Sample-efficient Reinforcement Learning Representation Learning with Curiosity Contrastive Forward Dynamics Model*

Thanh Nguyen†, Tung M. Luu†, Thang Vu and Chang D. Yoo

**Abstract**—Developing an agent in reinforcement learning (RL) that is capable of performing complex control tasks directly from high-dimensional observation such as raw pixels is a challenge as efforts still need to be made towards improving sample efficiency and generalization of RL algorithm. This paper considers a learning framework for a Curiosity Contrastive Forward Dynamics Model (CCFDM) to achieve a more sample-efficient RL based directly on raw pixels. CCFDM incorporates a forward dynamics model (FDM) and performs contrastive learning to train its deep convolutional neural network-based image encoder (IE) to extract conducive spatial and temporal information to achieve a more sample efficiency for RL. In addition, during training, CCFDM provides intrinsic rewards, produced based on FDM prediction error, and encourages the curiosity of the RL agent to improve exploration. The diverge and less-repetitive observations provided by both our exploration strategy and data augmentation available in contrastive learning improve not only the sample efficiency but also the generalization. Performance of existing model-free RL methods such as Soft Actor-Critic built on top of CCFDM outperforms prior state-of-the-art pixel-based RL methods on the DeepMind Control Suite benchmark.

---

*This work was partly supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data and No. 2021-0-01381, Development of Causal AI through Video Understanding and Reinforcement Learning)*

†Equal contribution

---

**I. INTRODUCTION**

In recent years, Reinforcement Learning (RL) has received considerable attention for its achievements in games, robotics, and autonomous driving. In particular, the Deep Q-Network [1] outperforms human on Atari games [2]. The AlphaGo [3] and AlphaZero [4] defeat professional Go players. Recent advances in deep RL has even enabled agents to perform complex control tasks directly from visual observation of the environment: solving complex task from first-person view observations [5], [6], autonomously performing robotic tasks [7], [8], [9], [10].

Despite the aforementioned successes, sample efficiency and generalization are two main challenges in performing robotic tasks directly from visual observation. There has been remarkable progress in improving sample efficiency and generalization such as CURL [11], RAD [12], DrQ [13], Planet [14], SAC-AE [15], SLAC [9]. Therein, combination of contrastive learning and data augmentation techniques from computer vision with model-free RL show certain improvements in sample efficiency on common RL benchmarks such as Atari [2], DeepMind control [16], ProcGen [17], and OpenAI gym [18]. However, these methods do not utilize temporal information of consecutive observations and strong exploration strategies, leading to limited performance.

This paper proposes Curiosity Contrastive Forward Dynamics Model (CCFDM), a framework that provides efficient image encoder learning for most of RL algorithms. CCFDM consists of an image encoder, a momentum encoder, a forward dynamics model (FDM), and a curiosity module. The connections of all modules are shown in Fig. 1.

CCFDM wisely incorporates data augmentation and the
FDM (a model that predicts next observation given current observation and action) to force the image encoder to capture both spatial and temporal information of visual observations by unsupervised contrastive learning. Consecutive observations potentially complement information each other or provide physics features. Image encoder has to utilize all information from provided transitions to extract meaningful features in order to satisfy both the FDM (temporal prediction) and the data augmentation (spatial disturbances) under CCFDM. In addition, contrastive learning helps extracted features more discriminative which is easier for training RL.

During training, the curiosity module provides intrinsic rewards in addition to extrinsic rewards from the environment. The intrinsic rewards encourage the curiosity of the RL agent to explore novel observations. The prediction error of the FDM is used as the intrinsic reward. Generating intrinsic rewards in this manner is efficient since it utilizes existing FDM and does not introduce extra models. Intuitively, the FDM prediction error (intrinsic reward) is high for novel observations, so that the agent is encouraged to explore these observations. Providing a more diverse and less repetitive set of observation not only makes the image encoder learning efficient but also improves generalization.

