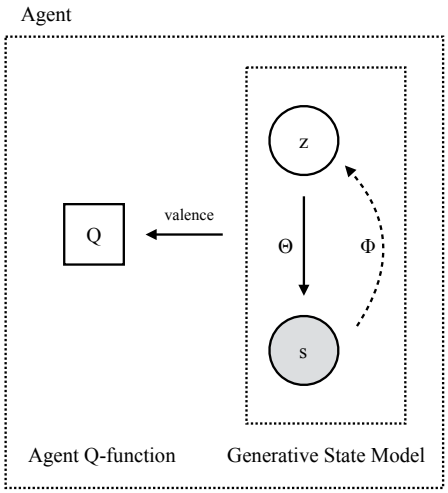

Figure 1: Agent equipped with generative state model. The generative model learns to represent the world from unsupervised data from the agent's interactions with its environment. The latent representation is z and the observed states are s. Theta represents the parameters of the forward generative model. Phi represents the parameters of the variational approximation to the posterior. Both the genenerative distribution and the inference distribution are parameterized by deep convolutional neural networks. In each state the agent receives a valence signal that is increasing with the unlikihood of the state under the generative model. The Q-function learns to approximate the valence of the states using a TD objective.

a *valence*, an intrinsic learned reward signal, to state-action pairs using the model. We consider the simple approach of rewarding actions that lead the agent to reach states that are unlikely under the generative model. This has the effect of rewarding the agent for visiting states that it has little experience in, and rewards actions that produce information that might lead to the adjustments to the parameters of the generative model of the state space.

2 MODEL

This work attempts to incentivize exploration of rarely-seen regions of statespace using a generative model of states. In many reinforcement learning tasks rewards can be quite sparse; in others, the expected reward for any action may be negative for a novice player. In these environments it can intuitively be helpful to encourage the agent to explore. A natural formulation of the exploration objective is "Go where you haven't been before."

We formalize this goal using a generative model of states which provides a lower bound on the likelihood of any state under the policy. This lower bound provides a practical way to incentivize an agent to visit regions of statespace which it sees only infrequently. This form of "intrinsic motivation" aligns nicely with the design of video games, which typically aim to give the player the feeling that they are exploring a world or trying to get further than they have before. Because of this structure, succeeding in the game (as measured by the score) and exploring statespace are in some cases nearly perfectly aligned.

2.1 Q-LEARNING AGENT

The reinforcement learning algorithm we use is Q-learning (Sutton & Barto, 1998) as instantiated in the DQN (Mnih et al., 2015). We use the DQN code released by the authors as a known-good agent. As our work centers on a modification of the reward signal, it is essential that we have a stable base to work from. By using the DQN to optimize our objective we can be confident that differences in

performance come from the difference in objective functions itself, not failures or advantages of the agent.

2.2 GENERATIVE MODEL

In order to give rewards for attaining previously unlikely states, we require a generative model which provides us a likelihood function over states, or at least a lower bound on this quantity. We use a variational autoencoder (Kingma & Welling, 2013) to optimize this lower bound.

Our generative process is as follows. We first sample an instance of z from the prior over our latent space:

$$p(z) = \mathcal{N}(0, \mathbb{1}) \tag{1}$$

We then calculate the conditional distribution according to

$$p(x|z) = \prod_{(i,j)\in x} \mathcal{N}(\mu_{ij}(z), \sigma_{ij}(z)) \tag{2}$$

where μ and σ are parametrized by a neural network and (i, j) range over the image. Each pixel is then sampled independently.

The lower bound to be optimized is the usual variational autoencoder objective, namely

$$\log p_\theta(x) \geq \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[-\log q_\phi(z|x) + \log p_\theta(x, z)] \tag{3}$$

$$= -D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \tag{4}$$

where q_ϕ is the inference network with parameters ϕ, $p(z)$ is our prior given by Equation 1, and $p_\theta(x|z)$ is our generative distribution with parameters θ given by Equation 2.

2.3 VALENCE SIGNALS

Tying together the components of our model is the valence of visiting a state, sometimes elsewhere called pseudo-rewards or intrinsic rewards. The valence of a state reflects its novelty; states which have been visited very often in the past have negative valence, while new states have positive valence.

For simplicity we assign positive valence to states which have below-average log probability and negative valence otherwise. To avoid issues of scaling between the game's reward signals and valence, we use the sign of this difference as our valence signal:

$$\nu(s) = \text{sign}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\gamma^{T-t}\mathcal{L}(s_{T-t})\right) - \mathcal{L}(s_T)\right\} \tag{5}$$

where $\mathcal{L}(s)$ is the variational lower bound on the log likelihood of state s (Equation 4) given by our generative model.

3 RELATED WORK

The problem of incentivized exploration for reinforcement learning, also called intrinsic motivation, is one of the most fundamental challenges in the field. It is accordingly very well studied in the literature.

Perhaps the simplest method for exploring the environment is given by an epsilon-greedy strategy (Sutton & Barto, 1998), in which the agent takes the action which it currently believes to be optimal with some probability epsilon, and otherwise selects an action uniformly at random. This is the strategy followed by Mnih et al. (2015), and we follow it with modifications to the reward function.

