Deep Bisimulation Dreaming: Combating Distractions with State Abstractions

Huy Ha
huy.ha@columbia.edu

Sian Lee Kitt
sll2190@columbia.edu

William Zheng
wjz2101@columbia.edu

1 Introduction

To achieve sample efficient deep reinforcement learning in high dimensional (i.e: visual) observation spaces and long-horizon tasks, recent approaches that learn latent world models have shown lots of promise. By encoding visual sensory information into a latent state, the policy, transition dynamics, and/or rewards model can operate on the resulting low-dimensional state more efficiently. Notably, Dreamer [1] uses a probabilistic latent world model [2] to generate more experience in the latent space, enabling policy optimization without interacting with the environment - hence, dreaming, or latent imagination - and achieves unprecedented sample efficiency. Since the latent world model is fully-differentiable, the policy can be optimized for the rewards model's predicted reward using analytical gradients, achieving high performance on long-horizon tasks. The performance of such latent world model approaches crucially depends on how good the latent states are, yet not much investigation has been done on the quality of these representations. What makes a good representation for reinforcement learning?

Consider the task of hammering a nail into a block. Even though visual observations may exhibit wide variance (i.e: due to varied lighting conditions and wallpaper colors), only the pose of the nail head, hammerhead, and block matters for the task. How can we distill these state abstractions from visual observations to allow for efficient policy learning? One approach is to optimize the latent states using reconstruction of high dimensional observations, such as the Variational Auto-Encoder (VAE) [3]. While such dimensionality reduction algorithms are generically applicable to a wide range of settings beyond RL, it assumes that all information in visual observations is not only useful but crucial to the task. However, in the hammering task, such approaches will try to encode the hammer color, which leads to poor generalization to, for instance, another hammering task with different hammer colors. Intuitively, we want a low dimensional state abstraction that explicitly ignores distracting information.

In this project, we aim to learn a generalizable representation for reinforcement learning without reconstruction which is robust against distractors by encoding only task-relevant information. We propose Deep Bisimulation Dreaming (DBD), a new representation learning algorithm for RL which regularizes the latent space with bisimulation metrics [4]. By enforcing distance between latent states to their bisimulation metric distance, latent states are close together if they have similar reward behaviors and dynamic transitions, and far apart otherwise. We compare our approach with a reconstruction-based baseline on a continuous action space robotic task from visual observations on the MuJoCo [5] benchmark. We report that our approach can generalize to novel visual distractors by extracting only task-relevant information from high-dimensional observations, while the baseline fails to achieve high rewards in both the training and testing environment.

2 Background

2.1 Partially-Observable Markov Decision Processes

RL is formalized as a Markov decision process (MDP), described by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{P}(s'|s,a)$ the probability of transitioning from state $s \in \mathcal{S}$ to state $s' \in S$ by executing action $a \in \mathcal{A}$, and $\gamma \in [0,1)$ a discount factor. An agent chooses actions according to a stochastic policy $\pi(a|s) \in [0,1]$, which yields the next state $s' \sim \mathcal{P}(s,a)$ and a scalar reward $r_s^a = \mathcal{R}(s,a)$. The agent's goal is to maximize the expected cumulative discounted rewards $\max_\pi \mathbf{E}_{\mathcal{P}}[\sum_{t=1}^{\infty}[\gamma^t \mathcal{R}(s_t)]]$.

A partially observable Markov decision process (POMDP) has, in addition, the set of observations Ω and the conditional observation probabilities O, such that at each time step, the agent receives the observation $o\in\Omega$ sampled from $O(o|s',a)$ conditioned on its action a and next state s'. While having access to only the observations means the environment is no longer Markovian from the agent's perspective, the agent can maintain its belief in the state (i.e: a distribution over states) and update its belief with its previous belief, the current observation, and the current action. Thus, the environment is Markovian in the agent's belief space.