The whole framework CCFDM is trained in an end-to-end manner with the RL algorithm is built on top of the query encoder and uses additional intrinsic reward signals from the curiosity module. CCFDM is evaluated on a diverse set of image-based continuous control tasks from DeepMind Control Suite [16]. Empirical results showed that CCFDM improves feature representation learning in terms of data-efficiency and generalization, indicated by RL performance, compared to recent state-of-the-art methods.

The contributions of this paper are threefold: (i) incorporate FDM, augmentation, and contrastive learning to force the image encoder to capture meaningful features of visual observation, (ii) introduce simple but effective curiosity module which can utilize existing FDM to achieve better exploration, and (iii) provide an end-to-end framework which is compatible with most of RL algorithm and easy to implement. We thoroughly analyze the results and show that CCFDM improves feature representation learning in terms of data-efficiency and generalization compared to recent state-of-the-art methods.

II. Related Work

Reinforcement Learning (RL). RL is a research field in machine learning that aims to form a software agent that can perform actions in an environment so as to maximize some notion of cumulative reward. RL algorithms can be classified into model-based RL (e.g: [20], [21], [22], [23]) and model-free RL (e.g: [24], [25]). Model-based RL has access to (or learns) a model of the environment and it provides better sample efficiency than model-free RL. In case model-based RL has to learn the model of the environment, it is extremely difficult to achieve performance better than that of model-free RL: the learned model is usually less accurate than that of model-free RL. By contrast, model-free RL foregoes the potential gains in sample efficiency for gaining convenience in design. These algorithms outperform humans at board games [3], [4], computer games [26], [27], and complex robotic tasks [28], [29]. CCFDM can be used for most model-free reinforcement learning (RL).

Intrinsic Reward exploration. Intrinsic rewards as exploration bonuses are one of the well-known approaches to better exploration, especially for solving hard-exploration problems [30]. Intrinsic reward exploration is also known as curiosity exploration and the methods diverge ranging from count-based exploration (counting by density model [30], [31], [32], counting after hashing [33]) to prediction-based exploration (forward dynamics [34], [35], [36], random networks [37], [38], physical properties [39]). Intrinsic reward exploration has improved sample efficiency in conventional RL. CCFDM utilizes prediction-based exploration using a forward dynamics model and incorporates it to the framework of contrastive learning for visual observation RL.

Dynamics Models for Sample-efficient RL. Modeling the dynamics of the environment has proved effective in improving sample efficiency. The dynamics model can be used for generating more data for Atari games [40], planing ahead [14], [41], shaping the representations using an auxiliary loss [5], [42], [9]. CCFDM estimates a forward dynamics model to force the image encoder to capture temporal information and learn meaningful representation which is sufficient for dynamics prediction.

Representation Learning for RL. Representation learning plays an important role in achieving high performance in visual-based RL. Many methods have been proposed with various approaches. [5] proposes representation auxiliary loss for improving RL performance in the DeepMind Control Suite [43]. Using a reconstruction-based task is also an effective approach. Beta variation autoencoder (β-VAE) [44] or deterministic autoencoder AE is leveraged for encoding features. [45], [15] jointly learn VAE/AE objectives and RL objectives while [46], [47] propose to train the two objectives in an alternating fashion. Recently, contrastive learning in

---

Algorithm 1 Curiosity Contrastive Forward Dynamic Model

**Input:** Batch size $K$, training steps $M$, EMA factor $\tau$, momentum update frequency $\eta$, intrinsic weight $C$, intrinsic decay weight $\gamma$

1. **Initialize:** Replay Buffer $D$, QE, $f_{\theta_q}$, $g_{\theta}$, FDM $g_{\omega}$, AE $h_{\omega}$

2. **for** $k = 1, M$ **do**

3. **Sample** transitions $B = \{(o_j, a_j, o'_j, r_j)\}_{j=1}^K$ from $D$

4. **Augment** $B$ to get $\hat{B} = \{(\hat{o}_j, a_j, \hat{o}'_j, r_j)\}_{j=1}^K$

5. **Compute** query observation feature, query, key:

