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

Chenyu Sun¹,²,³, Hangwei Qian²,⁴*, Chunyan Miao¹,²,⁴*  
¹Alibaba-NTU Singapore Joint Research Institute  
²School of Computer Science and Engineering, Nanyang Technological University  
³Alibaba Group  
⁴Joint 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 Contrastive-Curiosity-driven Learning Framework (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 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 Contrastive-Curiosity-driven Learning Framework (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.

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 [Schaal 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 $s_t \in S$, the action $a_t \in A$, 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 $r_t^e \in R$. More details are provided in Appendix A.

Soft Actor-Critic (SAC). SAC [Haarnoja et al., 2018] is an off-policy model-free algorithm that learns a stochastic policy $\pi_\phi$ (actor) with state-action value functions $Q_{\theta_1}, Q_{\theta_2}$ (critics), and a temperature $\alpha$ by encouraging exploration through a $\gamma$-discounted maximum-policy-entropy term. However, agents are often required to learn directly from high-dimensional observations $o_t \in O$ 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 ($\sim$) and dissimilar ($\neq$) encoded states. Given a batch of visual observations $o$, each is augmented twice and encoded into a query $q = f_\theta(o_a)$ and a key $k = f_\theta(o_k)$. The key encoder $f_\theta$ is a momentum-based moving average of the query encoder $f_\theta$ to ensure consistency and stability, and $f_\theta$ is learned by enforcing $q$ to match with $k^+ \equiv k$ while keeping far apart from $k^-$. 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 [Schaal 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 $s_t \in S$, the action $a_t \in A$, 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 $r_t^e \in R$. More details are provided in Appendix A.

Soft Actor-Critic (SAC). SAC [Haarnoja et al., 2018] is an off-policy model-free algorithm that learns a stochastic policy $\pi_\phi$ (actor) with state-action value functions $Q_{\theta_1}, Q_{\theta_2}$ (critics), and a temperature $\alpha$ by encouraging exploration through a $\gamma$-discounted maximum-policy-entropy term. However, agents are often required to learn directly from high-dimensional observations $o_t \in O$ 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 ($\sim$) and dissimilar ($\neq$) encoded states. Given a batch of visual observations $o$, each is augmented twice and encoded into a query $q = f_\theta(o_a)$ and a key $k = f_\theta(o_k)$. The key encoder $f_\theta$ is a momentum-based moving average of the query encoder $f_\theta$ to ensure consistency and stability, and $f_\theta$ is learned by enforcing $q$ to match with $k^+ \equiv k$ while keeping far apart from $k^-$. 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 [Schaal 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.
4 The Proposed CCLF

Contrastive-Curiosity-driven Learning Framework (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 re-purpose 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 $o$,

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

(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 $\text{IB}(g_i(o), g_j(o)) = \frac{\exp(q^T W k^+)}{\exp(q^T W k^+) + \sum_{l=1}^M \exp(q^T W k_l^+)}$, where $B$ is the batch size and $q$ is the query encoder $q = f_o(g_i(o))$. Moreover, we denote the key encoder $k = f_o(g_j(o))$ 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 $f_o$ 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 $o_i$ and similarly two are selected from $K$ augmented next observations $o'_j$ 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,

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

(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),

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

(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 \( \tau \) 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 \( B \). 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 \( s \),

\[
w_s = \beta w_{s-1} + \frac{1}{2} (1 - \beta) (c_t^{i_s, j_s} + c_t^{i_s', j_s'})
\]

where \( \beta \in [0, 1] \) is a momentum coefficient, and \( c_t^{i_s, j_s}, c_t^{i_s', j_s'} \) 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 \( \tau_i \) to be replayed is \( p(\tau_i) = \frac{w_i}{\sum_{n=1}^{N} w_n} \) and it becomes small only when \( \tau_i \) 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

\[
\mathcal{L}_Q(\phi) = \mathbb{E}_{\tau \sim B} \left[ (1 - c^{i_s, j_s}) \mathcal{E}^2_\tau + c^{i_s, j_s} \mathcal{E}^2_\tau \right],
\]

\[
\mathcal{T}_\tau = (1 - c^{i_s', j_s'}) \mathcal{T}_\tau^{i_s'} + c^{i_s', j_s'} \mathcal{T}_\tau^{i_s'},
\]

where \( \mathcal{E}_m = Q_\phi(g_m(o_i), o_{i-1}) - (r_i + \gamma (1 - d_i) T_i), m = i^*, j^* \), and

\[
\mathcal{T}_k = \min_{i=1,2} Q_{\phi}(g_k(o_i), a_k) - \alpha \log q_\psi(a_k|g_k(o_i)), k = i^*, j^*.
\]

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 \( c^{i_s, j_s} = c^{i_s', j_s'} = \frac{1}{2} \), 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 \( c^{i_s, j_s} = c^{i_s', j_s'} = 0 \), it is sufficient to update the Q-function with only one input; when the agent fails to encode any similarity and becomes extremely curious \( c^{i_s, j_s} = c^{i_s', j_s'} = 1 \), it should focus completely on the novel input instead. Both cases can reproduce the work of RAD [Laskin et al., 2020]. 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.

