CCLF: A Contrastive-Curiosity-Driven Learning Framework for Sample-Efficient Reinforcement Learning

Chenyu Sun[1,2,3], **Hangwei Qian**[2,4*] and **Chunyan Miao**[1,2,4*]
[1]Alibaba-NTU Singapore Joint Research Institute
[2]School of Computer Science and Engineering, Nanyang Technological University
[3]Alibaba Group
[4]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY)
{chenyu002, qian0045}@e.ntu.edu.sg, ascymiao@ntu.edu.sg

Abstract

In reinforcement learning (RL), it is challenging to learn directly from high-dimensional observations, where data augmentation has recently remedied it via encoding invariances from raw pixels. Nevertheless, we empirically find that not all samples are equally important and hence simply injecting more augmented inputs may instead cause instability in Q-learning. In this paper, we approach this problem systematically by developing a model-agnostic **C**ontrastive-**C**uriosity-driven **L**earning **F**ramework (CCLF), which can fully exploit sample importance and improve learning efficiency in a self-supervised manner. Facilitated by the proposed contrastive curiosity, CCLF is capable of prioritizing the experience replay, selecting the most informative augmented inputs, and more importantly regularizing the Q-function as well as the encoder to concentrate more on under-learned data. Moreover, it encourages the agent to explore with a curiosity-based reward. As a result, the agent can focus on more informative samples and learn representation invariances more efficiently, with significantly reduced augmented inputs. We apply CCLF to several base RL algorithms and evaluate on the DeepMind Control Suite, Atari, and MiniGrid benchmarks, where our approach demonstrates superior sample efficiency and learning performances compared with other state-of-the-art methods. Our code is available at https://github.com/csun001/CCLF.

1 Introduction

Despite the success of reinforcement learning (RL), extensive data collection and environment interactions are still required to train the agents [Laskin *et al.*, 2020b]. In contrast, human beings are capable of learning new skills quickly and generalizing well with limited practice. Therefore, bridging the gap of sample efficiency and learning capabilities between machine and human learning has become a main challenge in the RL community [Rakelly *et al.*, 2019; Schwarzer *et al.*, 2020; Yarats *et al.*, 2021c; Malik *et al.*, 2021; Sun *et al.*, 2022].

This challenge is particularly vital in learning directly from raw pixels. More recently, data augmentation methods are leveraged to incorporate more invariances, promote data diversity, and thereby enhance representation learning [Laskin *et al.*, 2020a; Laskin *et al.*, 2020b; Yarats *et al.*, 2021b]. Ideally, injecting a larger number of augmented samples should lead to a better model with invariances. Nevertheless, a noticeable trade-off is the computational complexity introduced. What's worse, simply increasing the number of augmented inputs may alter the semantics of samples, which has been empirically shown in our experimental results. Moreover, the samples used for data augmentation are uniformly drawn from the replay buffer which is inefficient as they are not equally important to learn. These assumptions deviate from human-like intelligence, where humans can learn efficiently by curiously focusing on novel knowledge and revisiting old knowledge less frequently. Therefore, replaying the most under-explored experiences and selecting the most informative augmented inputs are the keys to improving sample efficiency and learning capability.

To tackle these challenges, we propose a **C**ontrastive-**C**uriosity-driven **L**earning **F**ramework (CCLF) by introducing contrastive curiosity into four important components of RL including experience replay, training input selection, learning regularization, and task exploration without much computational overhead. Inspired by the psychological curiosity that can be externally stimulated, encompassing complexity, novelty, and surprise [Berlyne, 1960; Spielberger and Starr, 2012; Liquin and Lombrozo, 2020], we define the contrastive curiosity based on the surprise conceptualized by the agent's internal belief towards the augmented inputs. The internal belief is modeled by reusing the contrastive loss term in CURL [Laskin *et al.*, 2020a], which can quantitatively measure the curiosity level without introducing any additional network architecture. With the proposed contrastive curiosity, agents can sample more under-explored transitions from the replay buffer, and select the most informative augmented inputs to encode invariances. This process can significantly reduce the amount of data used in RL without sacrificing the invariances. Thereafter, CCLF further utilizes the contrastive curiosity to regularize both Q-function and encoder by concentrating more on the surprising inputs, and intrinsically rewards agents for exploring under-learned observations.

*Co-corresponding authors

Our contribution can be highlighted as follows. 1) We empirically demonstrate that not all samples nor their augmentations are equally important in RL. Thus, agents should learn curiously from the most important ones in a self-supervised manner. 2) A surprise-aroused type of curiosity, namely, contrastive curiosity, is proposed by reusing the representation learning module without increasing the network complexity. 3) The proposed CCLF is capable of improving the sample efficiency and adapting the learning process directly from raw pixels, where the contrastive curiosity is fully exploited in different RL components in a self-navigated and coherent way. 4) CCLF is model-agnostic and can be applied to model-free off-policy and on-policy RL algorithms. 5) Compared to other approaches, CCLF obtains state-of-the-art performance on the DeepMind Control (DMC) suite [Tunyasuvunakool *et al.*, 2020], Atari Games [Bellemare *et al.*, 2013], and MiniGrid [Chevalier-Boisvert *et al.*, 2018] benchmarks.