   $q = f_{\theta_q}(\hat{o}_j), q' = g_{\omega}(q, h_{\omega}(a))$

6. **Compute** intrinsic reward $r_i$ by CM using Eq. 9

7. **Update** $f_{\theta_q}, h_{\omega}$, and $g_{\omega}$ by contrastive loss in Eq. 8

8. **Update** $f_{\theta_q}, D$ by RL algorithm with intrinsic rewards $r_i$

9. **if** $k \mod \eta = 0$ **then**

10. **Update** momentum-based network $\hat{f}_{\theta_q}$:

11. $\hat{\theta}_k \leftarrow \tau \theta_k + (1 - \tau)\theta_k$

12. **end if**

13. **end for**
To obtain a good policy, the RL algorithm has to balance between exploration (discover novel observations) and exploitation (follow the current best policy). A simple approach is to provide exploration bonuses as rewards (a.k.a. intrinsic reward $r_i$) to encourage policy reaching novel observation during training in order to provide a diverse set of observations. This prevents the policy from getting stuck in the local minimal. Thus, during training, the policy has to maximize the new expected cumulative reward:

$$F(\pi) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_e(o_t, a_t) + r_i(o_t, a_t) \right].$$

## III. BACKGROUND

### A. Reinforcement Learning with Intrinsic Reward

Reinforcement Learning (RL) considers solving a Markov Decision Process (MDP). MDP is characterized by a set of state $O$, a set of action $A$, a transition probability $P$ mapping from current observation $o$ and action $a$ to the future observation $o'$, and the immediate reward $R$, also known as extrinsic reward (denoted by $r_e$). Given the policy $\pi \in \Pi$: $O \rightarrow A$, the goal of RL is to learn an optimal policy that maximizes the expected cumulative reward:

$$F(\pi) = \mathbb{E}_{o_t \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r_e(o_t, a_t) \right].$$

RL algorithms that explicitly estimate (or are given) transition probability are considered model-based RL, the others are considered model-free RL.

To balance exploration and exploitation, the RL algorithm has to consider both exploration (discover novel observations) and exploitation (follow the current best policy). A simple approach is to provide exploration bonuses as rewards (a.k.a. intrinsic reward $r_i$) to encourage policy reaching novel observation during training in order to provide a diverse set of observations. This prevents the policy from getting stuck in the local minimal. Thus, during training, the policy has to maximize the new expected cumulative reward:

$$F(\pi) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_e(o_t, a_t) + r_i(o_t, a_t) \right].$$

### B. Soft Actor-Critic

Soft Actor-Critic (SAC) [53] is an effective RL algorithm for learning robotics tasks. SAC learns a policy $\pi_\theta$ (a.k.a. actor) and critics $Q_{\phi_1}$ and $Q_{\phi_2}$. $\phi_i$ are learning by minimizing the Bellman error:

$$\mathbb{E}_{b \sim D} \left[ (Q_{\phi_1}(o, a) - (r_e + \gamma(1-d)T))^2 \right],$$

where $b = (o, a, o', r_e, d)$, $d$ is done signal, $D$ is the replay buffer, and $T$ is defined as:

$$T = \min_{i=1,2} \left[ Q_{\phi_i}(o', a') - \alpha \log \pi_\theta(a' | o') \right].$$

The $Q^*_\phi$ denotes the exponential moving average (EMA) of the parameters of $Q_\phi$, and $\alpha$ is a positive entropy coefficient. The actor is trained by maximizing the expected return of its actions as in:

$$\mathcal{L}(\theta) = \mathbb{E}_{a \sim \pi} \left[ Q^\pi(o, a) - \alpha \log \pi_\theta(a | o) \right].$$
C. Contrastive Learning

Contrastive learning is an approach to improve representation learning by a teaching model whereby pairs of data points are “similar” or “different”. To be specific, given a query \( q \) and keys \( K = \{ k_0, k_1, \ldots \} \) where \( K \) includes the positive key \( k^+ \) and the negative keys \( K \setminus \{ k^+ \} \). The goal of contrastive learning is to ensure that \( q \) matches with \( k^+ \) and is far apart from \( K \setminus \{ k^+ \} \). One example of contrastive learning loss [19] is as below:

\[
\mathcal{L}_q = - \log \frac{\exp (q \cdot k^+ / \tau)}{\sum_{i=0}^{K} \exp (q \cdot k_i / \tau)}
\]

(6)

where \( \tau \) is a temperature hyper-parameter.

