Intrinsic Motivation Based on Predictive Surprise Model

Matej Pecháč and Igor Farkaš

***Abstract*—Intrinsic motivation research is a promising part of reinforcement learning which can push artificial agents to completely new frontiers. Namely, from simple reaction agents depending on the reward engineered by human, to more autonomous agents with their own goals and skill development able to act successfully in the environments which are unknown to their human authors. In this paper, we introduce an extension of the intrinsic motivation model, which combines via gating two different motivational signals based on a prediction error and predictive surprise and thus allows the agent to accelerate the exploration of the environment and find sources of an external reward that the agent maximizes. We focus on the speed-up of the reinforcement learning process using three environments with dense and sparse rewards, showing the superior performance of an agent with a gated reward. Our experiments also confirm the hypothesis that intrinsic motivation can help stabilize the learning process. Last but not least, we introduce a visualization tool that allows us to analyze the model behavior and hence reveal its dynamics. The source code is available at https://github.com/Iskandor/MotivationModels.**

***Index Terms*—reinforcement learning, intrinsic motivation, prediction error, predictive surprise, active learning**

I. Introduction

The development of reinforcement learning [1] methods has recently achieved much success, since together with advances in computer vision (mainly convolutional networks), it has become possible to teach agents to solve various tasks, play simple computer games, even overcoming human players. Nevertheless, these are still concrete single tasks. A lot of computer time has to be spent, and the agents are given a lot of resources to manage to learn the aforementioned challenges in a reasonable time. However, coping with a complex (continuous) environment such as our world is still a challenge. There are several pathways offering research opportunities. One is the search for new optimization and learning methods that would shorten the learning time or reduce the amount of resources needed (CPU instead of GPU). Another is hardware development, which attempts to adapt to the requirements of neural networks that are currently being used in the field of reinforcement learning.

Yet another, conceptually different approach adds a new dimension to reinforcement learning, namely *intrinsic motivation* (IM) that is believed to provide means for open-ended learning [2]. IM has a strong psychological motivation, since children acquire skills and knowledge about the world using their own drive and experience without obvious reward from the outside environment. This obviously prepares the children for later play, for school and real life. Therefore, if we want to achieve an open-ended development with our agents, we have to master this first step and equip them with an ability to generate their own goals and acquire new skills. Only in this way will they be able to adapt effectively and in real time in an unknown environment (which is also our real world for them). Therefore, computational approaches concerned with IMs and open-ended development are thought to have the potential to lead to the construction of truly intelligent artificial systems, in particular systems that are capable of improving their own skills and knowledge autonomously and indefinitely [2].

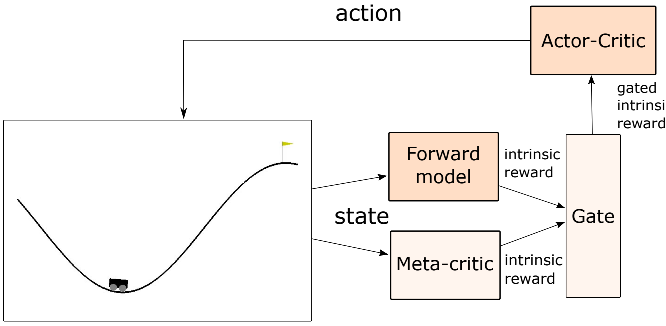


Fig. 1: The concept of an intrinsic motivation model with a gated reward signal.

A lot of IM-related research has been done, as briefly summarized in the next chapter. In this paper, we introduce a new version of a IM-based agent that is shown to efficiently learn the tasks at hand. In particular, we provide three main contributions:

- We extend an intrinsic motivation model based on prediction error introducing a *meta-critic* module.
- We explore a gating approach to exploit two motivation signals generated in the intrinsic module of the agent.
- We present a simple visualization tool for deeper analysis of intrinsic motivation modules behavior and monitoring the learning process in a model.

II. Related work

Motivation represents a complex of psychological phenomena such as novelty, surprise, incongruity or challenge. Therefore, various theories have been developed, out of which we will briefly describe some:

- Theory of drives [3] is based on a statement that humans are looking for options to secure their basic needs, to explore the environment [4], or look for ways to control it [5].