Schmidhuber (1991) may be the earliest work to propose incentivizing a reinforcement learning agent to generate experience which improves a model of the environment. They propose that using an agent this way allows for training such a world model with fewer examples; the agent is an auxillary tool to this objective. Today this would likely be discussed more in the context of *active learning* rather than intrinsic motivation.

More recently, Stadie et al. (2015) applied unsupervised learning techniques quite similar to ours. They learn a latent space by pretraining an autoencoder (Hinton & Salakhutdinov, 2006) on images sampled from an agent taking random actions, then learn a predictive model of the next latent state z' given the current latent state z and action a. They then give rewards to their agent proportional to the mean-squared error of these latent predictions. While this model appears to give some benefit in several of the games they tested, the improvement was fairly small. This model differs from ours in that a) they use predictive error instead of reconstruction error; b) they use Euclidean error instead of directly estimating the density function; and c) their model is not trained end to end.

A more sophisticated version of this idea is put forward by Houthooft et al. (2016), who propose to learn a model of the environment's transition function, then provide pseudorewards proportional to the reduction in that predictive model's entropy; this entropy is estimated by a variational approximation. This method provides rewards for reducing the uncertainty of a predictive model.

Another approach is that of Mohamed & Rezende (2015), who used a variational model to estimate the *empowerment* of an agent, defined as the mutual information between a sequence of actions and the resulting state. This empowerment score was then used as an intrinsic reward for the agent.

Bellemare et al. (2016) propose a model which is in principle quite similar to ours. It uses a complex density model to create a "pseudocount" function which is asymptotically related to the number of times that the agent has visited a particular state, but which converges faster. The agent then receives intrinsic rewards for visiting states which have low "pseudocounts"; that is, which are unlikely under the model. This high-level objective is the same as ours, but the model we employ is quite different.

4 EXPERIMENTS

We tested our hypothesis that achieving valence (novelty) is closely aligned with scoring well in the Atari game Breakout by training several versions of our model using valence alone without giving the agent any true reward signal at all. We compare the actual game rewards earned by a baseline DQN model trained on the normal reward signal to those earned by a DQN trained solely on valence.

This comparison tests exactly the question we are interested in: does optimizing for novel experiences encourage an agent to make substantial forward progress in the game? And if so, how does this progress compare to that of an agent trained using the game's own reward signal?

4.1 RESULTS

At least in the early phase of training, our valence signals allowed for dramatically faster learning than the true game rewards; this result is surprising given that the rewards in Breakout are relatively dense, with the agent earning a +1 for every block destroyed and a -1 for every death. Even in very limited training time we allowed (about 1/50 the number of frames used in the original DQN paper) the valence models achieved nearly 50% of human-level performance (as measured by Mnih et al. (2015)).

This confirms our hypothesis, at least for Breakout, that the novelty metric given by valence is well aligned with the rewards of real games. Furthermore, the extremely high density of these intrinsic rewards allows the valence-based model to actually outpace the model trained on real rewards.

5 DISCUSSION

The performance of the valence reward relative to the baseline is suggestive. Notably an agent trained purely on an unsupervised valence signal exceeds the Q-learning baseline in the early stages of learning. The only reward the agent receives is for the unlikelihood of its observations under its

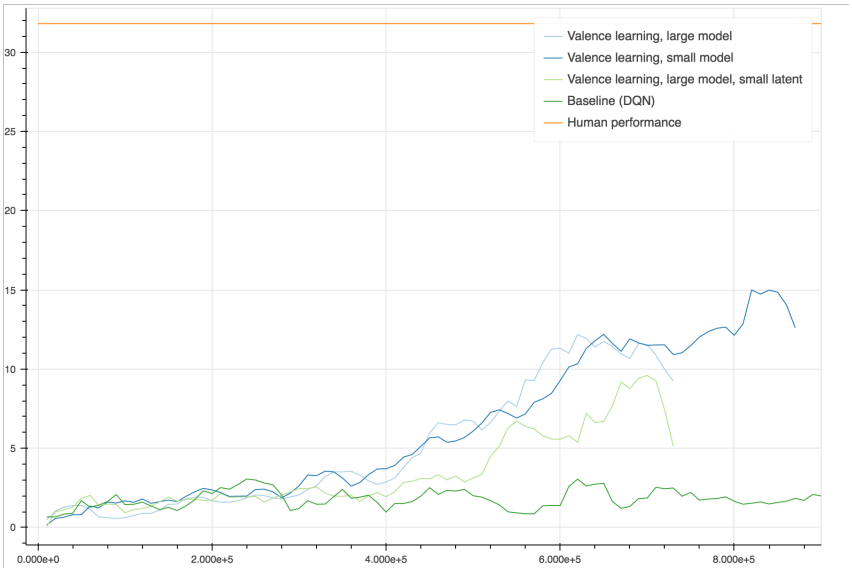

Figure 2: Learning curves for various versions of our model and a DQN baseline. DQN eventually reaches similar performance to our model, but takes about twice as long.

