Automated Curriculum Learning by Rewarding Temporally Rare Events

Niels Justesen
IT University of Copenhagen
Copenhagen, Denmark
noju@itu.dk

Sebastian Risi
IT University of Copenhagen
Copenhagen, Denmark
sebr@itu.dk

***Abstract*—Reward shaping allows reinforcement learning (RL) agents to accelerate learning by receiving additional reward signals. However, these signals can be difficult to design manually, especially for complex RL tasks. We propose a simple and general approach that determines the reward of pre-defined events by their rarity alone. Here events become less rewarding as they are experienced more often, which encourages the agent to continually explore new types of events as it learns. The adaptiveness of this reward function results in a form of automated curriculum learning that does not have to be specified by the experimenter. We demonstrate that this *Rarity of Events* (RoE) approach enables the agent to succeed in challenging VizDoom scenarios without access to the extrinsic reward from the environment. Furthermore, the results demonstrate that RoE learns a more versatile policy that adapts well to critical changes in the environment. Rewarding events based on their rarity could help in many unsolved RL environments that are characterized by sparse extrinsic rewards but a plethora of known event types.**

I. Introduction

Deep reinforcement learning and deep neuroevolution have achieved impressive results learning to play video games [17] and controlling both simulated and physical robots [2, 9, 15, 27]. These approaches, however, struggle to learn in environments where feedback signals (also called rewards) are sparse and/or delayed. A popular way to overcome this issue is to shape the reward function with prior knowledge such that the agent receives additional rewards to guide its learning process [22, 23, 31]. Another approach is to gradually increase the difficulty of the environment to ease learning through curriculum learning [4, 46]. Both approaches are time-consuming, require substantial domain knowledge and are especially difficult to implement for complex environments. In this paper, we propose a simple method that automatically shapes the reward function during training and performs a form of curriculum learning that adapts to the agent's current performance. The only required domain knowledge is the specification of a set of positive events that can happen in the environment (e.g. picking up items, moving, winning etc.), which is easy to implement if raw state changes are accessible.

The method introduced in this paper rewards a reinforcement learning (RL) agent by the rarity of experienced events such that rare events have a higher value than frequent events. The idea is to completely discard the extrinsic reward and instead motivate the agent intrinsically toward a behavior that explores the pre-defined events. As the agent first experiences certain types of events that are relatively easy to learn (e.g. moving around and picking up items) they will slowly become less rewarding, pushing the agent to explore rare and potentially more difficult events. Thus by only rewarding events for their rarity, the system performs a form of automated curriculum learning.

The goal of this approach is to learn through a process of *curiosity* rather than optimizing toward a difficult pre-defined goal. We apply our method, called *Rarity of Events* (RoE), to learn agent behaviors from raw pixels in the VizDoom framework [19]. While our approach could be applied to any reward-based learning method and possibly also fitness-based evolutionary methods, in this paper we train deep convolutional networks through the actor-critic algorithm A2C [29]. In the future, RoE could offer a new way to learn versatile behaviors in increasingly complex environments such as StarCraft, which is a yet unsolved reinforcement learning problem [44].

The paper is structured as follows. We first review relevant previous work, including related approaches in Section II. After explaining RoE (Section III), we demonstrate the usefulness of the method on five challenging VizDoom scenarios with sparse rewards and show how RoE learns a versatile behavior that can adapt to critical changes in the environment (Section V).

II. Previous Work

A. Deep Reinforcement Learning

Deep reinforcement learning allows learning agent behaviors in video games directly from screen pixels, including Atari games [28], first-person shooters [19, 22, 46], and car racing games [29]. These methods are typically variants of Deep Q Networks (DQN) [28] or actor-critic methods with parallel actor-learners such as Asynchronous Advantage Actor-Critic (A3C) [29]. Neuroevolution [13, 36] has also recently shown promising results in playing Atari games and can be easier to parallelize [39, 43].

