Constrained Contrastive Reinforcement Learning

Haoyu Wang HAOYU59@FOXMAIL.COM
Xinrui Yang XINRUI.YANG@STU.XJTU.EDU.COM
Yuhang Wang YHW.WERTY@QQ.COM
Xuguang Lan XGLAN@MAIL.XJTU.EDU.CN
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Editors: Emtiyaz Khan and Mehmet Gonen

Abstract

Learning to control from complex observations remains a major challenge in the application of model-based reinforcement learning (MBRL). Existing MBRL methods apply contrastive learning to replace pixel-level reconstruction, improving the performance of the latent world model. However, previous contrastive learning approaches in MBRL fail to utilize task-relevant information, making it difficult to aggregate observations with the same task-relevant information but the different task-irrelevant information in latent space. In this work, we first propose Constrained Contrastive Reinforcement Learning (C2RL), an MBRL method that learns a world model through a combination of two contrastive losses based on latent dynamics and task-relevant state abstraction respectively, utilizing reward information to accelerate model learning. Then, we propose a hyperparameter β to balance two kinds of contrastive losses to strengthen the representation ability of the latent dynamics. The experimental results show that our approach outperforms state-of-the-art methods in both the natural video and standard background setting on challenging DMControl tasks.

Keywords: Model-based reinforcement learning; Representation learning; Contrastive learning

1. Introduction

Deep Reinforcement Learning (DRL) achieves great success in various domains such as game playing (Mnih et al., 2013; Berner et al., 2019; Vinyals et al., 2019), autonomous driving (Chen et al., 2021). In model-based reinforcement learning (MBRL), a dynamic world model predicts the observation in latent space, improving sample efficiency and enabling generalization (Thrun and Littman, 2000; Lee et al., 2020; Hafner et al., 2019a). Previous MBRL methods (Hafner et al., 2019a) learn a latent world model by minimizing the reconstruction error of the past observations (Hafner et al., 2019a). However, pixel-level reconstruction leads to the waste of representation capacity to capture the information that is unpredictable or task-irrelevant. For example, a robot arm should ignore the random background pixels when grasping objects.

To avoid the limitation of reconstruction methods, recent work applies contrastive learning on RL for latent world model learning (Dwibedi et al., 2018; Zhang et al., 2020; Liu et al., 2021). Contrastive Learning maps the positive inputs to be close while mapping the negative inputs to be further away in latent space (Le-Khac et al., 2020; He et al., 2020). To achieve better performance, especially in tasks with noisy pixels, different approaches apply different division methods of positive and negative samples (Laskin et al., 2020; Ma et al., 2020; Okada and Taniguchi, 2021; Zhang et al., 2020; Nguyen et al., 2021). However, prior contrastive learning approaches in MBRL fail to utilize task-relevant information, making it difficult to aggregate observations with the same task-relevant information but with different task-irrelevant information in latent space (Nguyen et al., 2021). Since these methods aim to strengthen the ability to represent each observation, the observations happen to be mapped further away in the latent space. However, when the different observations refer to a similar state of the task in some cases, the latent dynamics may fail to capture available information for the agent. Considering a "reacher" task that a robot arm tries to reach a target and get rewards. In an episode, the arm may reach the target multiple times. Intuitively, all the observations when the robot arm reaches the target should be regarded as similar samples, namely positive samples. But with task-irrelevant division methods, they will be, by contrast, regarded as dissimilar samples, namely negative samples, which will result in poor representation learning.

In this paper, we propose Constrained Contrastive Variational Reinforcement Learning (C2RL), an MBRL method that learns the world model through a combination of two contrastive losses based on latent dynamics and task-relevant state abstraction respectively. We name the two loss functions same-segmentation contrastive learning (SSCL) and different-segmentation contrastive learning (DSCL). Inspired by Contrastive Variational Model-Based Reinforcement Learning (CVRL) (Ma et al., 2020), we introduce SSCL to increase the distance between one observation and the others in latent space to strengthen the representation ability of the latent dynamics. Compared with CVRL, SSCL considers only the observation itself and its corresponding latent state as the positive samples in each episode. To decrease the distance among the observations in latent space that contain similar task-relevant information, DSCL is proposed to aggregate the observations with similar returns while discriminating the observations with dissimilar returns in latent space. Inspired by Z^{π}-irrelevance abstraction (Liu et al., 2021), DSCL considers the observations in the whole replay buffer that has similar online scaled returns as the positive samples. In the case of the conflict sample division between SSCL and DSCL, we also introduce a hyperparameter β to balance the two different contrastive losses which will be detailed in Section 3.1. The key contributions of our algorithm are summarized:

- **Utilizing task-relevant information to accelerate model learning** We utilize task-relevant information for representation learning through state abstraction to accelerate model learning.

- **Balancing contrastive learning** We propose a hyperparameter to balance the two kinds of contrastive losses to strengthen the representation ability of the latent dynamics.