IV. CCFDM IMPLEMENTATION

The proposed framework extends the model-free RL algorithm by adding modules to aid image encoder parts. The whole architecture of the framework is shown in Fig 2. There are five main modules: query encoder (QE), key encoder (KE), action embedding (AE), forward dynamics model (FDM), and curiosity module (CM). The QE is a deep convolutional neural network that maps visual observation to a feature vector. The KE is the moving average version of the QE similar to MoCo [19] which helps stabilize the learning process of QE. The AE and the FDM is multi-layer perceptron networks. Therein, AE encodes action to form an appropriate action feature for FDM to predict the next observation feature (a.k.a query). The curiosity module is a non-learning module that provides intrinsic rewards.

Formally, we denote the QE is \( f_{\theta_q} \), the KE is \( f_{\theta_k} \), the AE is \( h_{\omega} \), the FDM is \( g_{\phi} \) and replay buffer is \( D \). A work flow of framework is as follow. A batch of transitions \( B \) is sampled from \( D \). Observations are augmented to form query observations \( o_q \) and next observations are augmented to form key observations \( o_k \). The query observations, the key observations, and the actions are then encoded with the QE, the KE, and the AE to form feature representations:

\[
q = f_{\theta_q}(o_q); \quad h_q = f_{\theta_k}(o_k); \quad a_c = h_{\omega}(a)
\]

(7)

Action feature (\( a_c \)) and query observation feature (\( q \)) are input to FDM to predict query (\( q' \)): \( q' = g_{\phi}(q, a_c) \). The query-key pairs (the query, the positive key, negative keys) are passed to the contrastive learning objective. Any contrastive learning objective can be used but we found that infoNCE [54] showed the best performance. The formulation is below:

\[
\mathcal{L}(q', k') = - \log \frac{\exp (\text{sim}(q', k^+))}{\exp (\text{sim}(q', k^+)) + \sum_{j=1}^{K} \exp (\text{sim}(q', k'_j))}
\]

(8)

Where \( j \) is the sample index in \( B \), \( K \) is the batch size and \( \text{sim}(\cdot) \) is the similarity measure function. It can be Bilinear products \( \text{sim}(q', k') = (q'^T W k') \) or dot products \( \text{sim}(q', k') = (q'^T k') \).

The query-positive key pairs are passed to the curiosity module to calculate intrinsic rewards \( r_i \). The curiosity module measures the similarity between the query and positive key then gives intrinsic reward proportional to the dissimilarity. The similarity is not bounded and differs from task to task. We normalize the similarity to make it task-agnostic. Furthermore, the intrinsic reward is decayed during training to make agents converge to the optimal solutions. The formulation for CM is below:

\[
r_i(q, k) = C e^{(-\gamma t)} \text{sim}(q, k) \frac{e^{\gamma \max}}{e^{\gamma \max}}
\]

(9)

Therein, \( t \) is environment step, \( C \) is temperature weight, \( \gamma \) is decay weight, \( e^{\gamma \max} \) and \( e^{\gamma \max} \) are the maximum extrinsic reward value and intrinsic reward value over \( t \), respectively.

Finally, the \( q \) and \( r_i \), \( B \) is passed to the RL algorithm. During the gradient update, the CCFDM updates the query encoder, action embedding, and FDM simultaneously. Note that \( D \) is updated during RL training. The whole algorithm is described the Alg. 1.