A key requirement for deep RL methods to work out of the box are frequent and easy obtainable reward signals from the environment that can guide learning toward an optimal behavior. An infamous Atari game where this is not the case

is Montezuma's Revenge; for this game with very sparse rewards, both DQN and A3C variants fail [28, 29].

The lack of frequent reward signals can be overcome by reward shaping, where a smoother reward function is designed using prior domain knowledge [22, 23], or by gradually increasing the difficulty of the environment (e.g. the level itself or the NPCs' behaviors) to ease learning through curriculum learning [4, 46]. Related to curriculum learning is a method called *Power Play* that searches for new unsolvable problems while the agent is trained to progressively match the difficulty of the environment [41]. Another related approach is hierarchical reinforcement learning where a meta-controller controls one or more sub-policies that are trained to reach sub-goals (equivalent to events) [7, 20].

B. Curiosity & Intrinsic Motivation

In curiosity-driven learning the agent seeks to explore new situations guided by intrinsic motivation [18, 32, 38]. One theory of intrinsic motivation is *reduction of cognitive dissonance*, i.e. the motivation to learn a cognitive model that can explain and predict sensory input [12, 33]. This theory has also been formalized in the context of RL in which agents are intrinsically rewarded when observing temporarily novel, interesting, or surprising patterns based on their own world model [40]. A related idea is *optimal incongruity*, where discrepancies between the currently perceived and what is usually perceived produce a high stimulus; thus novel situations that yet lie within our current understanding are highly rewarding [5, 16]. The prediction error of a learning model can thus be used directly to define the reward function [14].

One way of implementing intrinsic motivation is to model the expected learning progress $\zeta(s, a)$ of a state-action pair [25]. The Intrinsic Curiosity Module (ICM) is another approach that encodes states s_t and s_{t+1} into features $\Phi(s_t)$ and $\Phi(s_{t+1})$ and determines the intrinsic reward based on the prediction error of these features and the forward model's features [34]. State-density models that assign probabilities to screen images, can be learned together with a policy and then determine intrinsic motivation as the model's temporal change in prediction, such that surprising screen images produce higher rewards [3].

Rewarding RL agents based on the novelty of events has been explored earlier with tabular Q-learning in a simple 3D environment [26], where the reward is highest when novelty is moderate. A combination of habituation theory and self-organizing maps was employed to vary the agent's curiosity (the reward signal toward certain events).

C. Novelty Search

The pursuit of novel situations also shares some similarities with novelty search [24] in evolutionary computation. The idea of novelty search is to search for novel behaviors instead of optimizing toward a specific objective directly. Both novelty search and our approach RoE push the search toward unexplored areas; however, novelty search does so for a population of individuals where novelty is defined as the behavioral distance to other behaviors in the population. Our approach is trained through reinforcement learning and novelty (or rather rarity of events) is based on experiences of previous versions of the policy.

D. VizDoom

The approach in this paper is tested in VizDoom, an AI research platform based on the commercial video game Doom that allows learning from raw visual information [19]. The VizDoom framework includes several diverse environments, some of which are very challenging to learn due to their sparse and delayed rewards. Several deep RL approaches have been applied to Doom, which include auxiliary learning [21, 22], game-feature augmentation [6, 8, 11], manual reward shaping [8, 11, 22], and curriculum learning [46]. A very different approach by Alvernaz and Togelius applies neuroevolution on top of a pre-trained auto-encoder [1]. In this paper, we purposefully build on a vanilla implementation of the RL algorithm A2C, to set a baseline for how well RoE can help in challenging VizDoom scenarios.

III. Approach

This section describes our *Rarity of Events* (RoE) approach and its integration with A2C in VizDoom.

A. Rewarding Temporally Rare Events

The reward function in RoE adapts throughout training to the policy's ability to explore the environment. By rewarding events based on how often they occur during training, the agent is intrinsically motivated toward exploring new parts of the environment rather than aiming for a single goal that might be difficult to obtain directly. In effect, the approach performs a form of curriculum learning since events are rewarded based on the agent's current ability to obtain them. As the agent learns, it becomes less interested in events that are frequent and *curious* about newly discovered events.