Department of Applied Informatics, Comenius University in Bratislava, Slovak Republic, email:{matej.pechac,igor.farkas}@fmph.uniba.sk

- Theory of effectance [6] is described as the desire for effective interaction with the environment.
- Theory of cognitive dissonance [7] explains motivation as a reduction in the differences between the experience gained and the expectations that were created in internal cognitive structures.
- Optimal incongruity theory [8] proposes that a person is motivated by stimuli that differ from the standard stimuli he has already experienced.
- Theory of "Flow" [9] assigns the greatest motivation to solving problems with optimal difficulty. In simple tasks, a person begins to get bored quickly, and on the other hand, without the ability to find a solution a person becomes frustrated.
- Synchronicity detection is a crucial mechanism in object interaction skills [10] or self-modelling [11].

The concept of motivation can be divided into *external* and *internal*, depending on the mechanism that generates motivation for the agent. If the source of motivation comes from outside, we are talking about *external* motivation, and it is always associated with a particular goal in the environment. If the motivation is generated within the structures that make up the agent, it is an *internal* motivation. Another dimension for the distinction is not so obvious: *Intrinsic* motivation is defined as something goalless for its own sake, that is fun or challenge, whereas *extrinsic* motivation has some goal defined separately [12]. As emphasized by the authors, it is a vague definition of the difference and we are inclined to that view. Further we will use the concept of *intrinsic* motivation, but it can be understood as an *internal* motivation.

For simplicity, in our notation, we consider only *the state* as an input for the presented models. In a more general case, we should also consider transitions between states, observation of state conditioned by previous state and more events, which could also serve as inputs.

The computational approaches to *intrinsic* motivation, including the ones mentioned above, can be divided into three main categories. *Knowledge-based* approach is focused on exploration of the environment and contains information theoretic and distributional methods, predictive methods and learning progress methods. This approach is based on the theory of drives, theory of cognitive dissonance and optimal incongruity theory. *Competence-based* approach motivates the agent to achieve higher level of performance in the environment, which means to acquire desired actions to achieve self-generated goals. Its psychological basis includes the theory of effectance and the theory of flow. *Morphological* approach is based on the synchronicity detection theory and motivates the agent to stay in more or less (depending on the motivation signal) stable state according only to its sensory inputs.

Our model based on predictive surprise can be included into category of knowledge-based approaches focusing on exploration. Models from this category use mainly three different methods to generate intrinsic motivation. The first group is based on prediction error often using forward model as core component, e.g. [13] which was successfully tested on Atari games environments [14]. These methods have to deal with the so-called white-noise problem [15]. The ICM model [16] used additionally an inverse model for better feature extraction (to prevent white-noise problem) and was tested on challenging environments like VizDoom [17]. Features can be also extracted from variational autoencoder [18] and then used for computation of a prediction error. For more details we refer to [19]. Very good results were obtained by EMI model [20] combining the predictive approach with mutual information which is maximized and used as a motivation signal. The EMI architecture also includes a module estimating the prediction error, similarly to our model, but it is used differently to stabilize the EMI learning process.

The second group monitors the state novelty and the intrinsic signal is based on its value. The first models were based on count-based approach [21]. This method is impractical for large or continuous state spaces and it was extended by introducing pseudo-count and neural density models [22]–[24]. A similar method to pseudo-count was used by RND model [25] with a lower complexity.

The last group uses motivation signals based on information theory quantities like information gain [19], [26], [27], mutual information [28], predictive information [29] or empowerment [30], [31].

III. Preliminaries

Reinforcement learning [1] is a domain of machine learning focused on learning from interaction of the agent and environment solely from its own experience. The environment is often formalized as Markov decision process (MDP) which consists of a state space $\mathcal{S}$, action space $\mathcal{A}$, transition function $\mathcal{T}(s,a,s') = p(s_{t+1} = s'|s_t = s, a_t = a)$, reward function $\mathcal{R}$ and a discount factor γ. The main goal of the agent is to maximize the expected return

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

in each state. Stochastic policy is defined as a state dependent probability function $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$, such that

$$\pi_t(s,a) = p(a_t = a|s_t = s) \quad \text{and} \quad \sum_{a \in \mathcal{A}} \pi(s,a) = 1$$

and deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is defined as

$$\pi(s) = a$$

An agent following the optimal policy π^* maximizes the expected return R. The methods searching for the optimal policy can be divided into on-policy (family of actor–critic algorithms) and off-policy (family of Q-learning algorithms and its derivatives) methods.