We choose Standard DeepMind Control (DMC) tasks (Tassa et al., 2018) and Natural DeepMind Control tasks (Ma et al., 2020) which replace the background of DMC tasks with random videos for experiment. Compared with recent MBRL methods, C2RL achieved comparable or better performance in most tasks. We performed a detailed analysis of the Standard and Natural experiments in Section 3.

2. Backgrounds

2.1. Related Works

World Models: In MBRL, world models are trained as the core component to extract features from high-dimensional observations, which can be used to control with multi-step predictions or multi-step rewards. The majority relies on sequential variational autoencoders, which aim to reconstruct observations by optimizing the evidence lower bound (ELBO), to capture the stochastic dynamics of the environment via three kinds of losses: reconstructing observations in pixel spaces, reconstructing rewards, constraining latent dynamics (Ha and Schmidhuber, 2018). Lee et al. (2020) learn the world model with hierarchical stochastic latent states. Hafner et al. (2019b,a) propose RSSM models based on GRU (Cho et al., 2014) with stochastic latent states and deterministic latent states. Zhang et al. (2020) accelerate learning world models using bisimulation metrics. However, these methods all fail to reconstruct high-dimensional observations such as images with noisy pixel backgrounds in complex tasks, which leads to an accumulated compositional error of the world model (Ma et al., 2020; Nguyen et al., 2021).

Contrastive Learning in MBRL: Contrastive learning is widely used as an auxiliary task to construct world models in MBRL by scoring positive sample pairs and negative sample pairs, motivated from different perspectives. Van den Oord et al. (2018); Guo et al. (2018); Anand et al. (2019); Mazoure et al. (2020) propose different temporal contrastive losses in reinforcement learning environments. Laskin et al. (2020) propose data augmentation to construct positive samples. Ghosh et al. (2018); Castro (2020); Zhang et al. (2020); Liu et al. (2021) use elements based on reinforcement learning to construct positive samples and negative samples. Ma et al. (2020) optimizes contrastive losses by maximizing the mutual information between latent states and observations. Nguyen et al. (2021) present an information-theoretic approach through contrastive learning. Although existing methods have proposed different losses for contrastive learning, they failed to aggregate observations with the same task-relevant information but the different task-irrelevant information in latent space. Compared to the mentioned methods, our work focuses on the combination of world models and task-relevant contrastive learning based on state abstraction to solve this problem.

2.2. Preliminary

Sequential Latent World Model: Visual input contains only part of the information of the environment state, thus we assume the environment can be formulated as a partially observable Markov decision process (POMDP). We define discrete time step as t, high dimensional observation as o_t, action as a_t, reward as r_t and latent state as z_t. In POMDP, we can sample sequential data by interacting with the environments and build Sequential Latent World Model for learning representation and decision making. We use three parts of the recurrent state-space model (RSSM) (Hafner et al., 2019b) as our model:

$$\begin{aligned}
&\text{Deterministic state model}: && h_t = f(h_{t-1}, s_{t-1}, a_{t-1}) \\
&\text{Stochastic state model}: && s_t \sim p(s_t|h_t) \\
&\text{Reward model}: && r_t \sim p(r_t|h_t, s_t)
\end{aligned} \tag{1}$$

Where latent state $z_t = [s_t, h_t]$ includes the deterministic state and stochastic state and z_t are divided into posterior state and prior state. Previous works generalize Evidence Lower Bound (ELBO) of VAEs from Equation (2) (Kingma and Welling, 2013) to world models of MBRL (Ha and Schmidhuber, 2018), where $p(z)$ denotes prior distribution of z and $q(z|o)$ denotes the proposal distribution that samples z conditioned on the observation o.

$$\log p(o) = \log \int_z p(o|s)p(z)dz \geq E_{q(z|o)}[p(o|z)] - KL[q(z|o)||p(z)] \tag{2}$$

Given sequential samples from an episode, we can derive the variational bound from Equation (3). Maximizing the data log-likelihood $\log p(o_{1:T}, r_{1:T}|a_{1:T})$ can be converted to maximize the variational bound (Hafner et al., 2019b).

$$\begin{aligned}\log p(o_{1:T}, r_{1:T}|a_{1:T}) &= \log \int \prod_t p(z_t|z_{t-1}, a_{t-1})p(o_t|z_t)p(r_t|z_t)dz_{1:T} \\ &\geq \sum_{t=1}^{T} \Bigg(\underbrace{E_{q(z_t|o_{\leq t}, a_{<t})}[\log p(o_t|z_t)]}_{\text{observation reconstruction}} + \underbrace{E_{q(z_t|o_{\leq t}, a_{<t})}[\log p(r_t|z_t)]}_{\text{reward reconstruction}} \\ &\quad - \underbrace{E_{q(z_{t-1}|o_{\leq t-1}, a_{<t-1})}[KL[q(z_t|o_{\leq t, a_{<t}})||p(z_t|z_{t-1}, a_{t-1})]]}_{\text{latent dynamics}} \Bigg)\end{aligned} \tag{3}$$