2.2 Latent World Models

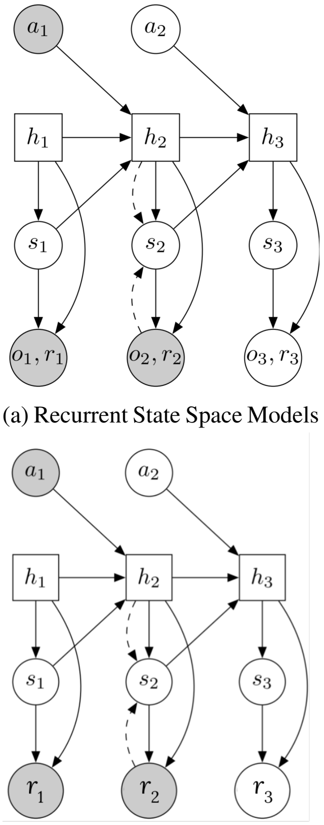
Latent World Models is a family of approaches for learning rewards and dynamics models of environments with high dimensional observations. Instead of working directly in the space of visual sensory data, an encoder infers low dimensional latent states from the high dimensional observations and learns the transition dynamics within this latent space. When used within RL, the policy and value networks can use the world model for optimization.

Since visual observations are partial observations of the ground state of any system, RL agents with access only to visual observations operate in POMDPs rather than MDPs. As such, their world is only Markovian in their belief. However, since their belief is sampled from their previous beliefs, their world is still Markovian in their entire past sequence of observations and actions, which can be approximated with recurrence.

Recurrent State Space Model (RSSM) [2] (Figure 1a) is a probabilistic recurrent dynamics model, which can be thought of as a sequential VAE. RSSM maximizes the Evidence Lower Bound (ELBO) of its observations $o_{1:T}$ and rewards $r_{1:T}$ in $\log p(o_{1:T},r_{1:T}|a_{1:T})$ with

$$\begin{aligned}&\log\int p(s_t|s_{t-1},a_{t-1})p_\theta(o_t|s_t)p_\theta(r_t|s_t)ds_{1:T}\\&\geq\sum_{t=1}^{T}\Big(\underbrace{\mathbb{E}[\log p_\theta(o_t|s_t)]+\mathbb{E}[\log p_\theta(r_t|s_t)]}_{\text{observation and rewards reconstruction}}\\&-\underbrace{\mathbb{E}[D_{KL}[q_\phi(s_t|o_{\leq t},a_{\leq t})||p(s_t|s_{t-1},a_{t-1})]]}_{\text{self consistent dynamics}}\Big)\end{aligned}\tag{1}$$

Here, the encoder $q_\phi(s_t|o_{\leq t},a_{\leq t})$ extracts a low dimensional latent state s_t from a sequence of observations and actions, and uses this latent state space for observation reconstruction $p_\theta(o_t|s_t)$, rewards inference $(r_t|s_t)$, and dynamics prediction $p(s_t|s_{t-1},a_{t-1})$.



(a) Recurrent State Space Models

(b) Our approach

Figure 1: Graphical models, where solid lines denote the generative process and dashed lines the inference model.

RSSM's accurate prediction has been exploited by Dreamer [1] to solve long horizon tasks in a sample efficient manner. To minimize its data requirements, Dreamer cycles between interacting with the environment, training its latent world model, then using its latent world model to "imagine" trajectories in latent space to optimize its policy. Since its latent representation, which comes from RSSM, is optimized for the reconstruction of the high dimensional observations, Dreamer is prone to visual distractors.

2.3 Bisimulation Metrics

In the RL state abstraction literature, stochastic bisimulation is an equivalence relation for partitioning an MDP's state space into behaviourally similar clusters. Two states s and s' are bisimilar if, for every action a, both states yield the same reward and the same distribution over the next states which are also bisimilar. The bisimulation metric for MDPs with continuous action spaces [4] softens the exact bisimulation relation, making it more applicable to noisy observation spaces. The bisimulation metric distance between two states s and s' is defined by [4] as

$$\max_{a\in\mathcal{A}}\big((1-\gamma)|r_s^a-r_{s'}^a|+\gamma W_1(\mathcal{P}(\cdot|s,a),\mathcal{P}(\cdot|s',a))\big)\tag{2}$$

where γ is the discount factor and W_1 is the Wasserstein (a.k.a "Earth Mover") probability metric. Bisimulation metrics have been applied to deep RL algorithms to learn robustly generalizable representations [6] without reconstruction. However, operating purely from visual observations, this approach can't be used for latent imagination for long-horizon RL tasks due to the inaccuracies of a single-step (i.e: non-recurrent) latent dynamics model.