Our method requires a set of pre-defined events, and the reward $R_t(\epsilon_i)$ for experiencing one of these events ϵ_i at time t is determined by its temporal rarity $\frac{1}{\mu_t(\epsilon_i)}$, where $\mu_t(\epsilon_i)$ is the temporal episodic mean occurrence of ϵ_i at time t, i.e. how often ϵ_i occurs per episode at the moment. The mean occurrences of events are clipped to be above a lower threshold τ (we used 0.01 such that the maximum reward for any event is 100). For a vector of event occurrences x, such that x_i is the number of times ϵ_i occurred in a game step, the reward is the sum of all event rewards:

$$R_t(x) = \sum_{i=1}^{|x|} x_i \frac{1}{\max(\mu_t(\epsilon_i), \tau)}. \quad (1)$$

The rarity measure $\frac{1}{\mu_t(\epsilon_i)}$ is not arbitrary but is designed such that all events have equal importance. If any event ϵ_i is experienced n times during an episode, and $n = \mu_t(\epsilon_i)$ (which is the expected amount), then the accumulated reward for ϵ_i is 1 regardless of the rarity. This means that in theory all events have equal importance. In practice, the policy might learn that some events have a negative or positive influence on the occurrence of others.

B. Determining the Temporal Episodic Mean Occurrence

There are arguably many ways to determine the temporal episodic mean occurrence $\mu_t(\epsilon_i)$; here we employ a simple approach that nevertheless achieves the desired outcome. Whenever an episode during training reaches a terminal state, a vector ϵ containing the occurrence of events in this episode is added to a buffer of size N. The size of the buffer determines the adaptability of the reward function. If N is small, the agent quickly becomes *bored* of new events as it easily forgets their rarity in the past. If N is large, the agent will stay *curious* for a longer period of time. The temporal episodic mean occurrence $\mu_t(\epsilon)$ is then determined as the mean of all records in the buffer, i.e. the episodic mean of the last N episodes.

C. Events in Doom

We track 26 event types in VizDoom by implementing a function that determines which events occur in every state transition (i.e. in each time step). The event types include movement (one unit), shooting (decrease in ammo), picking up an item (one event for each item type; health pack, armor, ammo, and weapons 0–9), killing (one for each weapon type 0–9 as well as one regardless of weapon type). Movement events are triggered when the agent has traveled one unit from the position of the last movement event (or the initial position if the agent has not yet moved).

D. Policy

The presented reward shaping approach can be applied to most (if not all) RL methods that learn from a reward signal. It could potentially also be applied to evolutionary approaches such as Evolution Strategies by defining fitness as the sum of rewards in an episode. A standard policy network is employed that has three convolutional layers followed by a fully connected layer of 512 units, and a policy and value output. We use filter sizes of [32, 64, 32] with strides [4, 2, 1], ReLU activations for hidden layers, and softmax for the policy output.

The input is a single frame of 160×120 pixels in grayscale, cropped by removing 10 pixels on top/bottom and 30 pixels on the sides and then resized to 80×80. In most of the scenarios, the agent can perform four actions: attack, move forward, turn left, and turn right. In this case, the policy output has $2^4 = 16$ values to allow any combination of the four actions. The event buffer is updated whenever a worker reaches a terminal state. The rewards from VizDoom, which vary between -100 and 100, are normalized to $[0, 1]$. Rewards based on our approach are not normalized and vary between $[0, 100]$ (due to $\tau = 0.01$), while for all events where $\mu_t(\epsilon_i) \geq 1$ the reward will be between 0 and 1 (following Equation 1 in Section III-A).

E. Advantage Actor-Critic (A2C)

The deep networks in this paper are trained with the deep reinforcement learning algorithm A2C, a synchronous variant of Asynchronous Advantage Actor-Critic (A3C) [29], which is able to reach state-of-the-art performance in a wide range of environments [42, 45, 47].