Similarly, we also regularize the representation learning in a curious manner, inspired by the solution to the supervised class imbalance problem. To deal with the 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 \( c^{i_s, j_s'} \) about the current observations as the weight for each log-loss class \( b \) to update the encoder \( f_\theta \),

\[
\mathcal{L}_f(\theta) = - \sum_{b=1}^{B} b^b \log \frac{\exp(q_{\phi}(Wk+b))}{\sum_{b=1}^{B} \exp(q_{\phi}(Wk+b))}
\]

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 \( f_\theta \). 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...
to the average contrastive curiosity about $CCLF$, we explicitly define an intrinsic reward proportional to 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 intrinsic reward $r^i$ 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.

Table 1: Performance scores (mean & standard deviation) on DM evaluation 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.

<table>
<thead>
<tr>
<th>100K Step Scores</th>
<th>SAC-Pixel</th>
<th>CURL</th>
<th>DrQ</th>
<th>CURL+</th>
<th>CURL++</th>
<th>Select</th>
<th>Select+</th>
<th>CCLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finger, Spin</td>
<td>230±194</td>
<td>686±113</td>
<td>784±173</td>
<td>780±96</td>
<td>735±120</td>
<td>699±138</td>
<td>768±90</td>
<td>944±42</td>
</tr>
<tr>
<td>Cartpole, Swingup</td>
<td>237±49</td>
<td>524±179</td>
<td>675±174</td>
<td>694±87</td>
<td>665±122</td>
<td>624±182</td>
<td>561±181</td>
<td>799±61</td>
</tr>
<tr>
<td>Reacher, Easy</td>
<td>239±183</td>
<td>556±226</td>
<td>682±86</td>
<td>541±190</td>
<td>479±216</td>
<td>646±171</td>
<td>616±284</td>
<td>738±99</td>
</tr>
<tr>
<td>Cheetah, Run</td>
<td>118±13</td>
<td>286±65</td>
<td>332±36</td>
<td>302±50</td>
<td>264±53</td>
<td>251±26</td>
<td>265±49</td>
<td>317±38</td>
</tr>
<tr>
<td>Walker Walk</td>
<td>95±19</td>
<td>482±237</td>
<td>492±267</td>
<td>484±61</td>
<td>504±142</td>
<td>453±91</td>
<td>408±170</td>
<td>648±110</td>
</tr>
<tr>
<td>Ball in Cup, Catch</td>
<td>85±130</td>
<td>667±197</td>
<td>828±131</td>
<td>687±260</td>
<td>728±143</td>
<td>732±223</td>
<td>739±132</td>
<td>914±20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>500K Step Scores</th>
<th>SAC-Pixel</th>
<th>CURL</th>
<th>DrQ</th>
<th>CURL+</th>
<th>CURL++</th>
<th>Select</th>
<th>Select+</th>
<th>CCLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finger, Spin</td>
<td>346±95</td>
<td>783±192</td>
<td>803±198</td>
<td>855±164</td>
<td>838±164</td>
<td>803±167</td>
<td>879±153</td>
<td>974±6</td>
</tr>
<tr>
<td>Cartpole, Swingup</td>
<td>330±73</td>
<td>847±28</td>
<td>858±19</td>
<td>853±22</td>
<td>852±17</td>
<td>855±26</td>
<td>837±38</td>
<td>869±9</td>
</tr>
<tr>
<td>Reacher, Easy</td>
<td>307±65</td>
<td>956±40</td>
<td>939±44</td>
<td>933±62</td>
<td>937±40</td>
<td>939±78</td>
<td>906±80</td>
<td>941±48</td>
</tr>
<tr>
<td>Cheetah, Run</td>
<td>85±51</td>
<td>440±14</td>
<td>536±115</td>
<td>518±24</td>
<td>495±97</td>
<td>417±59</td>
<td>470±78</td>
<td>588±22</td>
</tr>
<tr>
<td>Walker Walk</td>
<td>71±52</td>
<td>928±26</td>
<td>887±126</td>
<td>916±27</td>
<td>914±24</td>
<td>921±27</td>
<td>850±64</td>
<td>936±23</td>
</tr>
<tr>
<td>Ball in Cup, Catch</td>
<td>162±122</td>
<td>956±14</td>
<td>956±14</td>
<td>951±19</td>
<td>956±8</td>
<td>949±21</td>
<td>949±24</td>
<td>961±9</td>
</tr>
</tbody>
</table>

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++
- 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 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 $t^*$ 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 $2.5 \times$ 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.
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 sparse 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 × 7 × 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 (NRF 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 Wallenberg-NTU Presidential Postdoctoral Fellowship.

References