IV. Methods

In this section we describe the formal approach to the extension of an intrinsic module based on the prediction error (see Fig. 1). The extended module provides for a short time a larger amount of intrinsic reward to the agent, especially in the first phases of learning. This phenomenon can be described as predictive surprise and our hypothesis is that it can lead to significant improvement in the training speed in environments with sparse reward. Our expectations of performance in environments with dense rewards are that it can lead to a more stable policy because of exhaustive exploration performed mainly in the first period of learning process. Some of our first observations also suggested that the number of parameters in policy and value models could be reduced if our intrinsic motivation module is available and still reach optimal results contrary to the same agent without motivation. Both hypotheses are tested in experiments.

Meta-critic module

Our motivational module is based on two prediction modules. The first module is the forward model $\Pi(s_t, a_t)$ with parameters θ_{fm} which predicts the next state $\hat{s}_{t+1}$ from the current state and action

$$\Pi(s_t, a_t; \theta_{\mathrm{fm}}) = \hat{s}_{t+1} \tag{1}$$

The prediction error e_t is then defined as a (squared) distance between the predicted state $\tilde{s}_{t+1}$ and the next observed state s_{t+1}

$$e_t = \frac{1}{n} \|s_{t+1} - \hat{s}_{t+1}\|^2 \tag{2}$$

where n is the dimension of the state space. The intrinsic reward based on the prediction error is defined as

$$r_t^{\mathrm{ifm}} = e_t \tag{3}$$

The second module models predictive surprise motivation which rewards the states that occur but were not expected or do not occur and were expected. To formalize the expectations, another predictor $\mathrm{Meta}\Pi$ is introduced and in the text is referred to as *meta-critic*. It aims to predict the error e_t of the first predictor Π at time t

$$\mathrm{Meta}\Pi(s_t, a_t; \theta_{\mathrm{mc}}) = \tilde{e}_t \tag{4}$$

where $\tilde{e}_t$ is the predicted absolute error of the predictor Π and θ_{mc} are parameters of the *meta-critic*. Its input is identical to that of the prediction module – the current state and action, but the output is no longer an estimate of the next state, but an estimate of the prediction error. In this way, we obtain qualitatively new information about the state of the agent's internal model about the environment, which describes how confident the agent is in its predictions. Based on this information we designed a new intrinsic reward function

$$r_t^{\mathrm{imc}} = \begin{cases} \frac{e_t}{\tilde{e}_t} + \frac{\tilde{e}_t}{e_t} - 2, & \text{if } |e_t - \tilde{e}_t| < \sigma \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

If the *meta-critic* correctly estimates the prediction error, the reward is close to 0 due to constant 2 which is subtracted from the term. To prevent cases where the estimated error and prediction error are very small, but still generate some reward, we introduced a threshold σ which has to be exceeded. The reward function defined in this way can stimulate an agent if the prediction error is low and its estimate is high, or vice versa, when the prediction error is high and its estimate is low. The whole scheme is shown in Fig. 2. The training of the proposed intrinsic module is straightforward and can be approached as an optimization problem, formulated as

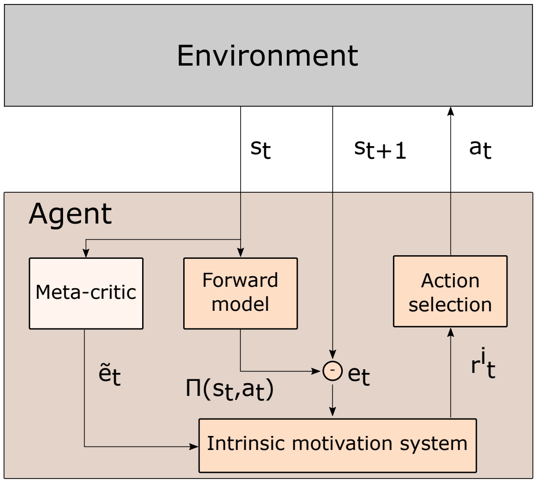


Fig. 2: Simplified scheme of the proposed intrinsic motivation model with highlighted meta-critic module.