3 Method

We propose Deep Bisimulation Dreaming (DBD), an algorithm that incorporates bisimulation metrics into latent world models to retain the generalizable and robust representations of the former and the sample efficiency of the latter.

3.1 Learning the world model

Our latent world model is built on top of RSSM [2] and includes all of its components except for its decoder. From a sequence of observations and actions from an episode, we use the encoder $q(s_{1:T}|o_{1:T},a_{1:T}) = \prod_{t=1}^{T} q(s_t|h_t,o_t)$ to sample the latent state s_t. The latent states s_t are used by our latent world model, which comprises of a deterministic state transition model $h_t = f(h_{t-1},s_{t-1},a_{t-1})$, a stochastic state transition model $s_t \sim p(s_t|h_t)$, and the rewards model $r_t \sim p(r_t|h_t,s_t)$. Our encoder, transition models, and rewards model are optimized to maximize the ELBO from equation 1 without the observation reconstruction term.

All our stochastic networks are Gaussian distributions, parameterized by the outputs of deep neural networks. Our encoder is deep convolutional neural network [7], our stochastic state model a multilayer perceptron (MLP), and our deterministic state model a recurrent neural network. The rewards model also outputs a Gaussian, but unlike the other three networks, parameterizes only the means, as opposed to both the means and variances.

Using a latent world model based on RSSM gives us accurate dynamic models that we can use for optimizing our policy through long horizon dreaming. Thus, our approach retains the sample efficiency of latent imagination.

3.2 Bisimulation Regularization

We wish to learn a latent state which encodes only task-relevant information. To do this, we follow a similar approach to [6] and optimize the L1 distance between two latent states s and s' to their bisimiulation metric distance

$$\mathcal{L}_{bisim} = \left(|s-s'| - (1-\gamma) \underbrace{|r_s^a - r_{s'}^a|}_{\text{rewards difference}} - \gamma W_2 \Big(\underbrace{p(\cdot|f(h,s,a)) \Big| p(\cdot|f(h',s',a))}_{\text{Transitions under our latent world model}} \Big) \right)^2 \tag{3}$$

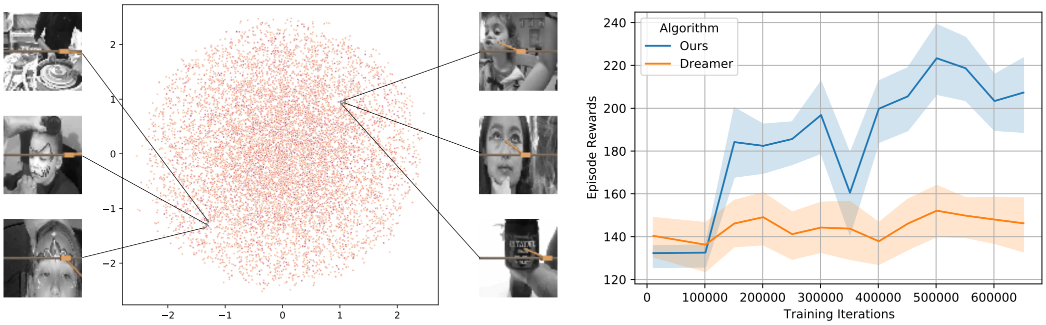
Intuitively, this loss encourages states which are behaviourally similar with respect to both their reward schemes and transition dynamics to be close together in latent space. Following [6], we use the 2-Wasserstein probability metric W_2 for the transition functions because it admits a simple closed-form for the Gaussian distributions from our encoder

$$W_2(\mathcal{N}(\mu_i,\sigma_i),\mathcal{N}(\mu_j,\sigma_j)) = \sqrt{||\mu_i - \mu_j||_2^2 + ||\sigma_i^{\frac{1}{2}} - \sigma_j^{\frac{1}{2}}||_{\mathcal{F}}^2}$$

where $||\cdot||_{\mathcal{F}}$ is the Frobenius norm. Given a batch of stochastic latent states $\{s_i\}$ sampled from our encoder, we minimize the mean pairwise bisimulation loss between the batch and a random permutation of the batch. Using the reparameterization trick, we can backpropagate through the random sampling of latent states.