However, reconstructing observations in pixel space means that we encode all the information of observations with thousands of dimensions into latent space regardless of whether the information is valid or not (Ha and Schmidhuber, 2018). To tackle this problem, observation reconstruction can be replaced with contrastive learning (Ma et al., 2020). The variational bound can be rewritten as Equation (4) where the latent state z_t maps the corresponding observation o_t and is distinguished from other observations with contrastive learning. In the Equation (4), O_t is a set of irrelevant observations sampled from a replay buffer, $f_\theta(o_t, z_t)$ is a trainable non-negative function that measures the compatibility between latent state z_t and observation o_t with parameter θ.

$$\begin{aligned}\log p(o_{1:T}, r_{1:T}|a_{1:T}) \geq \sum_{t=1}^{T} \Bigg(&\underbrace{E_{q(z_t|o_{\leq t}, a_{<t})}[\log \frac{f_\theta(o_t, z_t)}{\sum_{o'_t \in O_t,} f_\theta(o'_t, z_t)}]}_{\text{contrastive learning}} + \\ &\underbrace{E_{q(z_t|o_{\leq t}, a_{<t})}[\log p(r_t|z_t)]}_{\text{reward reconstruction}} - \underbrace{E_{q(z_{t-1}|o_{\leq t-1}, a_{<t-1})}[KL[q(z_t|o_{\leq t, a_{<t}})||p(z_t|z_{t-1}, a_{t-1})]]}_{\text{latent dynamics}} \Bigg)\end{aligned} \tag{4}$$

State Abstraction: State abstraction such as bisimulation metric (Givan et al., 2003) is a way to address high-dimensional sensory inputs, aggregating similar states to the one abstract state, which results the reduction of the state space. However, it is computationally expensive to directly perform bisimulation metric in reinforcement leaning, in which case Liu et al. (2021) demonstrate Z^π-irrelevance as the simplification of bisimulation.

When performing Z^{π}-irrelevance, the returns are empirically divided into K discrete equal bins $[R_0, R_1...R_K]$. The samples from the same discrete bin are mapped as the same abstraction while the samples from different bins are mapped different abstraction (Liu et al., 2021). In world models, the latent space can also be modeled as low-dimensional abstraction space in which similar samples should be aggregated. One reasonable latent space should satisfy both the variational bound and state abstraction. In this paper, we focus on the abstraction constraint of samples from different episodes.

3. Constrained Contrastive Reinforcement Learning

We introduce Constrained Contrastive Reinforcement Learning (C2RL), an MBRL framework for complex image-input tasks. The basic architecture of C2RL is built upon Dreamer (Hafner et al., 2019a), learning behaviors based on imagination in a latent world model. For decision-making, C2RL uses model predictive control (MPC) as CVRL does, maximizing the expected future return through latent gradients (Ma et al., 2020). As the major difference, C2RL introduces four main components for world model learning: reward reconstruction and latent dynamics as in Equation (4), the same-segment contrastive learning (SSCL), and the different-segment contrastive learning (DSCL). In this part, we first introduce SSCL and DSCL theoretically, then present practical implementation details of C2RL.

3.1. SSCL and DSCL

SSCL is namely the contrastive learning method proposed in CVRL (Ma et al., 2020) but chooses sample pairs from a single episode instead of the whole replay buffer. Thus, SSCL can be described as Equation (4) but with a different definition of O_t. SSCL increases the distance between one observation and the others in latent space to strengthen the representation ability of the latent dynamics. Since SSCL aims to discriminate each observation in an episode, the observations are naturally mapped further away in the latent space. However, it does not explicitly constrain whether the two latent states should be aggregated or not. When the different observations refer to a similar state of the task in some cases, the latent dynamics may fail to capture available information for the agent. To tackle this problem, we propose DSCL aggregate the observation with a similar return while discriminating the observation with dissimilar returns in latent space. Figure 1 demonstrates the different sample division methods of SSCL and DSCL. It is worth noting that DSCL includes the samples in the same episodes except for the samples from different episodes.

Considering a division set $\Phi_O = \left\{(o_i, o_j)|o_i, o_j \in O, o_i \sim p(o_i), o_j \sim p(o_j), \hat{\phi}(o_i) = \hat{\phi}(o_j)\right\}$, where $\hat{\phi}$ is the encoder in state abstraction, and $p(o_t) = p(o_t|o_{<t}, a_{<t})$. Given $z_t \in Z$, we define a set $\psi(z_t) = \{o_j|(o_t, o_j) \in \Phi_O\}$ which contains all the positive samples of o_t. We use DSCL constraint as an auxiliary task of major optimization in Equation(4). DSCL constraint based on a variant of original InfoNCE (Van den Oord et al., 2018) to aggregate the latent states considering state abstraction:

$$\sum_{t=1}^{T} \left(E_{q(z_t|o_{\leq t}, a_{<t})} \log \frac{\sum_{o_j \in O_t,} f_\theta(o_j, z_t)}{\sum_{o_j \in \psi(z_t)} f_\theta(o_j, z_t)} \right) \leq \epsilon \tag{5}$$