2 Related Works

Data Augmentation in Sample-Efficient RL. Data augmentation has been widely applied in computer vision but is only recently introduced in RL to incorporate invariances for representation learning [Laskin *et al.*, 2020b; Yarats *et al.*, 2021b; Laskin *et al.*, 2020a]. To further improve the sample efficiency, one approach is to automatically apply the most effective augmentation method on any given task, through a multi-armed bandit or meta-learning the hyper-parameters to adapt [Raileanu *et al.*, 2020]. However, the underlying RL algorithm can become non-stationary and it costs more time to converge. Another approach is to regularize the learning process with observations from different training environments [Wang *et al.*, 2020] or different steps [Yarats *et al.*, 2021a]. By injecting greater perturbations from other tasks and steps, the encoded features and learned policies can become more robust to task invariances. Different from these works, the proposed CCLF primarily focuses on perturbations generated in a single task and step, and selects the most under-learned transition tuples and their augmented inputs. As not all samples nor their augmented inputs are equally important, our work exploits the sample importance to adapt the learning process by concentrating more on under-explored samples. Most importantly, the amount of augmentations can be greatly reduced and the sample efficiency is improved without introducing complicated architectures.

Curiosity-Driven RL. In curiosity-driven RL, agents are intrinsically motivated to explore the environment and perform complex control tasks by incorporating curiosity [Aubret *et al.*, 2019; Sun *et al.*, 2022]. In particular, curiosity is mainly used as a sophisticated intrinsic reward based on state novelty [Bellemare *et al.*, 2016], state prediction errors [Pathak *et al.*, 2017], and uncertainty about outcomes [Li *et al.*, 2021] or environment dynamics [Seo *et al.*, 2021]. Meanwhile, it can also be employed to prioritize the experience replay towards under-explored states [Schaul *et al.*, 2015; Zhao and Tresp, 2019]. However, additional networks are required to model curiosity, which can be computationally inefficient and unstable for high-dimensional inputs with continuous controls. Moreover, none of these works have yet attempted to improve the sample efficiency and resolve the instability caused by data augmentation. CCFDM [Nguyen *et al.*, 2021] is a concurrent work that incorporates CURL with action embedding and forward dynamics to formulate an intrinsic reward. Different from CCFDM, our framework does not require any additional architecture but only reuses the contrastive term in CURL that predicts more stably. More importantly, the proposed CCLF seamlessly integrates the curiosity mechanism into experience replay, training input selection, learning regularization, and environment exploration to concentrate more on under-learned samples and improve the sample efficiency stably.

3 Background

In this paper, we consider a Markov Decision Process (MDP) setting with the state $s_t \in \mathcal{S}$, the action $a_t \in \mathcal{A}$, the transition probability P mapping from the current state s_t and action a_t to the next state s'_t, and the (extrinsic) reward $r^e_t \in \mathcal{R}$. More details are provided in Appendix A [1].

Soft Actor-Critic (SAC). SAC [Haarnoja *et al.*, 2018] is an off-policy model-free algorithm that learns a stochastic policy π_ψ (actor) with state-action value functions Q_{ϕ_1}, Q_{ϕ_2} (critics), and a temperature α by encouraging exploration through a γ-discounted maximum-policy-entropy term. However, agents are often required to learn directly from high-dimensional observations $o_t \in \mathcal{O}$ rather than states s_t in practice. In this paper, we demonstrate our framework mainly using SAC with raw pixel inputs as the base algorithm.

Contrastive Unsupervised RL (CURL). CURL [Laskin *et al.*, 2020a] utilizes data augmentation and contrastive learning to train an image encoder $f_\theta(o)$ in a self-supervised way, imposing an instance discrimination between similar ($^+$) and dissimilar ($^-$) encoded states. Given a batch of visual observations o, each is augmented twice and encoded into a query $q = f_\theta(o_q)$ and a key $k = f_{\bar{\theta}}(o_k)$. The key encoder $f_{\bar{\theta}}$ is a momentum-based moving average of the query encoder f_θ to ensure consistency and stability, and f_θ is learned by enforcing q to match with k^+ while keeping far apart from k^-.