4 Experiments

Through our experiments, we want to investigate the following questions: (1) Can our approach achieve high rewards in the presence of distractors and generalize to held-out distractors? (2) Does our representation lead to more accurate latent dynamics and rewards models? and (3) Does our encoder



(a) T-SNE embedding of feature space

(b) Performance against held-out distractors

Figure 2: From visual observations with distractors, our approach extracts representations where pairwise distances in the latent space reflects to differences in task-relevant details (a), enabling it to achieve high rewards even when facing novel distractors (b).

extract meaningful information from high-dimensional visual observations? We compare against Dreamer [1], which learns a latent state through the reconstruction of visual inputs.

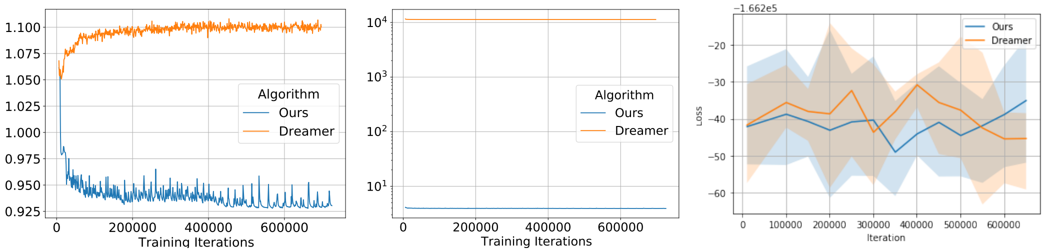
Tasks with Visual Distractors. We chose to test on the Deep Mind Control Suite's Cart Pole task because the ground truth states necessary to solve the task includes only 4 scalars, making it a very simple task given the ground truth state. However, agents must solve tasks from noisy visual sensory data with distractors. For each observation, we mask the agent out and replace the background with grayscaled videos from the Deep Mind Kinetic Dataset [8]. Our training background videos uses 25,000 frames from 1653 videos, coming from the categories "arranging flowers", "blowing glass", and "carving pumpkin" (Figure 2a). Our held-out testing background videos uses 25,000 frames from 414 videos from the categories "brush painting" and "clay pottery making".

Evaluation Metric. To evaluate the agent's performance, we measure the agent's cumulative discounted rewards, averaged over 100 episodes with distractors not seen during training.

Convergence Metric. Since maximizing the variational lower bound [9] includes an observation reconstruction term, we instead choose to monitor the correlation between our latent states and the 4 scalars describing the ground truth system state (which is hidden from the agent, but we have access to through the simulation environment). We say that a state representation encodes more task-relevant details if it correlates with the ground truth state. Specifically, given a collection of ground truth states and the corresponding collection of latent states, we measure the rank correlation in pairwise Euclidean distances between two collections. More details on this metric can be found in the supplementary material.

A task-relevant latent state. Quantitatively, our approach achieves high rewards in the presence of unseen background videos, suggesting that our representation ignores distractors and extracts task-relevant latent states (Figure 2b). Qualitatively, behaviourally similar visual states are encoded closer together in latent space (Figure 2a). In contrast, the baseline fails to separate task-relevant information from distractors, as a result of maximizing the ELBO of the visual observations through reconstruction. This result shows that generic representation techniques, such as VAEs [3], may not be well-suited for RL applications, where extracting only information about rewards and dynamics enables better generalization.

Efficient rewards and dynamics modeling. We compare our approach's rewards and dynamics model loss curve with the baseline in Figure 3a and Figure 3b respectively. The baseline's rewards loss is 15% higher than that of our approach, while its dynamics model loss is close to 4 orders of magnitude larger than ours. Recall the dynamics model is optimized for self-consistency with the KL divergence from equation 1. Since bisimilar states have identical transition probabilities, a dynamics model operating on our latent states has a much easier task in predicting dynamics. Indeed, bisimulation is a state abstraction technique to reduce large state spaces to clusters of bisimilar states, and any policy, value network, or dynamics model trained on this reduced bisimilar state space will learn more efficiently. Meanwhile, Dreamer encodes states that are bisimilar but with different visual distractors to different latent states, putting more burden on the dynamics model to fit this wider variance of latent states and all its transitions.