$$\min_{\theta_{\mathrm{fm}}, \theta_{\mathrm{mc}}} \left[\frac{1}{n} \|s_{t+1} - \hat{s}_{t+1}\|^2 + \|e_t - \tilde{e}_t\|^2 \right] \tag{6}$$

Intrinsic reward gating

The proposed motivation model has two prediction modules generating two different intrinsic motivation signals. We decided to introduce the gating of reward signals that in each step of an episode puts through only one of the two signals to model situations where the unexpected event suppressed the second signal which is more common in the agent's experience and is not so beneficial for it. The final intrinsic reward added to an external reward is defined as

$$r_t^{\mathrm{i}} = \max(\tanh(r_t^{\mathrm{ifm}}) \cdot \epsilon_{\mathrm{fm}}, \tanh(r_t^{\mathrm{imc}}) \cdot \epsilon_{\mathrm{mc}}) \tag{7}$$

where the reward signals from both modules are scaled to the interval $(-1, 1)$ and then independently scaled by a respective factor ϵ. This procedure was informed by an observation that predictive surprise motivation often overwhelms common predictive motivation and leads to effect of sudden surprise for the agent. Without surprise the agent is driven by predictive motivation.

The above mentioned types of reward were used in four different agents listed in Table I.

TABLE I: Agents with their respective motivation signal.

Agent	Motivation
Baseline	none
Forward model	r^{ifm} (eq. 3)
Meta-critic	r^{imc} (eq. 5)
Meta-critic gated	r^{i} (eq. 7)

Visual analysis of agent learning

To better understand the agent's behavior during the learning process we collected outputs of actor, critic, forward model and meta-critic at the end of each episode and rendered them as a video sequence. In the first step we prepared a set $\mathcal{S}_n$ of n states generated to cover the state space of each environment. For the ones with a low dimension, we covered the state space by a grid and for higher dimensions we collected states from one complete training of baseline agent and then we applied the k-means algorithm to the data which returned a desired number of points in the state space. At the end of each episode we collected outputs of actor, critic, forward model and meta-critic, which were activated on the prepared set $\mathcal{S}_n$. To perform visualization of training in case of higher dimensional state spaces we applied t-SNE [32], a popular non-linear embedding method, used for visualizing high-dimensional data by projecting them to 2D (or 3D) while (stochastically) preserving their relationships. Recently, we found t-SNE superior to other embedding methods in the task of visualizing high-dimensional hidden-layer activations in trained deep networks [33].

We applied t-SNE on a set $\mathcal{S}_n$ to be projected into 2D space. In the last step we rendered collected outputs as scatter plots, where the position of a point in the chart was the projection of a high-dimensional state in 2D space and the value of the output for this state was depicted by color. Then we were able to create video from the sequence of charts.

V. Experiments

To appreciate the behavior of the proposed models, we tested them in three environments of different complexity, namely *Mountain car*, *Lunar lander* available in OpenAI Gym [34] and *Half cheetah* from PyBullet Gym. The well-known continuous environments have already encoded features in the state vector. The first two environments present a challenge for exploration, because the agent receives a negative reward according to the size of its action, and if it does not find a positive reward fast enough, the policy will converge into the agent's inactivity in the extreme case. We chose the last *Half-cheetah* environment to verify our hypothesis of increasing stability of the agent by motivation modules.

We divided our experiments into two parts. The first consists of testing the agent with meta-critic module in well-known continuous state and action space environments to compare the results with the baseline models. In the second part, we focused on a deeper analysis using imaging tool that can capture the dynamics during learning and better understand the complex processes that take place during this phase.

Model training setup

All our agents use DDPG [35] training algorithm. In making a decision, we were inspired by [36] who compared DDPG with TRPO [37] and PPO [38] algorithms. As an off-policy algorithm, DDPG appears more sample efficient than on-policy TRPO that has higher variance than off-policy algorithms and therefore needs more samples to converge to a good solution. DDPG can outperform TRPO and PPO in stable learning environments. TRPO, however, has more stable convergence properties, including near-monotonic improvement in practice (theoretical monotonic improvement guarantees), whereas DDPG learns an unstable policy. PPO outperforms A3C [39] on most continuous environments and often performs better than TRPO. NAF [40] learns a smooth, stable policy and it is more suitable for robotics where precision is required. It can also perform better than DDPG, especially in unstable environments.