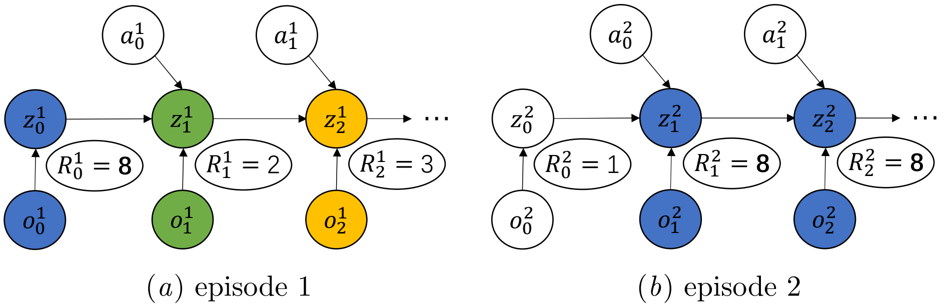


Figure 1: Division methods of SSCL and DSCL. In a single episode(**left**), SSCL considers observation and latent state pairs with the same color as positive samples and considers observations and latent state pairs with different colors as negative samples. In episodes in replay buffer(**left and right**), DSCL considers observation and latent state pairs with the same or similar returns as positive samples(marked as the same color also) and considers others as negative samples. R refers to the return of the state.

where $\epsilon > 0$ specifies the strength of the constraint, O_t is a set of irrelevant observations sampled from a replay buffer, $f_\theta(o_t, z_t)$ is a non-negative function that measures the compatibility between latent state z_t and observation o_t. For any $(o_i, o_j) \in \Phi_O$, we desire that the corresponding latent states (z_i, z_j) are aggregated. The intuition for Equation (5) is that we want to maximize the compatibility between the latent state z and a set of corresponding observations $\psi(z_t)$ (positive sample) while minimizing its compatibility between a set of irrelevant observations (negative samples), with the different division of the episodes (Ma et al., 2020; Liu et al., 2021). Based on Equation(4) and Equation(5), we can derive the joint optimization:

$$\begin{aligned}
&\max \log p(o_{1:T}, r_{1:T} | a_{1:T}) \\
&\text{subject to.} \sum_{t=1}^{T} \left(E_{q(z_t|o_{\leq t}, a_{<t})} \log \frac{\sum_{o_j \in O_t,} f_\theta(o_j, z_t)}{\sum_{o_j \in \psi(z_t)} f_\theta(o_j, z_t)} \right) \leq \epsilon
\end{aligned} \tag{6}$$

Equation(6) can be rewritten as a Lagrangian using KKT conditions (Bertsekas, 1997):

$$\begin{aligned}
L &= \log p(o_{1:T}, r_{1:T} | a_{1:T}) - \beta \sum_{t=1}^{T} \left(E_{q(z_t|o_{\leq t}, a_{<t})} \log \frac{\sum_{o_j \in O_t,} f_\theta(o_j, z_t)}{\sum_{o_j \in \psi(z_t)} f_\theta(o_j, z_t)} - \epsilon \right) \\
&> \log p(o_{1:T}, r_{1:T} | a_{1:T}) - \beta \sum_{t=1}^{T} \left(E_{q(z_t|o_{\leq t}, a_{<t})} \log \frac{\sum_{o_j \in O_t,} f_\theta(o_j, z_t)}{\sum_{o_j \in \psi(z_t)} f_\theta(o_j, z_t)} \right)
\end{aligned} \tag{7}$$

where $\beta > 0$ is the regularisation coefficient for constraining state abstraction. Since $\beta, \epsilon > 0$, we can obtain the lower bound in Equation(7). Introducing the lower bound in Equation(4), we obtain the final lower bound optimization in Equation(9), which includes four models: same-segment contrastive learning(SSCL), reward reconstruction, latent dynamics, and different-segment contrastive learning(DSCL).

$$\begin{aligned}
&\text{posterior model} : q_\phi \\
&\text{SSCL model} : f_{\theta_1} \\
&\text{reward reconstruction} : p_\theta \\
&\text{prior model} : p_\omega \\
&\text{DSCL} : f_{\theta_2}
\end{aligned} \tag{8}$$

$$\begin{aligned}
L > \sum_{t=1}^{T} \Bigg(& \underbrace{E_{q_\phi(z_t|o_{\leq t}, a_{<t})}[\log \frac{f_{\theta_1}(o_t, z_t)}{\sum_{o_t^{'} \in O_t,} f_{\theta_1}(o_t^{'}, z_t)}]}_{\mathcal{L}_{\text{SSCL}}} + \underbrace{E_{q_\phi(z_t|o_{\leq t}, a_{<t})}[\log p_\theta(r_t|z_t)]}_{\mathcal{L}_{\text{RR}}:\text{reward reconstruction}} \\
& - \underbrace{E_{q_\phi(z_{t-1}|o_{\leq t-1}, a_{<t-1})}[KL[q_\phi(z_t|o_{\leq t}, a_{<t})||p_\omega(z_t|z_{t-1}, a_{t-1})]]}_{\mathcal{L}_{\text{LD}}:\text{latent dynamics}} \Bigg) \\
& - \underbrace{\beta \sum_{t=1}^{T} \left(E_{q_\phi(z_t|o_{\leq t}, a_{<t})} \log \frac{\sum_{o_j \in O_t,} f_{\theta_2}(o_j, z_t)}{\sum_{o_j \in \psi(z_t)} f_{\theta_2}(o_j, z_t)} \right)}_{\mathcal{L}_{\text{DSCL}}}
\end{aligned} \tag{9}$$