(a) Training Rewards Loss (b) Training Dynamics Loss (c) Rank Correlation with True States

Figure 3: Our representations achieves more accurate latent world models (a) (b), despite seemingly not correlating with the ground truth states (c).

Ground Truth State Recovery. A policy observing ground-truth states only requires 4 scalars (the pole's angle, angular velocity, the cart's position, and velocity) to solve the task. In this experiment, we investigate whether our latent representation correlates with hidden ground truth states. From figure 3c, we do not observe any meaningful relationship between our latent states or the baseline's latent states with ground truth states. We hypothesize that this might be reflective of our rank correlation measure. Perhaps, a better way to test for correlation might be to train a single layer MLP with no non-linearity to regress ground truth states from latent embeddings and compare the final regression accuracy.

5 Conclusion

In this work, we have proposed Deep Bisimulation Dreamer, an algorithm for learning task-relevant latent representations using the bisimulation metric. We show that it performs favorably on continuous action space robotics tasks within the framework of planning in latent spaces.

Due to the compute requirements of each experiment (approximately 1 day to converge on a GTX 1080 Ti) and limited time constraints, we were not able to test our approach on a wide variety of tasks with multiple different seeds. Moving forward, it is crucial to make sure our approach learns stably and robustly achieves high performance on a wide variety of tasks, including more challenging tasks requiring long horizon reasoning. Given the promising results, we would also like to test with more challenging colored visual distractors as opposed to just grayscaled distractors. Additionally, though our approach outperforms Dreamer by a large margin, its absolute performance is still low for the Cartpole task, whose average successful rewards are at least 600.

An interesting tangent to this current research would be to consider the bisimulation metric as an add-on service rather than a replacement for observation reconstruction. The bisimulation metric could be used to optimize the latent space, as performed in this report, in conjunction with the reconstruction loss. Also, our evaluation excluded optimizing the ELBO since we removed the reconstruction loss. However, there is budding research [10] in a robust optimization objective that avoids reconstructing the observations while still lower bounding the original ELBO. This could offer a more robust metric and optimization objective for representation learning in reinforcement learning.

References

[1] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

[2] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

[3] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2014.

[4] N. Ferns, P. Panangaden, and D. Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

[5] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi:10.1109/IROS.2012.6386109.

[6] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[7] Y. LeCun, Y. Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[8] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[9] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[10] X. Ma, S. Chen, D. Hsu, and W. S. Lee. Contrastive variational model-based reinforcement learning for complex observations. *arXiv preprint arXiv:2008.02430*, 2020.

6 Supplementary Material

6.1 Rank Correlation Metric

To evaluate the quality of an encoded representation, we devise a metric, order distance. For a set of encoded representations $v_1,v_2,...,v_n$ and their corresponding ground truth states $g_1,g_2,...,g_n$, we can compute the matrices, D_{True} and D_{Rep}, that specify the pairwise euclidean distances between the ground truth vectors and between the encoded representations respectively.

By considering the pairwise distance matrices, D_{True} and D_{Rep}, our metric can be invariant with respect to different transformations of the data (e.g. rotations, translations, ...). To determine if the pairwise distance information is preserved, we can consider the order of the nearest neighbors for v_i and g_i. If the encoded representations and the ground truth states have the same qualitative pairwise distance information, then the order of the nearest neighbors for v_i and g_i will be the same. Otherwise, we can impose a loss based on how different the orders are. For this implementation, our loss is the following:

$$L_i = -\sum_{j=1}^{n} w(\pi(j,v_i)) \cdot |\pi(j,v_i) - \pi(j,g_i)|$$

where $\pi(i,v)$ represents the position of data point i in the nearest neighbor ordering of $v_1,...,v_n$ with respect to v. Additionally, we can weight the accumulated losses by a function w so that bad orderings for the points nearby v incur a higher penalty than those that are further away.