Data Regularized Q-Learning (DrQ). Based on SAC settings, DrQ [Yarats *et al.*, 2021b] incorporates optimality invariant image transformations to regularize the Q-function, improving robust learning directly from raw pixels. Let $g(o)$ represent the random image crop augmentation on observations o. It should ideally preserve the Q-values s.t. $Q(o,a) = Q(g_i(o),a), \forall o \in \mathcal{O}, a \in \mathcal{A}, i = 1,2,3,\cdots$. DrQ then applies data augmentation to each transition tuple $\tau_t = (o_t, a_t, r^e_t, d_t, o'_t)$ that is uniformly sampled from the replay buffer $\mathcal{B}$, where d_t is the done signal. With K augmented next observations $g_k(o'_t)$ and M augmented current observations $g_m(o_t)$, the critic Q_ϕ can be regularized by averaging over M augmented inputs from o_t,

$$\mathcal{L}_Q(\phi) = \mathbb{E}_{\tau\sim\mathcal{B}}\left[\frac{1}{M}\sum_{m=1}^{M}\left(Q_\phi(g_m(o_t),a_t) - (r^e_t + \gamma(1-d_t)\mathcal{T}_t)\right)^2\right] \quad (1)$$

where $\mathcal{T}_t$ is the soft target value and it can also be regularized by averaging over K augmented inputs from o'_t,

$$\mathcal{T}_t = \frac{1}{K}\sum_{k=1}^{K}\left[\min_{i=1,2} Q_{\bar{\phi}_i}(g_k(o'_t), a'_k) - \alpha \log \pi_\psi(a'_k|g_k(o'_t))\right]. \quad (2)$$

[1] Appendix is available at http://arxiv.org/abs/2205.00943.

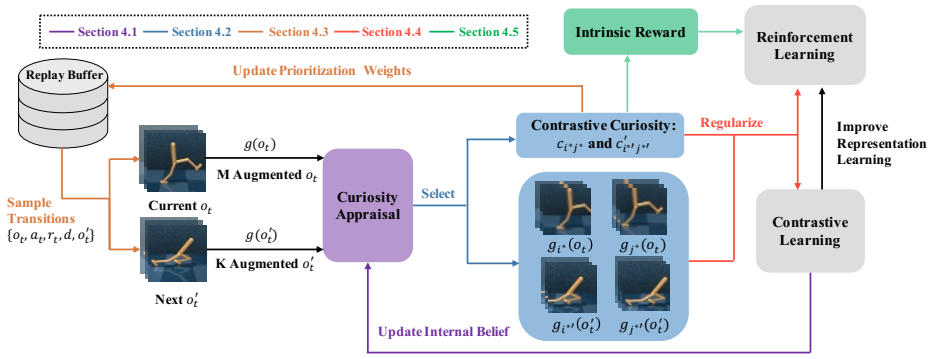

Figure 1: **C**ontrastive-**C**uriosity-driven **L**earning **F**ramework (CCLF): a batch of transitions are sampled w.r.t. their prioritization weights. Image augmentation is performed to obtain M augmented o_t and K augmented o'_t. The curiosity appraisal module quantitatively evaluates the contrastive curiosity and selects the two most informative inputs for both current and next observations. More importantly, the contrastive curiosity is simultaneously utilized to update the prioritization weights, construct an intrinsic reward, and adaptively regularize the contrastive learning and Q-learning modules. The contrastive learning module improves the representation learning and updates the agent's internal belief.

4 The Proposed CCLF

Contrastive-**C**uriosity-driven **L**earning **F**ramework (CCLF) extends the model-free RL to further improve sample efficiency when learning directly from the raw pixels. In particular, it fully exploits sample importance for agents to efficiently learn from the most informative data. Firstly, we repurpose the contrastive term in CURL [Laskin *et al.*, 2020a] without additional architectures to quantify the contrastive curiosity (Section 4.1). Subsequently, this contrastive curiosity is coherently integrated into four components to navigate RL with minimum modification: augmented input selection (Section 4.2), experience replay (Section 4.3), Q-function and encoder regularization (Section 4.4), and environment exploration (Section 4.5), as illustrated in Figure 1. Without loss of generosity, we apply CCLF on the state-of-the-art off-policy RL algorithm, SAC [Haarnoja *et al.*, 2018] as summarized in Algorithm 1. Extensions on other base algorithms are carried out in Section 5.2 and B.2 and B.3.

4.1 Contrastive Curiosity

Curiosity can be aroused by an unexpected stimulus that behaves differently from the agent's internal belief. To quantify such surprise-aroused curiosity, we define the agent's curiosity c_{ij} by the prediction error of whether any two augmented observations $g_i(o), g_j(o)$ are from the same observation o,

$$c_{ij} = 1 - \text{IB}(g_i(o), g_j(o)) \in [0, 1] \tag{3}$$

where IB represents agent's internal belief of whether $g_i(o), g_j(o)$ are augmented (*e.g.*, randomly cropped) from the same o with similar representations. Since the contrastive loss can be viewed as the log-loss of a softmax-based classifier to match a query q with the key k from the same observation in a batch, it becomes a natural choice for measuring agent's internal belief $\text{IB}(g_i(o), g_j(o)) = \frac{\exp(q^T W k^+)}{\exp(q^T W k^+) + \sum_{l=1}^{B-1} \exp(q^T W k_l^-)}$, where B is the batch size and q is the query encoder $q = f_\theta(g_i(o))$. Moreover, we denote the key encoder $k = f_{\bar{\theta}}(g_j(o))$ as k^+ if its input is the same as that in the query encoder q; otherwise, we denote it as k^-. An immediate merit is that the contrastive curiosity does not require any additional architecture or auxiliary loss because IB is updated directly through representation learning in a self-supervised way.

