

ATTENTION-BASED CURIOSITY-DRIVEN EXPLORATION IN DEEP REINFORCEMENT LEARNING

Patrik Reizinger and Márton Szemenyei*

Budapest University of Technology and Economics
Department of Control Engineering and Information Technology
Budapest, Hungary
rpatrik1996@gmail.com

ABSTRACT

Reinforcement Learning enables to train an agent via interaction with the environment. However, in the majority of real-world scenarios, the extrinsic feedback is sparse or not sufficient; thus, intrinsic reward formulations are needed to successfully train the agent. Our aim is to develop means for the better incorporation of state- and/or action-dependent information into existing intrinsic reward formulations with the attention mechanism. We propose new methods, such as Attention-aided Advantage Actor-Critic, an extension of the Actor-Critic framework. Second, a curiosity-based approach - Intrinsic Curiosity Module - is extended. The proposed model utilizes attention to emphasize features for the dynamic models within Intrinsic Curiosity Module; moreover, we also modify the loss function, resulting in a new curiosity formulation, which we call rational curiosity (RCM). The corresponding implementation can be found at <https://github.com/rpatrik96/AttA2C/>.

Index Terms— Reinforcement Learning, curiosity, exploration, attention

1. INTRODUCTION

Reinforcement Learning (RL) agents are generally trained based on the feedback, i.e., using rewards collected from the environment. However, the environment is not guaranteed to have an inner dynamics that rewards the agent in a way that corresponds with its goals. Thus, only having an extrinsic incentive may be impractical. Furthermore, rewards are in most real-life scenarios sparse, so it is more difficult for the agent to assign credit to useful actions and omit disadvantageous ones.

To mitigate the above-mentioned problems, several methods were developed to supplement the objective function of

an RL agent. The contribution of this paper concerns the extension of existing methods for intrinsic reward formulation to better incorporate state-/action-dependent information. Before reviewing our work in detail, let us summarize the main contributions in this field. In model-based RL, the World model architecture [1] is one answer for challenges mentioned above. On the other hand, curiosity-based methods modify the loss function (or even the network architecture) by adding terms to incentivize exploration. The Intrinsic Curiosity Module (ICM) module of [2] is a major contribution, which will be used extensively in this work to build upon. ICM introduces inner dynamics (forward and inverse) to quantify the prediction error of the next state and action, which is utilized as an intrinsic reward - a more extensive study can be found in [3]. Another promising work in this field is the notion of disagreement, which is an ensemble-based curiosity formulation [4].

2. ATTENTION-BASED CURIOSITY

In this section, we combine curiosity-driven exploration and attention [5] to develop new, adaptive means for intrinsic rewards.

2.1. Attention-aided Advantage Actor-Critic (AttA2C)

RL agents often include some variant of Actor-Critic network, which has two components: the actor predicts action probabilities, while the critic provides an estimate about how good the state is.

We conjecture that separating the feature space into two parts can be advantageous (inspired by the Generative Adversarial Network (GAN)[6]–Actor-Critic correspondence [7]), as different features may be useful for the actor and for the critic. The proposed architecture aims to overcome this by utilizing attention. As shown in Figure 1, the input from the environment is fed through separate attention layers, thus both heads can put emphasis onto the most important parts of the latent process, which is formulated as follows: $a_t =$

*Patrik Reizinger is the main author of this work. His contribution consist of developing the basis for the proposed methods in this work. Supported by the ÚNKP-19-2 New National Excellence Program of the Ministry for Innovation and Technology

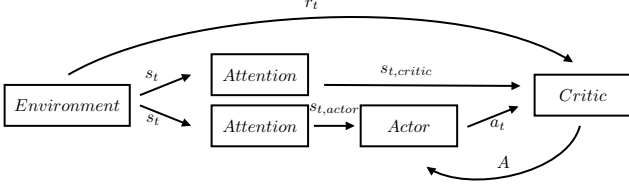


Fig. 1: The AttA2C architecture

$Actor(Attention_{\pi}(\phi(s_t)))$ for the actor and $A_{\pi}(s_t, a_t) = Critic(Attention_A(\phi(s_t)))$ for the critic, where ϕ is the feature transform of the input. Subscripts denote that two separate attention layers are used to predict the policy π and the advantage function A .

2.2. Feature- and action selective ICM

Being able to separate the most useful features cannot only be advantageous for the Actor-Critic network, but also for the curiosity formulation. To do that, attention is applied onto the concatenation of the inputs of the forward and inverse models (denoted by the fwd and inv subscripts, respectively), i.e. $\hat{\phi}_{t+1} = Forward(Attention_{fwd}([\phi(s_t), a_t]))$ and $\hat{a}_t = Inverse(Attention_{inv}([\phi(s_t), a_t]))$, where $\hat{\cdot}$ stands for predicted values. The equations above describe the single attention case, but we also experimented with double attention, in which we swapped the order of concatenation and attention to separate the weighting between the two domains.