ϵ and β: DSCL is weighted with a parameter β. Varying the parameter β changes the influence of state abstraction pressure during training, which encourages similar observations can be encoded into the corresponding latent states. We optimize the parameter ϵ in Equation(7) indirectly by changing the value of β . For different tasks, the optimal ϵ is different and β depends on the value of ϵ, thus the optimal β is difficult to determine. In this paper, we empirically choose β as a hyperparameter that is not optimized during training.

Different compatibility functions: The two contrastive learning losses may conflict in some samples. This is because SSCL maps the latent state s_t to o_t while DSCL maps the latent state s_t to other observations except for o_t. Since the compatibilities in SSCL and DSCL are not equivalent, the functions $f_\theta(s_t, o_t)$ that measure the compatibility between latent state and observation are different in SSCL and DSCL.

3.2. Practical Implementation Details

Sample division: When optimizing the loss $\mathcal{L}_{\text{DSCL}}$ in experiments, we introduce two hyperparameters to divide the positive and negative samples. The first hyperparameter is N, the number of bins for discrete limited returns. The second hyperparameter M is the number of adjacent samples for calculating the limited returns. Given sequential samples, we calculate the limited returns by summing rewards of the adjacent M samples. Then we discretize the limited returns in nth bin $n \in \{0, 1, 2...N\}$. We record the maximum and minimum limited returns of samples R_{max} and R_{min} during training. The positive samples of $\mathcal{L}_{\text{DSCL}}$ are determined by discrete limited returns: given sequential samples, we first calculate limited returns by summing rewards of the adjacent M samples, and then the limited returns are discretized into N values.The discretized limited return of the sample $\{s_t, a_t, r_t, s_t+1\}$ can be calculated:

$$R_t^{lim} = [\frac{\sum_{i=t-M+1}^{t} r_i}{\delta + \frac{(R_{max}-R_{min})}{N}}] \tag{10}$$

where δ is a small value to deal with the case in which $R_{max} = R_{min}$ and the denominator becomes zero; [.] indicates rounding. The samples with the same discretized limited return are considered as the positive samples of $\mathcal{L}_{\text{DSCL}}$ in Equation(9).

Behavior Learning: In C2RL, we use latent imagination to learn a parameterized policy for control following (Ma et al., 2020). For selfcontainedness, this section gives a summary of this approach. Firstly, the action model and the value model are the main components for behavior learning. The action model $a_\tau \sim q_\eta(a_\tau|s_\tau)$ is parameterized as a tanh-transformed Gaussian model. The value model $v_\mu(s_\tau)$ is estimated using the imagined trajectories$\{s_\tau, a_\tau, s_\tau\}_{\tau=t}^{t+H}$, where H is the imagine horizon based on the latent dynamics.

$$\begin{aligned} &a_\tau = \tanh(\mu_\eta(s_\tau) + \sigma_\eta(s_\tau)\epsilon), \epsilon \sim \text{Normal}(0, \mathbb{I}) \\ &v_\mu(s_\tau) = E_{q_\eta} \sum_{t=\tau} \gamma^{t-\tau} r_t \end{aligned} \tag{11}$$

To learn the action and value model, in this work, we use value estimation presented in (Ha and Schmidhuber, 2018). The values are estimated by TD(λ) (Thrun and Littman, 2000) which trade-off bias and variance.

$$\begin{aligned} V_N^k(s_\tau) &\doteq E_{q_\eta, p_\theta}(\sum_{t=\tau}^{h-1} \gamma^{t-\tau} r_t + \gamma^{h-t} v_\mu(s_h)) \\ V_\lambda(s_\tau) &\doteq (1-\lambda) \sum_{n=1}^{H-1} \lambda^{n-1} V_N^n(s_\tau) + \lambda^{H-1} V_N^H(s_\tau) \end{aligned} \tag{12}$$

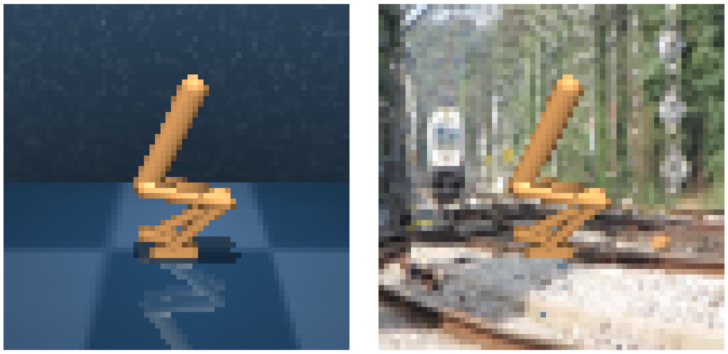
where $h = \min(\tau + k, t + H)$. The values are estimated under the imagined trajectories. The actor model is optimized to maximize the imagined value estimates. The value model is trained to regress the value estimates.