A2C is an actor-critic method that optimizes both a policy π (the actor) and an estimation of the state-value function $V(s)$ (the critic). Parallel worker threads share the same model parameters and synchronously collect trajectories $(s_t, s_{t+1}, a_t, r_{t+1})$ for t_{max} game steps where after the model's parameters are updated. Threads restart new episodes individually when they are done. The discounted return $R_t = \sum_{i=1}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k})$, where k is the number of trajectories collected after t, and the advantage $A(s_t, a_t) = R_t - V(s_t)$ is determined for each step, for every worker. A2C then uses the traditional A3C update rules in [29] based on the policy loss $\log \pi(a_i|s_i)A(s_i)$ and value loss; the mean squared error between the experienced R_t and the predicted $V(s_t)$: $\frac{1}{2}(R_t - V(s_t))^2$. In contrast to A3C, A2C updates the parameters synchronously in batches.

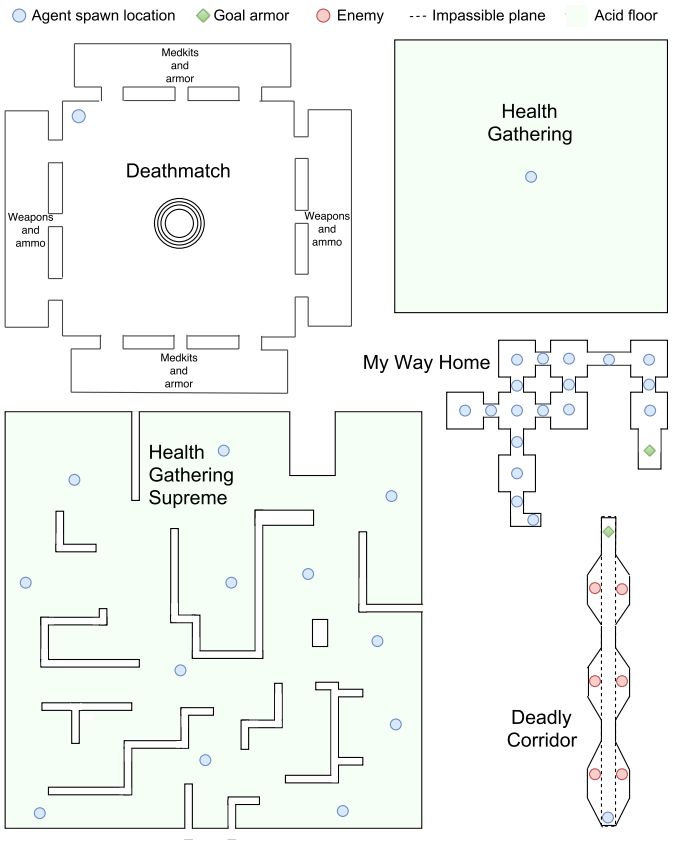

Fig. 1: The five ViZdoom scenarios. Scenarios with multiple spawning positions randomly select one of them at the start of an episode. The episode ends when the goal armor, which only appears in *My Way Home* and *Deadly Corridor*, is picked up. The agent periodically looses health when standing on acid floors.

IV. VizDoom Testing Scenarios

This section describes the five VizDoom scenarios used in our experiments. They all have sparse and/or delayed rewards and are therefore a good test domain for our approach. The scenarios are from the original VizDoom [19] repository[1].

[1]https://github.com/mwydmuch/ViZDoom/tree/master/scenarios

For each scenario we also detail the extrinsic reward from the environment, which is used when training models without RoE. Some of these extrinsic rewards were rescaled to be coherent across scenarios. If not stated otherwise, the agent can move forward, turn left, turn right, and shoot. Screenshots from these scenarios are shown in Figure 2, with top-down views in Figure 1.

1) Health Gathering: The goal is to survive as long as possible in a square room with an acid floor that deals damage periodically. Medkits spawn randomly in the room and can help the agent to survive as they heal when picked up. The agent is rewarded 1 for every time step it is alive, and -100 for dying. The maximum episode length is 2,100 time steps. The agent cannot shoot.