A higher contrastive curiosity value indicates that the agent does not believe q is similar to k^+ or the agent mistakenly matches q with some k^- instead, which ultimately results in a surprise in a self-supervised way. It further implies that the sampled transition tuple contains novel information that has yet been learned by the agent, and the encoder f_θ is not optimal to extract a meaningful state representation from raw pixels. With the proposed contrastive curiosity in-place, we can integrate different curiosity-driven mechanisms in the proposed CCLF to achieve sample-efficient RL, which is discussed in the following sub-sections.

4.2 Curiosity-Based Augmented Input Selection

Although DrQ [Yarats *et al.*, 2021b] has shown that increasing $[K, M]$, *i.e.*, the amounts of augmented inputs on next and current observations respectively, can potentially improve agent's performances through regularized Q-learning, a crucial trade-off is the introduced higher computational complexity. In addition, more augmented data does not necessarily lead to better performance, as data transformations might alter the semantics and result in the counterproductive performance. To tackle these challenges, we aim to select the most informative inputs for the subsequent learning. Without loss of generality, we assume two inputs are selected from M augmented current observations o_t and similarly two are selected from K augmented next observations o'_t in this paper.

It should be noted that there are various ways to select the most informative augmented inputs, where one straightforward way is to select by least overlap in pixels,

$$i^*, j^* = \arg\min \text{Overlap}(g_i(o), g_j(o)) \; \forall i, j, i \neq j. \tag{4}$$

However, a more human-like way is to select the most representative inputs based on the curious level conceptualized by the internal belief rather than simple visual overlaps. Therefore, we propose to select the augmented inputs that cause highest contrastive curiosity as defined in Eq. (3),

$$i^*, j^* = \arg\max c_{ij} \; \forall i, j, i \neq j. \tag{5}$$

In this way, the augmented inputs that are most challenging for matching can be curiously identified since they potentially

contain novel knowledge that has yet been learned; meanwhile, this selection mechanism can help to encode more representative state information from the selected inputs, while the agent's internal belief can be jointly updated. As a result, an improved encoder that is robust to different views of observations can be trained with fewer inputs, potentially yielding an improvement in the sample efficiency.

4.3 Curiosity-Based Experience Replaying

In the conventional off-policy RL, agents uniformly sample transitions τ from the replay buffer to learn policies. Although they can eventually perform a complex task by repeatedly practicing in a trial-and-error fashion, we hypothesize that a more sample-efficient and generalizable way is to revisit the transitions that are relatively new or different more frequently. Therefore, we prioritize the experience replay by assigning different prioritization weights $w \in [0, 1]$ to all transitions stored in the replay buffer $\mathcal{B}$. In particular, the prioritization weight is initialized to $w_0 = 1$ for any newly added transition tuple. Thereafter, we propose to update the weights of transitions with the overall contrastive curiosity at each training step s,

$$w_s = \beta w_{s-1} + \frac{1}{2}(1-\beta)(c_{i^*j^*} + c'_{i^{*\prime}j^{*\prime}}) \tag{6}$$

where $\beta \in [0, 1]$ is a momentum coefficient, and $c_{i^*j^*}, c'_{i^{*\prime}j^{*\prime}}$ are the contrastive curiosity about o and o' respectively. The intuition of the momentum update is to maintain a stable update such that the transitions arousing low curiosity will be gradually de-prioritized for learning. Mathematically, the probability of τ_i to be replayed is $p(\tau_i) = \frac{w_i}{\sum_{n=1}^{N} w_n}$ and it becomes small only when τ_i has been sampled many times. Hence, more recent and surprising transitions arousing high curiosity can be sampled more frequently to learn.

4.4 Curiosity-Based Regularization

Although agents can benefit from learning the selected complex inputs, it imposes challenges for agents as well, which may cause unstable and poor performances. Hence, it is crucial to adapt the learning process by concentrating more on under-learned knowledge. To achieve this, we propose an adaptive regularization for both Q-function and the observation encoder, guided by the contrastive curiosity in order to learn more from the selected inputs arousing high curiosity. In particular, we modify Eq. (2) and Eq. (1) as

$$\begin{aligned}
&\mathcal{L}_Q(\phi) = \mathbb{E}_{\tau\sim\mathcal{B}}\left[(1 - c_{i^*j^*})\mathcal{E}^2_{i^*} + c_{i^*j^*}\mathcal{E}^2_{j^*}\right], \\
&\mathcal{T}_t = (1 - c'_{i^{*\prime}j^{*\prime}})\mathcal{T}_t^{i^{*\prime}} + c'_{i^{*\prime}j^{*\prime}}\mathcal{T}_t^{j^{*\prime}}, \\
&\text{where } \mathcal{E}_m = Q_\phi(g_m(o_t), a_t) - (r^e_t + \gamma(1-d_t)\mathcal{T}_t),\ m = i^*, j^*, \\
&\text{and } \mathcal{T}_t^k = \min_{l=1,2} Q_{\bar{\phi}_l}(g_k(o'_t), a'_k) - \alpha \log \pi_\psi(a'_k|g_k(o'_t)), k = i^{*\prime}, j^{*\prime}.
\end{aligned} \tag{7}$$