$$\max_\eta E_{q_\eta, p_\theta}(\sum_{\tau=t}^{t+H} V_\lambda(s_\tau)) \tag{13}$$

$$\min_\mu E_{q_\eta, p_\theta}(\sum_{\tau=t}^{t+H} \frac{1}{2} \|v_\mu(s_\tau) - V_\lambda(s_\tau)\|^2) \tag{14}$$

4. Experiments

In this section, we empirically evaluate our C2RL model in various settings. First, we design experiments to compare the relative performance of our model with the state-of-the-art methods in both the natural-video and standard background setting on challenging



(*a*) Standard Mujoco control tasks (*b*) Natural Mujoco control tasks

Figure 2: We show the observation of Standard Mujoco control task walker walk and Natural Mujoco control task walker walk. The difference is that the Natural Mujoco control task includes more noisy pixels.

DMControl tasks (Tassa et al., 2018), as is shown in Figure 2. Second, we evaluate the cosine similarity in both positive and negative samples of C2RL and discuss how our model is superior in the latter case compared to other existing methods. Finally, we analyze the ability of the hyperparameter, in which we test the performance of the agent with different values of β.

4.1. DMControl

For the model training, two types of hyperparameters need to be determined in C2RL. The first pair is the hyperparameters M and N to calculate discrete limited returns. We determine the values of the hyperparameters by searching (M, N) in the set $\{(10, 30), (10, 40), (10, 50), (10, 60), (60, 60)\}$. We choose $(M, N) = (10, 60)$ in the benchmark tasks except for the Cup Catch task; $(M, N) = (60, 60)$ in the benchmark task cup catch. The second hyperparameter β is searched in the candidate set $\{0.0, 0.5, 1.5, 3.0\}$. We have found that $\beta = 1.5$ performs best in the candidate set. We performed a search for the two types of hyperparameters on the tasks Walker Walk task and Cup Catch task. All of the tasks except the Cup Catch task with sparse reward share the same hyperparameters. We compare C2RL with the following algorithms: CVRL(Ma et al., 2020), which is the baseline of C2RL; TPC(Nguyen et al., 2021), which is a recent model-based method based on information theory.

Natural Mujoco control tasks: Figure 3 shows that C2RL (ours) achieves better performance on 4 out of 5 tasks in final scores and data efficiency compared with the other two baselines. C2RL (ours) achieves comparable performance in the remaining task. Since the framework of C2RL is an extension of CVRL, we can see that C2RL brings improvement over the base algorithm in the five tasks.

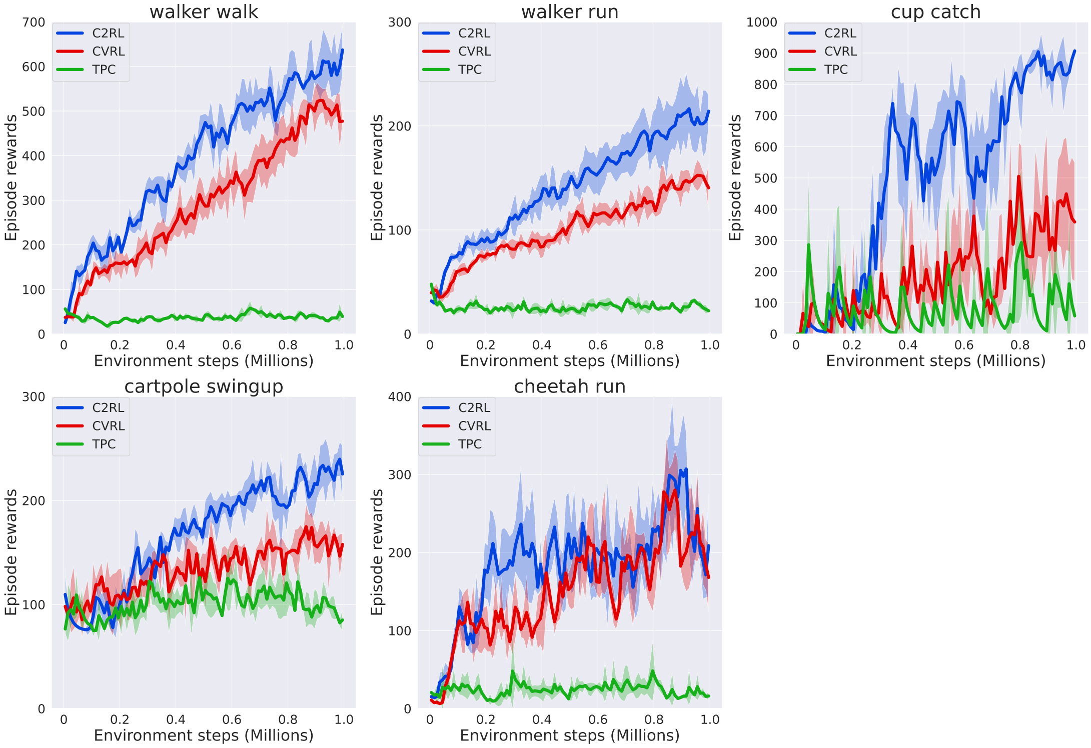


Figure 3: Experiments on five Natural Mujoco control tasks, where the background is replaced with images from the natural videos. The solid line and the shaded area indicate the mean and the standard deviation over 3 different seeds respectively.