2) Health Gathering Supreme: Same as *Health Gathering* but within a maze.

3) My Way Home: The goal is to pick up an armor, which gives a reward of 100 and ends the scenario immediately. The agent cannot shoot and is rewarded -0.1 for every time step it is alive. The agent starts an episode at one of the randomly chosen spawn locations with a random rotation.

4) Deadly Corridor: Similarly to *My Way Home*, the goal is to pick up an armor, which gives a reward of 100 and ends the scenario immediately. The armor is located at the end of a corridor, which is guarded by enemies on both sides. The agent must kill most, if not all of the enemies to reach it, and receives a -100 reward if it dies. The original reward shaping function (the distance to the armor) has been removed to make it harder and to compare RoE with a baselines that does not use any reward shaping. The maximum episode length is 2,100 time steps.

Fig. 2: From top-left to bottom-right: Screenshot from *Deathmatch*, *My Way Home*, *Health Gathering Supreme*, and *Deadly Corridor*. Notice that in some scenarios the agent cannot shoot. The scenario *Health Gathering* is similar to *Health Gathering Supreme* but without walls within the room.

5) Deathmatch: The agent spawns in a large battle arena with an open area in the middle and four rooms, one in each direction that contain either medkits and armor, or weapons (chainsaw, super shotgun, chaingun, rocket launcher, and plasma gun) and ammunition for each weapon. The maximum episode length is 4,200 time steps. The agent is rewarded the following amounts when killing an enemy: Zombieman (100), ShotgunGuy (300), MarineChainsawVzd (300), Demon (300), ChaingunGuy (400), HellKnight (1,000). These enemies spawn randomly on the map when the scenario starts.

A2C	
Learning rate	7e-4
γ (discount factor)	0.99
Entropy coefficient	0.01
Value loss coefficient	0.5
Learning rate	0.0007
Max. gradient-norm	0.5
Worker threads	4 (16 in DM)
t_{max} (Steps per. update)	20
Batch size	64
Frame skip	4
RMSprop Optimizer	
ϵ	1e-5
α	0.99
RoE	
N (event buffer size)	100
τ (mean threshold)	0.01

TABLE I: Experimental configurations for A2C and A2C+RoE. 16 worker threads were used in *Deathmatch*.

To test how well the approach can adapt to new scenarios, five variations of *Deathmatch* were also created that only include a certain weapon type. These scenarios are called *Deathmatch Chainsaw*, *Deathmatch Chaingun*, *Deathmatch Shotgun*, *Deathmatch Plasma*, and *Deathmatch Rocket* to denote which weapon that remains on the map. The ammunition for the other weapons was also removed.

V. Results

We tested A2C with our approach *Rarity of Events* (A2C+RoE) on the five VizDoom scenarios described in Section IV. The *Deathmatch* variations were not used for training. As a comparison baseline, A2C was also trained using the extrinsic reward from the environment as described in Section IV. Due to computational constraints we only trained each method once on each scenario.

When training with A2C+RoE, the agent did not have access to the extrinsic reward throughout training but only the intrinsic reward based on the temporal rarity of the pre-defined events. The algorithms ran for 10^7 time steps for each scenario and 7.5×10^7 for the Deathmatch scenario. For both A2C and A2C+RoE we save a copy of the model parameters whenever the mean extrinsic reward across all workers improves. The last copy is considered to be the final model that we use in our tests. The complete configurations for A2C and A2C+RoE are shown in Table I and the code for the experiments and trained models are available on GitHub[2]. Videos of the learned policies are available on YouTube[3].