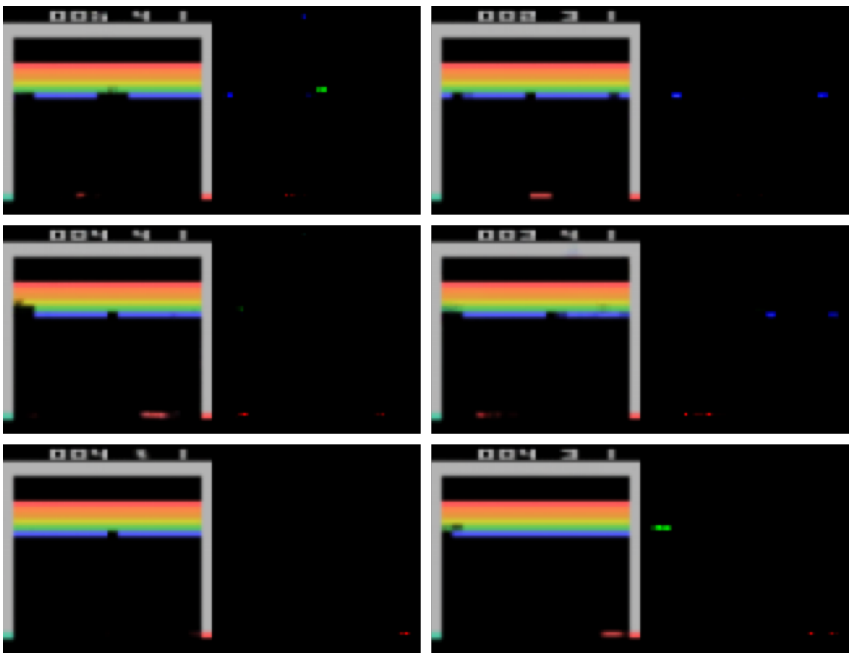

Figure 3: Random samples from our generative model. The first and third columns show the means of the samples, while the second and fourth show the variances. The model displays uncertainty about the presence or absence of certain blocks. It is also uncertain about the exact location of the paddle as evidenced by increased variance on each end.

Model	Layers	Hidden
Large	64x3x3-128x3x3-128x3x3-128x3x3-16x3x3	400
Bottlneck	64x3x3-128x3x3-128x3x3-128x3x3-16x3x3	10
Small Model	32x3x3-64x3x3-64x3x3-64x3x3-8x3x3	10

Table 1: Deep neural network architecture. Three different models were trained on 80x80 images downsampled from the full 160x210 Atari frames. The large model has a high capacity in the intermediate convolutional layers and large latent space. The bottleneck model has the same intermediate capacity as the large model and a compressed latent space. The capacity of the intermediate layers in the small model is reduced by a factor of two and has hidden space that is the same as the bottleneck model. Each of the convolutional layers in the inference network is followed by spatial max pooling with a stride of 2. The generative network uses spatial convolutions followed by spatial upsampling.

generative model of those observations. We are continuing to train the model and are interested in seeing how high performance can be acheived without providing any reward signal from the game.

Intuitively the performance we observe makes sense. In order for the agent to reach states that it is not yet good at representing, the agent must progress further and further into the game. In breakout this can only be achieved one way: by continuing to knock down bricks. We are also interested in trying this approach for games that are known to be difficult even for state of the art RL.

Our models are intensive to train which limited the number of architectures, games, and comparisons that we could perform. We are also training a model that combines valence reward with in-game reward, and are interested in seeing whether it can ultimately exceed the DQN baseline. We are also interested seeing whether we can acheive gains in games that are difficult for RL agents.

Montezuma's Revenge is an adventure game that is especially difficult for RL algorithms. The agent starts on a platform in the middle of a room; falling off the platform leads to immediate death; attempting to leave the room leads to death; and there is a key. Only if the agent retrieves the key can it progress beyond the first room, but in order to reach the key the agent must get past a skull. It is immediately obvious to a human what needs to be done. But to the agent the expected value of any action that it can take appears negative. This illustrates the potential gains for reinforcement learning from having good representations of the world.

Our project is also the beginning of an exploration of how to combine unsupervised learning with reinforcement learning in a principled way. Here we use the representational power of generative models to build a latent representation of the world and use the model to inform the state value function that the agent is optimizing.

Our experiments in this project along with related research discussed above, provide evidence that principled approaches to exploration and learning from the environment could improve standard reinforcment learning techniques. We belive that this is an exciting future direction for research. Our work can be extended in many ways, one direction we are especially interested in is the possibility of learning richer models of the causal dynamics of the environment that could be used for planning.

REFERENCES

Bellemare, Marc G, Srinivasan, Sriram, Ostrovski, Georg, Schaul, Tom, Saxton, David, and Munos, Remi. Unifying count-based exploration and intrinsic motivation. *arXiv preprint arXiv:1606.01868*, 2016.

Hinton, Geoffrey E and Salakhutdinov, Ruslan R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

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

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Mohamed, Shakir and Rezende, Danilo Jimenez. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2125–2133, 2015.

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

Stadie, Bradly C, Levine, Sergey, and Abbeel, Pieter. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

Sutton, Richard S and Barto, Andrew G. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.