Standard Mujoco control tasks: As is shown in Figure 4, C2RL achieves limited performance improvement on 3 out of 5 tasks in final scores compared with the other two baselines, while achieving comparable performance in the remaining two tasks.

4.2. Analysis of Latent States

To analyse how the two different losses $\mathcal{L}_{SSCL}$ and $\mathcal{L}_{DSCL}$ affect the latent space, we use cosine similarity to measure the similarity of different latent states compared with CVRL. Given two samples $o_i \sim p(o_t|o_{<i}, a_{<i})$ and $o_j \sim p(o_t|o_{<j}, a_{<j})$, the corresponding latent states are s_i and s_j respectively. The cosine similarity $S(s_i, s_j) \in [-1, 1]$ can be written as:

$$S(s_i, s_j) = \frac{s_i^T \cdot s_j}{\|s_i\| \|s_j\|} \tag{15}$$

The larger the cosine similarity, the more similar the two latent states are. We show the similarities between positive pairs and negative pairs by using cosine similarity. We only

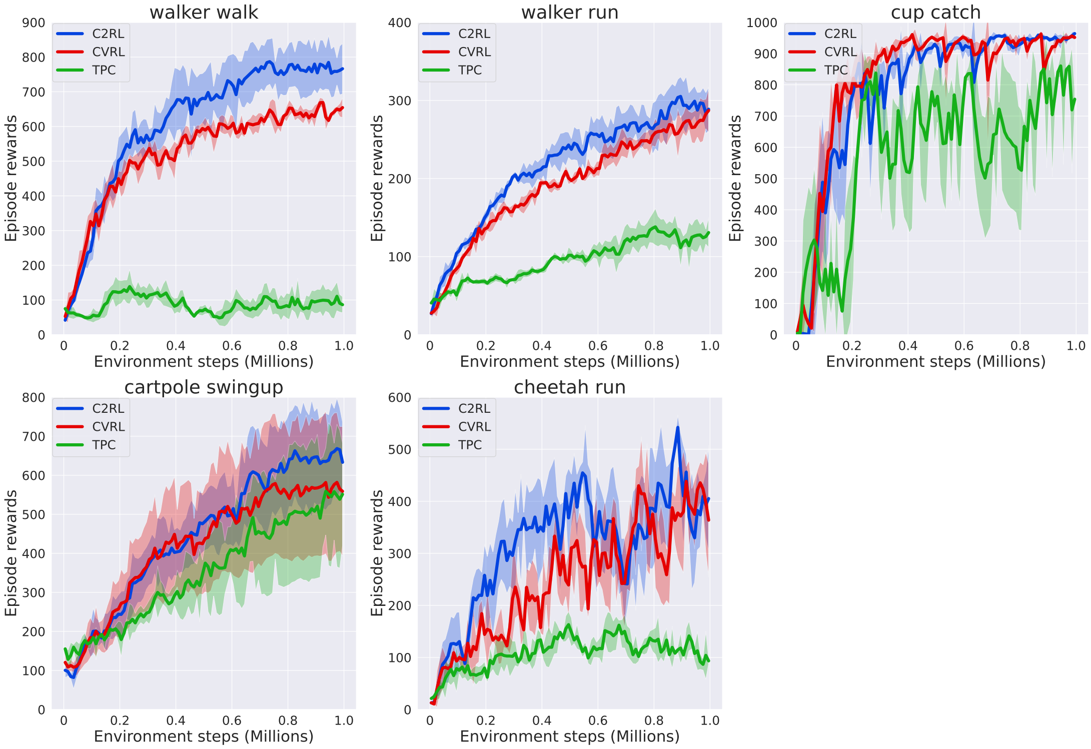
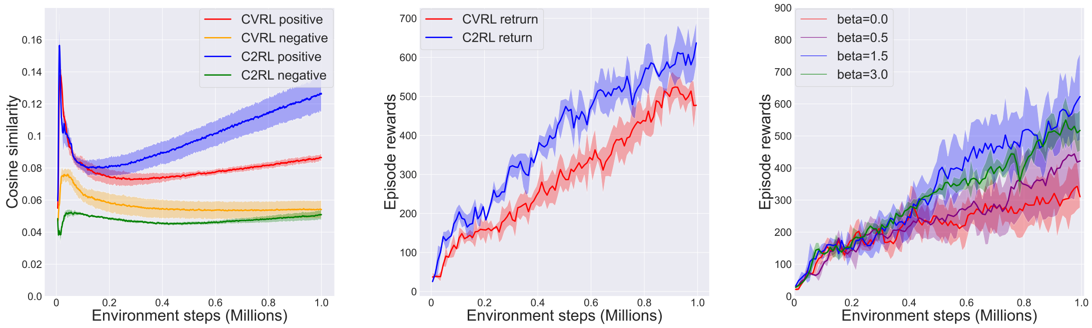