2.3. Rational Curiosity Module (RCM)

The curiosity-driven exploration strategies - such as that of [2] and [4]- are based on the assumption that enforcing curiosity-driven exploration is a good choice for every state and action, even though this is not true in every scenario, e.g. consider the noisy TV experiment of [3].

To mitigate the above case, an attention network [5] is used on the forward loss function of the ICM module [2]. The new forward loss term is illustrated in Figure 2, where the round-headed arrow denotes that $\phi(s_{t+1})$ is used as the control term in the attention layer. Thus, the loss function is modified in the following manner $J_{rcm} = J_{A2C} + \beta Attn(J_{fwd}, \phi(s_{t+1})) + (1 - \beta) J_{inv}$, where RCM is the term for the new model, J_{A2C} is the objective function for the Advantage Actor-Critic (A2C) network, J_{fwd} for the forward and J_{inv} for the inverse dynamics, weighted with a

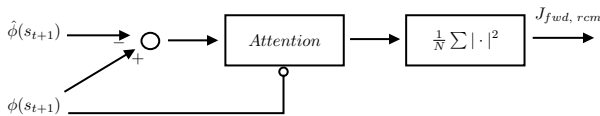


Fig. 2: Forward loss formulation of the RCM agent

Agent	Mean normalized reward	Std. dev.
Baseline	92.33	4.98
AttA2C	88.01	13.36
$ICM_{1 \times Att}$	94.88	4.46
$ICM_{2 \times Att}$	92.17	5.68
RCM	95.31	6.48

Table 1: Comparison of the agents' normalized mean performance (higher is better)

scalar β . We hypothesize that this formulation can help to utilize curiosity only in situations where the agent can benefit from that on the long run, but omits curiosity otherwise.

2.4. Implementation

The proposed methods are implemented in PyTorch [8], the agents are based on the implementation of [2]. Five configurations were implemented: AttA2C, single- and double-attention ICM and RCM, and the traditional ICM agent is used as a baseline [2].

For test purposes, the Atari environments of the OpenAI Gym [9] are used via the Stable-baselines [10] package. Three environments were chosen, i.e. Breakout, Pong and Seaquest (as in [4]). All agents were trained on 4 environments in a parallel manner for 2,500,000 rollouts with 5 steps each, using 4 stacked frames.

3. RESULTS

We monitored two metrics to compare performance and generalization: mean reward for the former, and mean standard deviation of the features for the latter. Due to lack of space, only numerical results for are shown here (figures are available on GitHub).

To be able to provide a concise summary of the agents' performance in all environments, we normalized the highest rewards of each agent, with 100 denoting the best performance. This way, we were able to compare the mean performances, summarized in Table 1. As it shows, the RCM agent performed the best (even in a statistically significant manner), followed by the single attention ICM.

4. DISCUSSION

This work investigated the paradigm of curiosity-driven exploration in RL, extended by the attention mechanism. We proposed three different methods to incorporate attention for utilizing curiosity in a selective manner. The new models were tested in the OpenAI Gym environment and have shown consistent improvement to the baseline models used for comparison.

5. REFERENCES

- [1] David Ha and Jürgen Schmidhuber, “World models,” *arXiv preprint arXiv:1803.10122*, 2018.
- [2] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell, “Curiosity-driven exploration by self-supervised prediction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 16–17.
- [3] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros, “Large-scale study of curiosity-driven learning,” *arXiv preprint arXiv:1808.04355*, 2018.
- [4] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta, “Self-supervised exploration via disagreement,” *arXiv preprint arXiv:1906.04161*, 2019.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” 2017.
- [6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [7] David Pfau, “Connecting GANs, Actor-Critic methods and multilevel optimization,” 2017.
- [8] Adam Paszke, Gregory Chanan, Zeming Lin, Sam Gross, Edward Yang, Luca Antiga, and Zachary Devito, “Automatic differentiation in PyTorch,” *Advances in Neural Information Processing Systems 30*, , no. Nips, pp. 1–4, 2017.
- [9] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba, “OpenAI Gym,” 2016.
- [10] Ashley Hill, Antonin Raffin, Maximilian Ernestus, Adam Gleave, Rene Traore, Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, and Yuhuai Wu, “Stable baselines,” <https://github.com/hill-a/stable-baselines>, 2018.