It is worth noting that this regularized Q-function is rather general to recover other state-of-the-art works as special cases. When all augmented inputs arouse exactly moderate level of curiosity $c_{i^*j^*} = c'_{i^{*\prime}j^{*\prime}} = \frac{1}{2}$, the proposed regularization is equivalent to DrQ with $[K, M] = [2, 2]$. Moreover, when the agent can perfectly match the two augmented inputs with no contrastive curiosity $c_{i^*j^*} = c'_{i^{*\prime}j^{*\prime}} = 0$, it is sufficient to update the Q-function with only one input; when the agent fails to encode any similarity and becomes extremely curious $c_{i^*j^*} = c'_{i^{*\prime}j^{*\prime}} = 1$, it should focus completely on the novel input instead. Both cases can reproduce the work of RAD [Laskin *et al.*, 2020b]. Most importantly, our proposed Q-function regularization enables the agent to adapt the learning process in a self-supervised way such that it is fully controlled by the conceptualized contrastive curiosity to exploit sample importance and stabilize the learning process.

Algorithm 1 An Implementation of CCLF on SAC

Input: MDP $\tau_t = (o_t, a_t, r^e_t, d_t, o'_t)$, numbers of augmented inputs $[K, M]$, replay buffer $\mathcal{B}$, training step T, batch size B
Parameter: Observation encoder network θ, actor network ψ, critics networks ϕ_i, temperature coefficient α, and bilinear product weight W
Output: Optimal policy π^*_ψ

for $t = 1$ **to** T **do**
  $a_t \sim \pi_\psi(\cdot|g(o_t))$
  $\mathcal{B} \cup (o_t, a_t, r^e_t, d_t, o'_t) \to \mathcal{B}$ with $w_t = 1$
  Sample a minibatch $\{(o_l, a_l, r^e_l, d_l, o'_l)\}_{l=1}^{B} \overset{w_l}{\sim} \mathcal{B}$ based on the prioritization weight w_l
  for each sample τ_l in the minibatch **do**
    Augment o_l and o'_l via $g(\cdot)$ to obtain M and K inputs
    Evaluate the contrastive curiosity c_{ij} by Eq. (3)
    Select $g_{i^*}(o_l), g_{j^*}(o_l)$ from M augmented o_l and select $g_{i^{*\prime}}(o'_l), g_{j^{*\prime}}(o'_l)$ from K augmented o'_l by Eq. (5)
    $r_l = r^e_l + r^i_l$ with r^i_l from Eq. (9)
    Update w_l according to Eq. (6)
  end for
  Update critics Q_{ϕ_i} by Eq. (7)
  Update the actor π_ψ and temperature coefficient α
  Update encoder f_θ and W by Eq. (8)
  $o_{t+1} = o'_t$
end for

Similarly, we also regularize the representation learning in a curious manner, inspired by the solution to the supervised class imbalance problem. To deal the with training set containing under-represented classes, a practical approach is to inversely weight the loss of each class according to their sizes. We follow this motivation to incorporate the contrastive curiosity $c^b_{i^*j^*}$ about the current observations as the weight for each log-loss class b to update the encoder f_θ,

$$\mathcal{L}_f(\theta) = -\sum_{b=1}^{B} c^b_{i^*j^*} \log \frac{\exp(q_b^T W k^+)}{\exp(q_b^T W k^+) + \sum_{l=1}^{B-1} \exp(q_b^T W k_l^-)} \tag{8}$$

where samples arousing high contrastive curiosity will be considered as under-represented classes and therefore agents need to adaptively pay more attention during the representation learning by optimizing f_θ. Meanwhile, agents also jointly re-calibrate a proper internal belief by updating W.

4.5 Curiosity-Based Exploration

Intrinsic rewards can motivate agents to explore actively [Sun *et al.*, 2022], improving the sample efficiency in the conventional RL. While SAC alone can be viewed as the entropy

100K Step Scores	SAC-Pixel	CURL	DrQ	CURL+	CURL++	Select	Select+	CCLF
Finger, Spin	230±194	686±113	784±173	780±96	735±120	699±138	768±90	**944±42**
Cartpole, Swingup	237±49	524±179	675±174	694±87	665±122	624±182	561±181	**799±61**
Reacher, Easy	239±183	566±226	682±86	541±190	479±216	646±171	616±284	**738±99**
Cheetah, Run	118±13	286±65	**332±36**	302±50	264±53	251±26	265±69	317±38
Walker Walk	95±19	482±237	492±267	484±61	504±142	453±91	408±170	**648±110**
Ball in Cup, Catch	85±130	667±197	828±131	687±260	728±143	732±223	739±132	**914±20**
500K Step Scores	SAC-Pixel	CURL	DrQ	CURL+	CURL++	Select	Select+	CCLF
Finger, Spin	346±95	783±192	803±198	855±164	838±164	803±167	879±153	**974±6**
Cartpole, Swingup	330±73	847±28	858±19	853±22	852±17	855±26	837±38	**869±9**
Reacher, Easy	307±65	**956±40**	939±44	933±62	937±40	939±78	906±80	941±48
Cheetah, Run	85±51	440±144	536±115	518±24	495±97	417±59	470±78	**588±22**
Walker Walk	71±52	928±26	887±126	916±27	914±24	921±27	850±64	**936±23**
Ball in Cup, Catch	162±122	**956±14**	**956±14**	951±19	**956±8**	949±21	949±24	**961±9**