Figure 4: Experiments on five Standard Mujoco control tasks, where the background is not changed. The solid line and the shaded area indicate the mean and the standard deviation over 3 different seeds respectively.

consider the samples for $\mathcal{L}_{DSCL}$ since they can be from different episodes. Figure 5 (a,b) shows 1) The similarity of positive pairs of C2RL (blue) is greater than that of CVRL (red) and the similarity of positive pairs of C2RL (yellow) is close to that of CVRL (green). To a certain extent, the distance among the observations in latent space that contain similar task-relevant information has been successfully achieved. 2) The cosine similarity is much less than 1 for both C2RL and CVRL, due to our simultaneous optimization of two losses $\mathcal{L}_{SSCL}$ and $\mathcal{L}_{DSCL}$. In latent space, SSCL may increase the distance between one observation and the others, while DSCL may decrease the distance between the observations with the same returns. 3) Combining the two losses, the performance has been significantly improved by aggregating the observations with a similar return while discriminating the observations with dissimilar returns in latent space.



(a) similarity of sample pairs (b) performance on Natural Mujoco (c) performance with different β

Figure 5: (a,b) The experiments on Natural Mujoco control task Walker Walk. We experiment with the seeds $(0, 1, 2)$. (c) Experiments on Natural Mujoco control task Walker Walk over the same seeds. All of the experiments share the same parameters except for β

4.3. Analysis of β

Since the conflict sample division between SSCL and DSCL, we introduce the hyperparameter β to balance the two different contrastive losses. To analyze the joint optimization of SSCL and DSCL affected by β, we conduct an ablation experiment to verify the effect of β on performance. We test β from the set $\{0.0, 0.5, 1.5, 3.0\}$ on Walker Walk environment with the seeds $(0, 1, 2)$ and found that $\beta = 1.5$ works well. So we choose $\beta = 1.5$ for all experiments and the performance is better than baselines, which means we don't need to tune β much for different tasks.

5. Conclusion

In the paper, we proposed Constrained Contrastive Reinforcement Learning (C2RL), a model-based reinforcement learning method for image-input tasks. The key contribution is that we optimize the world model through a combination of two contrastive losses based on latent dynamics and task-relevant state abstraction respectively, utilizing reward information to accelerate model learning. We also introduce a hyperparameter to balance the two optimization targets. Experiments on both the natural-video and standard background setting DMControl tasks demonstrate that our algorithm achieves superior performance compared with other state-of-the-art methods.

However, since the task-relevant state abstraction we use is based on Z-learning (Liu et al., 2021), which is the coarse method of (Givan et al., 2003), the performance of our method is sensitive to the hyperparameter β. Another disadvantage is that our method is sensitive to the tasks with sparse rewards which can be improved by using other state abstraction methods such as bisimulation metric (Givan et al., 2003).

Acknowledgments

This work was supported in part by National Key RD Program of China under grant No. 2021ZD0112700,NSFC under grant No.62125305, No.62088102, and No.61973246,the Fundamental Research Funds for the Central Universities under Grant xtr072022001.

References

Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. *Advances in Neural Information Processing Systems*, 32, 2019.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10069–10076, 2020.

Jianyu Chen, Shengbo Eben Li, and Masayoshi Tomizuka. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Debidatta Dwibedi, Jonathan Tompson, Corey Lynch, and Pierre Sermanet. Learning actionable representations from visual observations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1577–1584. IEEE, 2018.

Dibya Ghosh, Abhishek Gupta, and Sergey Levine. Learning actionable representations with goal-conditioned policies. *arXiv preprint arXiv:1811.07819*, 2018.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A Pires, and Rémi Munos. Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*, 2018.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019b.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems*, 33:741–752, 2020.

Guoqing Liu, Chuheng Zhang, Li Zhao, Tao Qin, Jinhua Zhu, Jian Li, Nenghai Yu, and Tie-Yan Liu. Return-based contrastive representation learning for reinforcement learning. *arXiv preprint arXiv:2102.10960*, 2021.

Xiao Ma, Siwei Chen, David Hsu, and Wee Sun Lee. Contrastive variational reinforcement learning for complex observations. *arXiv preprint arXiv:2008.02430*, 2020.

Bogdan Mazoure, Remi Tachet des Combes, Thang Long Doan, Philip Bachman, and R Devon Hjelm. Deep reinforcement and infomax learning. *Advances in Neural Information Processing Systems*, 33:3686–3698, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding for model-based planning in latent space. In *International Conference on Machine Learning*, pages 8130–8139. PMLR, 2021.

Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4209–4215. IEEE, 2021.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Sebastian Thrun and Michael L Littman. Reinforcement learning: an introduction. *AI Magazine*, 21(1):103–103, 2000.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2, 2019.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.