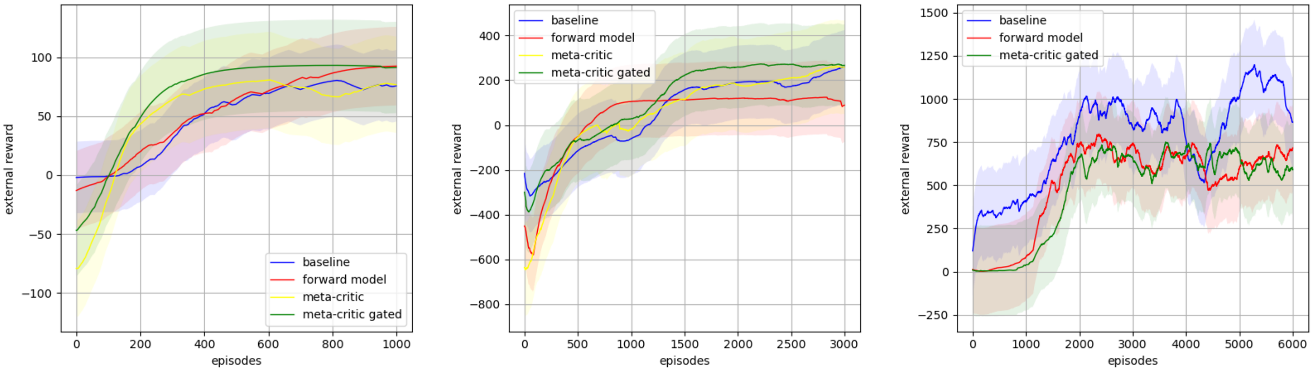
In our agents, the deterministic policy is approximated by an *actor* and Q-value function is approximated by a *critic*. All modules, i.e. actor, critic, forward model, and meta-critic are represented by two-layer neural networks. For parameter optimization of all modules we used Adam algorithm [41]. The learning rate of actor and critic was $\alpha_{\mathrm{act}} = 0.0001$ and $\alpha_{\mathrm{crit}} = 0.0002$ respectively in all environments. For forward model and meta-critic we chose the same values $\alpha_{\mathrm{fm}} = 0.0001, \alpha_{\mathrm{mc}} = 0.0002$ for mountain car and Lunar lander environments. Forward model and meta-critic in Half-cheetah environment were learnt from memory replay buffers and their learning rate was $\alpha_{\mathrm{fm}} = 0.001, \alpha_{\mathrm{mc}} = 0.002$. Exploration was performed by adding a noise to actor's output generated by the Ornstein–Uhlenbeck stochastic process in the case of mountain car environment and Gaussian noise in the other two environments. All environments had discount factor set to $\gamma = 0.99$ and in all our experiments we let $\epsilon_{fm} = \epsilon_{mc} = 1$. More hyper-parameters and further details of learning process can be found in our source codes.

For Mountain car environment we performed 10 training runs of each variant: baseline was without any motivation, forward model used prediction error motivation, meta-critic had only predictive surprise motivation and finally the gated meta-critic combined both predictive motivations. The results are shown in Fig. 3a. Each curve represents average cumulative external reward for each episode smoothed by exponential decay $\tilde{r}_t = 0.99\,\tilde{r}_{t-1} + 0.01\,r_t$ with $\tilde{r}_0 = 0$. The model with gated meta-critic was the best of all with steep ascent of an external reward between episodes 200 and 400 after which it stabilized. Average cumulative reward per episode is shown in Table II.

TABLE II: Average cumulative reward per episode in case of four agent types trained in three environments.