Table 1: Performance scores (mean & standard deviation) on DMC evaluated at 100K and 500K environment steps. CCLF outperforms other approaches on 5 out of 6 tasks in both sample efficiency (100K) and asymptotic performance (500K) regimes, across 6 random seeds.

maximization of agent's actions intrinsically, in the proposed CCLF, we explicitly define an intrinsic reward proportional to the average contrastive curiosity about o_t and o'_t,

$$r_t^i = \lambda \exp(-\eta t) \frac{r^e_{max}}{r^i_{max}} \frac{c_{i^*j^*} + c'_{i^{*\prime}j^{*\prime}}}{2} \tag{9}$$

where λ is a temperature coefficient, η is a decay weight, t is the environment step, r^e_{max} and r^i_{max} are respectively the maximum extrinsic and intrinsic rewards over step t.

With the proposed r_t^i to supplement r_t^e in Eq. (7), agents can be encouraged to explore the surprising states that arouse high contrastive curiosity substantially. In particular, higher r_t^i rewards agents for exploration when different views of the same observations produce inconsistent representations. Meanwhile, r_t^i is decayed with respect to the environment step t to ensure the convergence of policies. As the extrinsic reward r^e differs across different tasks, the normalization is performed to balance r^e and r^i. This formulation is similar to the intrinsic reward in CCFDM [Nguyen *et al.*, 2021], but the proposed CCLF does not require a forward dynamic model or action embedding that increases the model complexity.

5 Experiments and Results

5.1 Experimental Setup

We empirically evaluate the proposed CCLF in terms of sample efficiency and ultimate performance, on 6 continuous control tasks from the DMC suite [Tunyasuvunakool *et al.*, 2020], 26 discrete control tasks from the Atari Games [Bellemare *et al.*, 2013] and 3 navigation tasks with sparse extrinsic rewards from the MiniGrid [Chevalier-Boisvert *et al.*, 2018]. In this section, we mainly present the experimental results in the DMC suite with SAC being the base algorithm while detailed settings and results in the Atari Games and the MiniGrid are included in Appendix B.2, B.3, C.2, and C.3. For a comprehensive evaluation in the DMC suite, we include the following baselines to compare against:

- Pixel-based SAC (SAC-Pixel) [Haarnoja *et al.*, 2018]
- CURL [Laskin *et al.*, 2020a].
- DrQ [Yarats *et al.*, 2021b] with $[K, M] = [2, 2]$ and a modified augmentation method for consistency.
- Hybrids of CURL and DrQ: CURL+ and CURL++, where contrastive representation learning is integrated to DrQ for $[K, M] = [2, 2]$ and $[5, 5]$ respectively.
- Augmented input selection models: 2 out of 5 inputs for each sample are selected by pixel overlap (Select) via Eq. (4) and contrastive curiosity (Select+) via Eq. (5) without the other curiosity-based components.

The detailed setting of hyper-parameters is provided in Appendix B.1. For our proposed CCLF, we initialize it with $[K, M] = [5, 5]$ to generate a sufficiently large amount of augmented inputs. For simplicity, we fix i^* randomly and only select j^* via Eq. (5) for the augmented input selection.

5.2 Results and Discussion

Not all Samples are Equally Important. In CURL+, data augmentation is applied twice for each sampled transition while 5 times in CURL++. Since CURL++ injects 2.5× larger amount of inputs than CURL+, its computational complexity increases dramatically. Table 1 shows that CURL++ performs worse than CURL+ in 4 tasks at 100K steps and slightly outperforms CURL+ in only 2 tasks at 500K steps. In Figure 2 and Appendix Figure 5, the learning curve of CURL++ is clearly below CURL+ at first and gradually approaches to the same level as CURL+. Since more augmented inputs may not guarantee the consistency of semantics, additional training is often required for convergence. Therefore, we can empirically validate the hypothesis that not all augmented inputs are equally important and simply increasing the number of augmentations is instead inefficient. A similar result can be found in DrQ Appendix F [Yarats *et al.*, 2021b].

Main Results on the DMC Suite

The average sample efficiency and asymptotic performance are shown in Table 1 at 100K and 500K environment steps, respectively. Meanwhile, Figure 2 demonstrates the agent's learning capability over 500K steps. Compared SAC-Pixel to the other models in Figure 2, its performance is not improved in all 6 tasks even until 500K steps while the other models can asymptotically perform well. Thus, it is challenging for the conventional SAC to learn directly from raw pixels and a sample-efficient RL method is needed to aid that.