[2]https://github.com/njustesen/rarity-of-events
[3]https://youtu.be/YG-lf732a0U

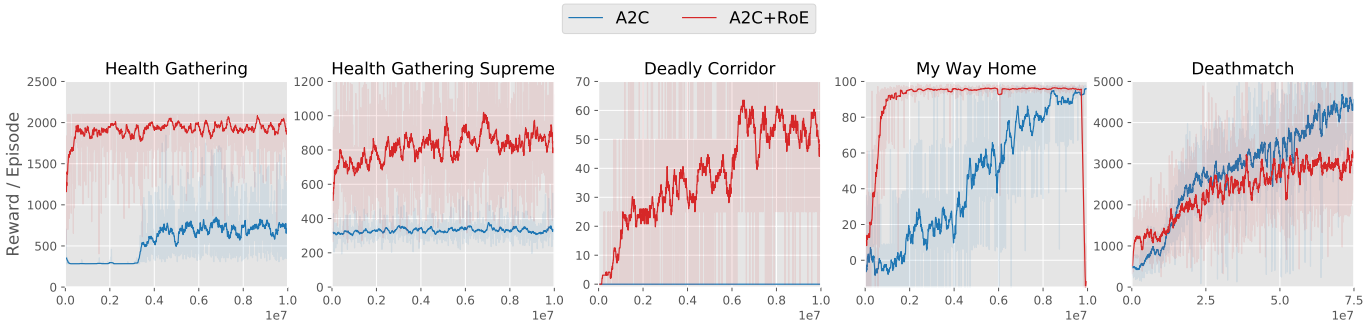

Fig. 3: The reward per episode of A2C and A2C+RoE during training in five VizDoom scenarios (smoothed). A2C is trained from the environment's extrinsic reward while A2C+RoE uses our proposed method without access to the reward. The drop in performance seen in the My Way Home scenario is discussed in-depth in Section V-A.

Fig. 4: Episodic mean occurrence during training for a subset of the event types in the five VizDoom scenarios. Notice the last spike in the *My Way Home* scenario with A2C+RoE, in which the policy ignores the final goal (armor pickup) to prioritize continuous movement around the maze.

Scenario	A2C	A2C+RoE	t-test
Health Gathering	399 ± 107	**1261** ± 533	$p < 0.0001$
Health Gathering Supr.	305 ± 60	**1427** ± 645	$p < 0.0001$
Deadly Corridor	0.00 ± 0.0	**40** ± 49	$p < 0.0001$
My Way Home	96.69 ± 0.12	**97.89** ± 0.01	$p < 0.0001$
Deathmatch	**4611** ± 2595	4062 ± 2442	$p = 0.1250$
Deathmatch Chainsaw	1025 ± 809	**3750** ± 3130	$p < 0.0001$
Deathmatch Chaingun	1487 ± 1189	**2852** ± 2038	$p < 0.0001$
Deathmatch Shotgun	1375 ± 941	**1832** ± 1752	$p = 0.0226$
Deathmatch Plasma	**4538** ± 1537	3248 ± 2701	$p < 0.0001$
Deathmatch Rocket	616 ± 583	**1463** ± 1449	$p < 0.0001$

TABLE II: Shown are average scores based on evaluating the best policies found for A2C and A2C+RoE 100 times each. The best results are shown in bold. The five last rows show how the policies that were trained on the original *Deathmatch* scenario generalize to five variations where only one weapon type is available. Standard deviations are shown for each experiment and two-tailed p-values from unpaired t-tests.

A. Learned Policies

The A2C baseline did not learn a good policy in *Health Gathering Supreme* and *Deadly Corridor*, and only improved slightly in *Health Gathering* (Figure 3). A2C learned a weak policy in three out of five scenarios, which demonstrates that they are indeed difficult to master guided by the extrinsic rewards alone. In *My Way Home*, A2C does learn a strong behavior that consistently locates and picks up the armor but only after 8–9 million training steps. In *Deathmatch*, A2C learned a very high-performing behavior that directly walks to the plasma gun (the most powerful weapon in this scenario) and shoots from cover toward the center of the map. The behavior is simple but effective until it runs out of ammunition, after which it attempts to find more ammunition and sometimes fails.