Environment	***Baseline***	***Forward model***	***Meta-critic***	***Meta-critic gated***
Mountain car	56.4 ± 16.6	64.4 ± 12.3	68.8 ± 11.5	**82.9** ± **6.2**
Lunar lander	70.5 ± 50.8	54.8 ± 63.8	79.5 ± 76.9	**130.8** ± **21.7**
Half cheetah	**782.6** ± **174.2**	523.6 ± 133.3	-	485.3 ± 178.3

For Lunar lander we performed 5 training runs and here the difference was not so significant (see Fig. 3b), where we can see that the model found optimal policy after 1500 episodes but then exhibited stable behavior compared to the baseline where several runs ended in a sub-optimal policy. The forward model and the meta-critic without gating performed similarly. We also performed 5 training runs of each agent in the Half-cheetah environment with during 6000 episodes. From the results (see Fig. 3c) it is obvious that the baseline agent outperformed both agents with motivation. It is partially



(a) Mountain car

(b) Lunar lander

(c) Half cheetah

Fig. 3: Simulation results on three different environments for each of the four agent types. Baseline agent is only driven by an external reward, the agent with a forward model uses the prediction error as an intrinsic motivation signal, the agent with a meta-critic makes use of predictive surprise, and the gated meta-critic agent combines both predictive signals into one intrinsic reward (you can find a brief overview in Table I). The agents using some kind of intrinsic motivation overcame the baseline agent in the first two environments (Mountain car and Lunar lander), and the most effective was the agent with a gated meta-critic module. In the third environment (Half cheetah) a baseline agent outperformed both agents with motivation, but did not reach stable policy. Both agents with motivation modules were more stable but the intrinsic signals prevented them from finding a closer-to-optimal policy.

caused by the fact that the environment has dense reward and motivation signals could introduce noise into the critic estimating the value function. But we also supposed that the agent with reduced number of parameters and motivation modules could reach more stable and optimal policy which is in contradictory with performance of both agents. The agent with motivation based on prediction error was able to stabilize it's policy but was unable to find optimal policy or just better policy than baseline agent. Our gated meta-critic agent ended with worse policy than baseline and it's stability was very similar as well.

Visual analysis of the gated meta-critic agent learning

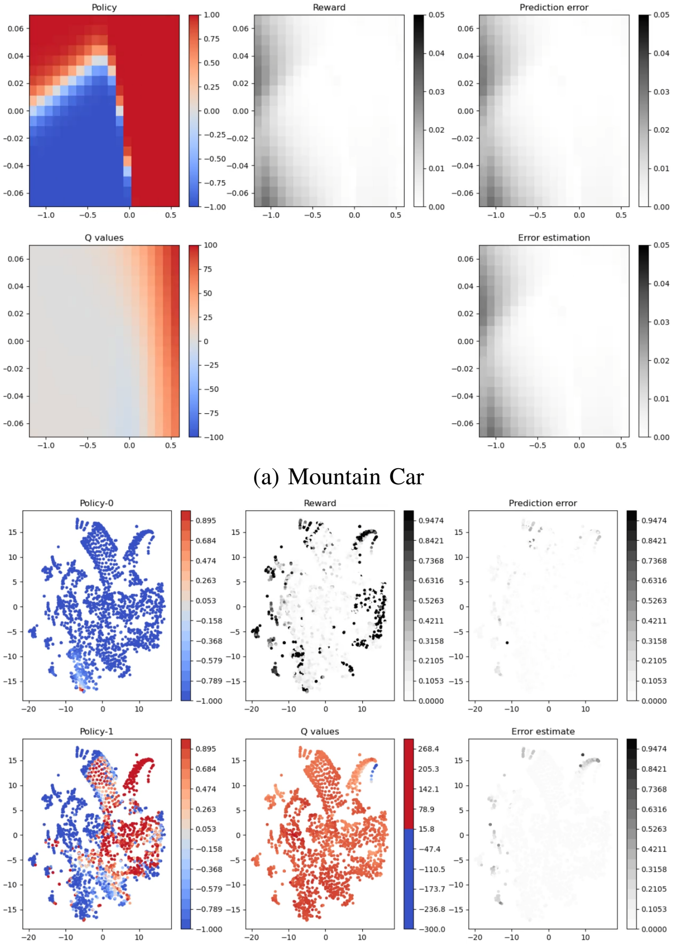
To better understand the behavior of our agents we propose a visualization of the training process step by step and using videos, created by our tool, providing a global picture over the process (snapshots are shown in Fig. 4). We focused on the dynamics of prediction error and its estimate which are the source of both intrinsic signals. This analysis revealed that there are several moments when the agent visits states where the meta-critic generates very large values of an intrinsic signal. It mainly occurs at the beginning of the training when the forward model and meta-critic are not synchronized well, but then it happens when the agent visits an unusual state which differs from the previous ones for the first time. Such large intrinsic rewards seemed to have positive effect and pushed the agent to further explore the revealed subspace of state space.