V. EXPERIMENT AND RESULT

A. Experiment setup

The proposed framework is benchmarked on six continuous control tasks from DeepMind Control suite (DMC) [16]. DMC is considered as a standard benchmark for evaluating visual observation RL algorithms [11], [12], [13] in term of sample efficiency and generalization. DMC provides excellent physical model environments referred to as domains. Each domain has different tasks associated with a particular MDP structure. Our experiments are conducted on six well-known domains as shown in Figure 3. The following tasks
Forward dynamics model (FDM) and action embedding (AE) layers with ReLU encoder (QE) architecture consists of four convolutional layers with ReLU activation followed by a fully connected projection layer which is similar to SAC-AE [15]. The forward dynamics model (FDM) and action embedding (AE) are modeled as Multi-Layer Perceptron with two hidden layers of 50 ReLU units. The key encoder (KE) architecture is identical to QE. The KE weights are the moving average of the KE weights which are similar to MoCo [19]. The EMA coefficient $\tau = 0.1$. Soft Actor Critic (SAC) [53] is used as a base RL algorithm similar to [12], [15]. The actor and critic use the QE for feature extraction. As with previous algorithms [15], [12], the batch size is set to 512, the target critic and the target actor are updated every two updates of the main critic. We use random cropping for data augmentation throughout the experiments. The FDM is optimized using Adam optimizer [44] with default parameters and initial learning rate $1e^0$. Intrinsic reward decay weight is set to 2e-5 and intrinsic weight is set to 0.2. All other settings are the same as mentioned in Curl [11].

### B. Experiment result

During training, we simultaneously evaluated the RL agent every 10K environment steps with 10 episodes and logged the average test returns. For each task, we trained our algorithm five times with different seeds and report the result as Table I. The result demonstrates CCFDM significantly improved performance over the baselines: SAC-CURL [11], PlaNet [14], SAC-AE [15], SLAC, SAC-Pixel and SAC State [53].

The sample-efficiency is evaluated by the performance at 100k environment step and 500k environment step as proposed in CURL [11]. The average result over six tasks is shown in Fig 4. According to the result, at the 100K environment step, CCFDM gains $3.3\times$ higher median performance than learning from pixel only (SAC Pixel), around $1.3\times$ higher than SLAC and CURL, CCFDM converges close to an optimal score of 1000 on all six tasks within 500K steps. It also nearly matches the SAC State together with the DrQ. The Fig 5 shows evaluation score curves throughout the