Our approach A2C+RoE learns effective behaviors in all five scenarios. The learned behavior in *Deathmatch* does not exclusively use the powerful plasma gun, which results in a slightly but not significantly worse performance than A2C ($p = 0.125$ using two-tailed t-test). The policy is still effective with over 10 kills per episode. These kills are spread across all weapons that are available, resulting in a behavior that is more varied (and interesting to watch). As we will show in Section V-B, the versatile behavior learned by A2C+RoE allows it to adapt to critical changes in *Deathmatch* in contrast

to policies trained through A2C.

The episodic mean occurrence of events (Figure 4) allows us to analyze how the policies change over time. In *Health Gathering* and *Health Gathering Supreme*, A2C+RoE quickly learns to move $\sim$80 and $\sim$30 units per episode, respectively. This behavior might explain why the agent also quickly learns to pick up medkits. A2C, on the other hand, learns the relationship between movement, medkits, and survival at a much slower pace, at least in the *Health Gathering* scenario. In *Deadly Corridor* A2C+RoE discovers an interesting behavior. After the agent learns to kill all six enemies (the red line) and to pick up armor (purple line), it still manages to increase the movement and the shooting events; the agent learned to walk back to its initial position while shooting and then afterwards to return to pick up the armor. This result is not unexpected as the agent is intrinsically motivated to experience as many events as possible during an episode.

In *My Way Home*, after the A2C+RoE policy has learned to routinely pick up the armor, it shifts into a different behavior toward the end of training. The agent learned to avoid the armor to instead continuously move around in the maze. We suspect that the policy would shift back to the previous behavior if training was continued, as the movement reward is now decreasing and the armor reward is increasing. Since our rarity measure is temporal, loops between these two behaviors could emerge as well. As policies with the highest extrinsic reward are saved during training, these sudden changes do not affect the final policy. In fact, one might argue that this is a useful feature of RoE: a network that has converged to some optimum can escape it to find other interesting behaviors.

B. Ability to Adapt

A2C+RoE motivates the agent intrinsically to learn a balanced policy that strives to experience a good mix of events. Reinforcement learning algorithms that exclude pre-training or proper reward shaping, including our A2C baseline, can easily converge into local optima with very *narrow* behaviors. In this context, *narrow* refers to behaviors that act in a very particular way, only utilizing a small subset of the features in the environment. This handicap prevents the learned policies from adapting to critical changes in the environment as they only know one way of behaving.

To test for such adaptivity, the learned policies are evaluated on five *Deathmatch* variations in which critical weapons and ammunition packs have been removed. Note that the policies were not directly trained on these variations. The results in Table II show that A2C+RoE learned a policy that significantly outperforms A2C ($p < 0.0001$ using two-tailed t-test) in four out of five *Deathmatch* variations. A2C+RoE learned a policy that is more versatile, capable of using all the weapons in the map, which is the reason it can easily adapt. Figure 5 shows heat maps (i.e. the proportional time spent at each map location) during the evaluations of the two policies on *Deathmatch* and its variations. The A2C+RoE policy expresses different strategies depending on the weapon available on the map, while the A2C policy mostly circles around the plasma gun location, regardless of it actually being there. However, if the plasma gun is present, A2C alone does execute a fairly effective strategy, shooting toward enemies in the middle of the map.

The heat maps show that the A2C policy has learned to stay at only one location on the map from which it can pick up the powerful plasma gun and thereafter shoot efficiently toward enemies in the middle of the map (see the video of the learned policies). In the *Deathmatch* variations, in which the map only contains two weapons of the same type, the A2C-policy fails to adapt to use the other weapons and instead walks around the area where the plasma gun would have been located.

The A2C+RoE policy has learned to explore a larger part of the maps in a more uniform way (Figure 5,bottom). In the different *Deathmatch* variations, a clear change in behavior can be observed when only a certain type of weapon is available. For example, in the *DM Rocket* scenario, the agent lures enemies into the map's top and bottom room while efficiently using the rocket's splash damage.