This process could be very dynamic and could lead to oscillations in the generated intrinsic signal, because some time is needed for adaptation of both predictive models. In particular, we observed such oscillatory behavior in Half-cheetah environment and partially in Lunar lander. It seems reasonable to conclude that increasing dimension of the state space leads to difficulties for the forward model to correctly estimate the next state. We also observed that the meta-critic started to underestimate the prediction error due to instabilities in the forward model estimations when the length of trajectories increased. It led to constant generation of an intrinsic reward which fed the critic with a noisy signal and degraded the agent's performance. Therefore, we decided to introduce the replay buffers to both models which slightly stabilized the predictive modules but did not eliminate the problem completely. It seems that after some steps the noise reached such levels that the agent was not able to modify its policy to reach a higher external reward because of saturation of the critic with an intrinsic reward and the agent got stuck. Here we see the room for further experiments and hyper-parameter tuning as well as thorough elaboration of a better forward model – meta-critic interaction.

We can consider each intrinsic motivation signal as a type of behavior, which competes for agent's "attention" what is very simply represented by the gating mechanism. The results of our analysis suggest that for an improvement of the learning speed (and the stability as well) there is a need for different behaviors in different states which can have a positive effect on the overall performance of the agent.

VI. Conclusion

We proposed an extension of the intrinsic motivation model based on prediction error, by introducing another predictor – meta-critic – that estimates an error of the forward model. With this qualitatively new information we created a new intrinsic signal which can be interpreted as predictive surprise. We performed tests of models with motivation based on predictive surprise and models combining prediction error motivation



(a) Mountain Car

(b) Lunar lander

Fig. 4: Snapshots from the visualisation of the gated meta-critic model. Each scatter plot shows the outputs of different modules at certain moment during learning and its axes represent the vehicle coordinates in 2D space, either original horizontal-vertical (Mountain car), or projected by t-SNE method (Lunar lander) from the original state space. This way we can see how each module responds to different states. In Lunar lander two policy plots (on the left) are needed because the action space is 2D. The more interesting part is the precision of the forward model predictions and the meta-critic estimations. Then we are able to analyze the dynamics of changes of an intrinsic signal based on the changes of outputs of both predictive modules and their influence on critic and actor, respectively.

and surprise by simple gating. With the last approach we obtained the best results for tested environments and this seems encouraging for our next research. We also described a visualization tool, usable also for higher-dimensional state spaces, for monitoring the training process which provided necessary insights into evolution of models and was very useful in subsequent analysis. We are able to construct intrinsic signals based on the outputs of our predictive modules which can refer to different types of behavior. Our next goal will be to elaborate the gating mechanism in a more detail and perform tests on challenging environments with very sparse reward and high-dimensional state space.

Acknowledgment

This research was supported by Slovak Grant Agency for Science, project VEGA 1/0796/18.

References

[1] R. S. Sutton and A. G. Barto, Introduction to Reinforcement Learning. MIT press Cambridge, 1998, vol. 135.

[2] G. Baldassarre, T. Stafford, M. Mirolli, P. Redgrave, R. M. Ryan, and A. Barto, “Intrinsic motivations and open-ended development in animals, humans, and robots: An overview,” Frontiers in Psychology, 2014.

[3] C. L. Hull, Principles of Behavior: An Introduction to Behavior Theory. New York, London: D. Appleton-Century Company, Inc., 1943.

[4] K. C. Montgomery, “The role of the exploratory drive in learning,” Journal of Comparative and Physiological Psychology, vol. 47, no. 1, pp. 60–64, 1954.

[5] H. F. Harlow, “Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys,” Journal of Comparative and Physiological Psychology, vol. 43, no. 4, pp. 289–294, 1950.

[6] R. W. White, “Motivation reconsidered: The concept of competence,” Psychological Review, vol. 66, no. 5, pp. 297–333, 1959.

[7] L. Festinger, A Theory of Cognitive Dissonance. Stanford university press, 1962.

[8] J. Hunt, “Intrinsic motivation and its role in psychological development,” in Nebraska symposium on motivation, vol. 13. University of Nebraska Press, 1965, pp. 189–282.