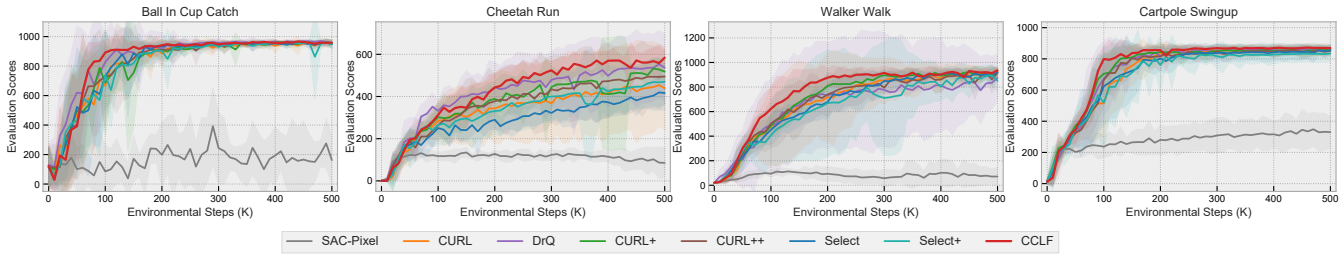

Figure 2: Learning performances on the continuous control tasks from the DMC Suite (Selected). The proposed CCLF on SAC outperforms the other baseline methods in terms of sample efficiency and converges much faster, averaged by 6 random runs.

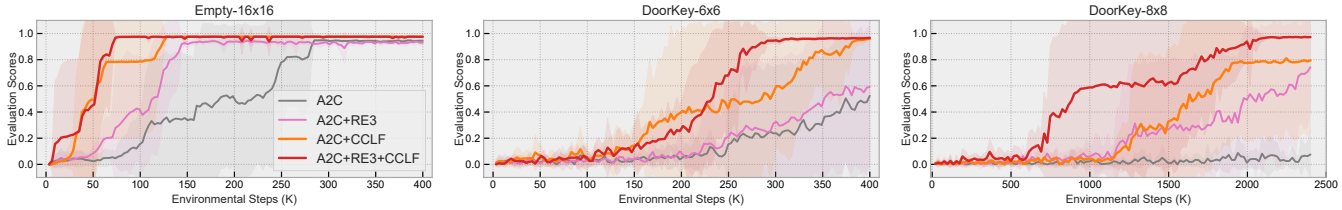

Figure 3: Learning performances on the navigation tasks from the MiniGrid. The proposed CCLF can be applied to both A2C and A2C+RE3. It significantly outperforms the other baselines in terms of sample efficiency and converges much faster, averaged by 5 random runs.

According to Table 1, Select performs better than Select+, on 3 tasks at 100K and 4 tasks at 500K, with more stable learning curves as shown in Figure 2. Indeed, the inputs in Select may contain some invariances to improve sample efficiency and learning capability. However, more under-learned inputs with even richer invariances are present in Select+, and agents cannot adapt the learning process in this model, causing the instability issue in Select+.

To tackle this issue, the proposed CCLF collaboratively adapts the learning process with the selected inputs and contrastive curiosity, so the learning curves become more smooth than others in Figure 2. In particular, CCLF outperforms all baselines in 5 tasks at both 100K and 500K regimes in Table 1. Moreover, it converges much faster than Select+ according to the results at 100K steps. In fact, the proposed CCLF only requires about 50% environment steps to converge to desirable performances on 3 tasks (Ball in Cup, Walker, and Cartpole) as the other baselines. In addition, it even benchmarks on Cheetah-Run and Finger-Spin tasks at 500K steps. Therefore, we can conclude that our proposed CCLF can improve the sample efficiency and learning capabilities of RL agents, with fewer environment interactions and 60% reduced augmented inputs. We also analyze the computational complexity on the Cartpole task by model sizes and training time in Appendix C.1, where CCLF can avoid increasing the training cost dramatically.

Additional Experiments in Atari Games. In addition to continuous control tasks, CCLF can also be incorporated into Rainbow DQN [Hessel *et al.*, 2018] to perform discrete control tasks. As shown in Appendix C.2, the proposed CCLF attains state-of-the-art performances in 8 out of 26 Atari Games at 100K steps. In particular, CCLF is superior to CURL in 11 games and DrQ in 18 games, which favorably indicates the effectiveness of improving sample efficiency.

Further Investigation on MiniGrid. Apart from off-policy algorithms, we also investigate the compatibility on the on-policy algorithm. More specifically, we extend the proposed CCLF to A2C [Mnih *et al.*, 2016] and RE3 [Seo *et al.*, 2021] to perform navigation tasks with spares rewards in MiniGrid. We first adapt CCLF to the on-policy algorithm by removing the experience replay component. Note that the input from MiniGrid is already a compact and efficient $7 \times 7 \times 3$ embedding of partially-observable 7×7 grids, so even slight augmentation will induce highly inconsistent learned features. Thus, we directly duplicate the embedding without random augmentations to obtain contrastive curiosity for regularization and intrinsic reward. Figure 3 shows that CCLF exhibits superior sample efficiency and learning capabilities in all three tasks, even model-agnostic with the state-of-the-art curiosity-driven method RE3. In the Empty-16×16 task, our CCLF can reach the optimal level in about 50% and 55% of the training steps of RE3 and A2C, respectively. By comparing the ultimate performance scores, the proposed CCLF obtains 1.63× higher average performance than RE3 in DoorKey-6×6 and 1.3× in DoorKey-8×8.