---

<table>
<thead>
<tr>
<th>500K STEP SCORES</th>
<th>CCFDM (Ours)</th>
<th>DrQ</th>
<th>SAC-CURL</th>
<th>PLANET</th>
<th>SAC-AE</th>
<th>SLAC</th>
<th>SAC-Pixel</th>
<th>SAC State</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINGER,SPIN</td>
<td>906±152</td>
<td>938±103</td>
<td>874±151</td>
<td>418±382</td>
<td>884±128</td>
<td>771±203</td>
<td>179±166</td>
<td>927±43</td>
</tr>
<tr>
<td>CARPOTLE,SWINGUP</td>
<td>875±38</td>
<td>868±10</td>
<td>861±30</td>
<td>464±50</td>
<td>735±63</td>
<td>-</td>
<td>419±40</td>
<td>870±7</td>
</tr>
<tr>
<td>REACHER,EASY</td>
<td>973±36</td>
<td>942±71</td>
<td>904±94</td>
<td>351±483</td>
<td>627±58</td>
<td>-</td>
<td>145±30</td>
<td>975±5</td>
</tr>
<tr>
<td>CHEETAH,RUN</td>
<td>552±130</td>
<td>660±96</td>
<td>500±91</td>
<td>321±104</td>
<td>550±34</td>
<td>629±74</td>
<td>197±15</td>
<td>772±60</td>
</tr>
<tr>
<td>WALKER,WALK</td>
<td>929±68</td>
<td>921±46</td>
<td>681±68</td>
<td>293±114</td>
<td>847±48</td>
<td>865±97</td>
<td>41±12</td>
<td>974±8</td>
</tr>
<tr>
<td>BALL IN CUP,CATCH</td>
<td>979±17</td>
<td>963±9</td>
<td>958±13</td>
<td>352±467</td>
<td>794±58</td>
<td>959±4</td>
<td>312±63</td>
<td>979±6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>100K STEP SCORES</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FINGER,SPIN</td>
<td>880±142</td>
<td>901±104</td>
<td>779±108</td>
<td>95±164</td>
<td>740±64</td>
<td>680±130</td>
<td>179±66</td>
<td>672±76</td>
</tr>
<tr>
<td>CARPOTLE,SWINGUP</td>
<td>785±87</td>
<td>759±92</td>
<td>592±170</td>
<td>303±71</td>
<td>311±11</td>
<td>-</td>
<td>419±40</td>
<td>812±45</td>
</tr>
<tr>
<td>REACHER,EASY</td>
<td>811±220</td>
<td>601±213</td>
<td>517±113</td>
<td>140±256</td>
<td>274±14</td>
<td>-</td>
<td>145±30</td>
<td>919±123</td>
</tr>
<tr>
<td>CHEETAH,RUN</td>
<td>274±98</td>
<td>344±67</td>
<td>307±48</td>
<td>165±123</td>
<td>267±24</td>
<td>391±47</td>
<td>197±15</td>
<td>228±95</td>
</tr>
<tr>
<td>WALKER,WALK</td>
<td>634±132</td>
<td>612±164</td>
<td>323±43</td>
<td>125±57</td>
<td>394±22</td>
<td>428±74</td>
<td>42±12</td>
<td>604±317</td>
</tr>
<tr>
<td>BALL IN CUP,CATCH</td>
<td>962±28</td>
<td>913±53</td>
<td>772±241</td>
<td>198±442</td>
<td>391±82</td>
<td>607±173</td>
<td>312±63</td>
<td>957±26</td>
</tr>
</tbody>
</table>

TABLE I: Scores achieved by CCFDM (mean & standard deviation) and baselines on DMC evaluated at 500k environment step and 100k environment step. CCFDM achieves state-of-the-art performance on the 4 out of 6 environment and just below DrQ on Finger-Spin and Cheetah-Run. The baseline are DrQ [13], SAC-CURL [11], PlaNet [14], SAC-AE [15], SLAC [9], SAC-Pixel and SAC State [53].
whole training process. This clearly proves the effectiveness of CCFDM compared to other methods.

C. Discussion

The SAC State is trained on hand-crafted features, considered the upper bound for performance and sample efficiency. However, according to Fig 5, CCFDM even outperforms SAC state in the Finger-Spin and Ball in Cup-Catch to some extent. It implies that hand-crafted (human-made) features are sometimes not the optimal features and thus learned features are potentially better features and produce better results without human effort.

The batch size has an important effect on our performance. During training and fine-tuning hyper-parameters, we realized that the larger batch size gives better performance. This is reasonable since CCFDM uses contrastive learning. However, a large batch size results in much lower training wall time. Thus, we should use it wisely.

We tried to incorporate CURL [11] to CCFDM since doing so is quite straightforward. Interestingly, we found out that the performance remains the same while the training wall time increases due to additional computational cost. We argue that CCFDM contains all the CURL’s effects. However, we need to work more to prove it with comprehensive experiments. The authors leave this task for future research.

VI. Conclusion

This paper proposed a Curiosity Contrastive Forward Dynamics Model (CCFDM) framework, a contrastive reinforcement learning (RL) based framework that provides efficient image encoder (IE) learning for most RL algorithms. CCFDM showed that incorporating a forward dynamics model into contrastive learning framework helps IE capture both spatial and temporal information in order to improve IE learning. CCFDM also provides a smart exploration strategy based on FDM error which is simple but effective to encourage agents to better explore novel observations. CCFDM has proved effective in terms of improving sample efficiency and generalization. CCFDM is easy to incorporate with different modules to further improve RL performance. Future research should focus on developing more efficient and lightweight modules that can be integrated with CCFDM.

REFERENCES