[9] M. Csikszentmihalyi, Flow: The Psychology of Optimal Experience. New York, NY: Harper Perennial, 1991.

[10] J. S. Watson, “Smiling, cooing, and ”the game”,” Merrill-Palmer Quarterly of Behavior and Development, vol. 18, no. 4, pp. 323–339, 1972.

[11] P. Rochat and T. Striano, “Perceived self in infancy,” Infant Behavior and Development, vol. 23, no. 3-4, pp. 513–530, 2000.

[12] P.-Y. Oudeyer and F. Kaplan, “What is intrinsic motivation? a typology of computational approaches,” Frontiers in Neurorobotics, vol. 1, p. 6, 2009.

[13] B. C. Stadie, S. Levine, and P. Abbeel, “Incentivizing exploration in reinforcement learning with deep predictive models,” 2015, arXiv:1507.00814.

[14] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, “The arcade learning environment: An evaluation platform for general agents,” Journal of Artificial Intelligence Research, vol. 47, pp. 253–279, 2013.

[15] J. Schmidhuber, “Formal theory of creativity, fun, and intrinsic motivation (1990–2010),” IEEE Transactions on Autonomous Mental Development, vol. 2, no. 3, pp. 230–247, 2010.

[16] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, “Curiosity-driven exploration by self-supervised prediction,” CoRR, vol. abs/1705.05363, 2017.

[17] M. Wydmuch, M. Kempka, and W. Jaśkowski, “Vizdoom competitions: Playing doom from pixels,” IEEE Transactions on Games, 2018.

[18] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2013, arXiv:1312.6114.

[19] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, “Large-scale study of curiosity-driven learning,” arXiv preprint arXiv:1808.04355, 2018.

[20] H. Kim, J. Kim, Y. Jeong, S. Levine, and H. O. Song, “Emi: Exploration with mutual information,” 2018, arXiv:1810.01176.

[21] H. Tang, R. Houthooft, D. Foote, A. Stooke, O. X. Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel, “# exploration: A study of count-based exploration for deep reinforcement learning,” in Advances in neural information processing systems, 2017, pp. 2753–2762.

[22] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos, “Count-based exploration with neural density models,” in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 2721–2730.

[23] J. Martin, S. N. Sasikumar, T. Everitt, and M. Hutter, “Count-based exploration in feature space for reinforcement learning,” 2017, arXiv:1706.08090.

[24] M. C. Machado, M. G. Bellemare, and M. Bowling, “Count-based exploration with the successor representation,” 2018, arXiv:1807.11622.

[25] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," 2018, arXiv:1810.12894.

[26] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, "Vime: Variational information maximizing exploration," in Advances in Neural Information Processing Systems, 2016, pp. 1109–1117.

[27] P. Shyam, W. Jaśkowski, and F. Gomez, "Model-based active exploration," in International Conference on Machine Learning, 2019, pp. 5779–5788.

[28] R. Zhao, V. Tresp, and W. Xu, "Mutual information-based state-control for intrinsically motivated reinforcement learning," 2020, arXiv:2002.01963.

[29] G. Montúfar, K. Ghazi-Zahedi, and N. Ay, "Information theoretically aided reinforcement learning for embodied agents," 2016, arXiv:1605.09735.

[30] T. Jung, D. Polani, and P. Stone, "Empowerment for continuous agent-environment systems," Adaptive Behavior, vol. 19, pp. 16–39, 2011.

[31] R. Zhao, P. Abbeel, and S. Tiomkin, "Efficient online estimation of empowerment for reinforcement learning," 2020, arXiv:2007.07356.

[32] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.

[33] T. Kuzma and I. Farkaš, "Embedding complexity of learned representations in neural networks," in Proceedings of the 28th International Conference on Artificial Neural Networks (ICANN), vol. 2, 2019, pp. 518–528.

[34] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016, arXiv:1606.01540.

[35] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.

[36] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[37] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in International Conference on Machine Learning, 2015, pp. 1889–1897.

[38] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[39] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in International Conference on Machine Learning, 2016, pp. 1928–1937.

[40] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous deep q-learning with model-based acceleration," in International Conference on Machine Learning, 2016, pp. 2829–2838.

[41] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations, 2015.