Effectiveness of the Proposed RL Components. One might wonder if the proposed CCLF benefits mainly from one or several curiosity-based components in practice. Hence, we empirically examine the effectiveness of all possible combinations of the four curiosity-driven components on the Cartpole task from the DMC suite. The results are included in Appendix C.4, where it can be concluded that all four components are necessary and important to attain state-of-the-art performances. Our proposed CCLF can navigate all four RL components together to improve the sample efficiency and resolve instability, which demonstrates effective collaboration.

6 Conclusion

In this paper, we present CCLF, a contrastive-curiosity-driven learning framework for RL with visual observations, which can significantly improve the sample efficiency and learning capabilities of agents. As we empirically find that not all samples nor their augmented inputs are equally important for RL, CCLF encourages agents to learn in a curious way, exploiting sample complexity and importance systemically.

Acknowledgments

This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is supported in part by the Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, and in part by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017) and (NIC Project No. MOH/NIC/HAIG03/2017). H.Qian thanks the support from the Wallenberg-NTU Presidential Postdoctoral Fellowship.

References

[Aubret *et al.*, 2019] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv:1908.06976*, 2019.

[Bellemare *et al.*, 2013] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *JAIR*, 47, 2013.

[Bellemare *et al.*, 2016] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *NeurIPS*, 29:1471–1479, 2016.

[Berlyne, 1960] Daniel E. Berlyne. *Conflict, arousal, and curiosity.* McGraw-Hill Book Company, 1960.

[Chevalier-Boisvert *et al.*, 2018] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018. Accessed: 2021-12-05.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[Hessel *et al.*, 2018] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI*, volume 32, 2018.

[Laskin *et al.*, 2020a] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *ICML*, pages 5639–5650. PMLR, 2020.

[Laskin *et al.*, 2020b] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *NeurIPS*, 33, 2020.

[Li *et al.*, 2021] Kevin Li, Abhishek Gupta, Ashwin Reddy, Vitchyr H Pong, Aurick Zhou, Justin Yu, and Sergey Levine. Mural: Meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning. In *ICML*, pages 6346–6356, 2021.

[Liquin and Lombrozo, 2020] Emily G Liquin and Tania Lombrozo. Explanation-seeking curiosity in childhood. *Current Opinion in Behavioral Sciences*, 35:14–20, 2020.

[Malik *et al.*, 2021] Dhruv Malik, Aldo Pacchiano, Vishwak Srinivasan, and Yuanzhi Li. Sample efficient reinforcement learning in continuous state spaces: A perspective beyond linearity. In *ICML*, pages 7412–7422. PMLR, 2021.

[Mnih *et al.*, 2016] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937. PMLR, 2016.

[Nguyen *et al.*, 2021] Thanh Nguyen, Tung M Luu, Thang Vu, and Chang D Yoo. Sample-efficient reinforcement learning representation learning with curiosity contrastive forward dynamics model. *arXiv preprint arXiv:2103.08255*, 2021.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, pages 2778–2787. PMLR, 2017.

[Raileanu *et al.*, 2020] Roberta Raileanu, Max Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv preprint arXiv:2006.12862*, 2020.

[Rakelly *et al.*, 2019] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *ICML*, pages 5331–5340. PMLR, 2019.

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Schwarzer *et al.*, 2020] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *ICLR*, 2020.

[Seo *et al.*, 2021] Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. *arXiv preprint arXiv:2102.09430*, 2021.

[Spielberger and Starr, 2012] Charles D Spielberger and Laura M Starr. Curiosity and exploratory behavior. In *Motivation: Theory and research*, pages 231–254. 2012.

[Sun *et al.*, 2022] Chenyu Sun, Hangwei Qian, and Chunyan Miao. From psychological curiosity to artificial curiosity: Curiosity-driven learning in artificial intelligence tasks. *arXiv preprint arXiv:2201.08300*, 2022.

[Tunyasuvunakool *et al.*, 2020] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

[Wang *et al.*, 2020] Kaixin Wang, Bingyi Kang, Jie Shao, and Jiashi Feng. Improving generalization in reinforcement learning with mixture regularization. *NeurIPS*, 33:7968–7978, 2020.

[Yarats *et al.*, 2021a] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

[Yarats *et al.*, 2021b] Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *ICLR*, 2021.

[Yarats *et al.*, 2021c] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *AAAI*, 2021.

[Zhao and Tresp, 2019] Rui Zhao and Volker Tresp. Curiosity-driven experience prioritization via density estimation. *arXiv preprint arXiv:1902.08039*, 2019.