VI. Discussion

While the presented approach worked well in VizDoom it will be important to test its generality in other domains in the future. RoE is designed to work well in challenging environments that have a plethora of known events and sparse and/or delayed rewards. Video games are thus a very suitable domain and we plan to test RoE in Montezuma's Revenge and StarCraft in future work. For domains in which reward shaping is not necessary, i.e. the extrinsic reward smoothly leads to an optimal behavior, our approach might add less value. We imagine that RoE should also work well in domains with deceptive reward structures, just as novelty search outperformed traditional evolutionary algorithms in mazes with dead ends [24] or deceptive meta-learning tasks [37]. Novelty search and RoE have the ability to learn interesting behaviors without the need for a goal. In the future, our approach could also be extended to reward the agent for both the rarity of events as well as the environment's original objective, inspired by *quality diversity methods* [35] that use a combination of diversity and objective-based search [10, 30].

The specification of adequate events is intimately tied to the success of our approach; events that lead to direct negative performance should be avoided. For example, if the extrinsic reward is negative when the agent wastes ammunition, it should not be intrinsically rewarded for shooting event. A benefit of the presented method is that events that contribute to the occurrence of other events (e.g. such as movement leads to medkit pickups), can lead to a system that performs automated curriculum learning. However, it is not guaranteed that this effect will occur, and it might require a bit of trial and error during the specification of events. Some events can also be contradicting, such as killing with the chainsaw and killing with the plasma gun, as the agent cannot do both at the same time. Our approach is designed to learn a policy that can balance their occurrences which results in a more versatile

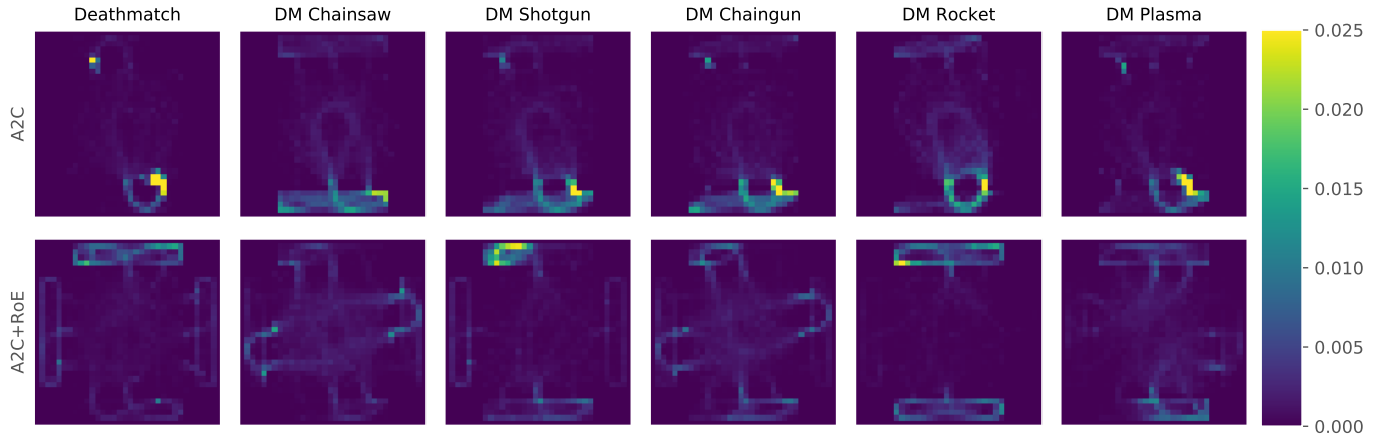

Fig. 5: Heat maps showing the proportional time spent at each location on the map in the *Deathmatch* scenario and its five variations. The values are based on evaluating the two trained policies 100 times each and clipped at 0.025. The heat maps show that the A2C-policy prefers to stay near the plasma gun, even in the map variations where it is not present, while the A2C+RoE-policy has learned distinct behaviors for each weapon type. The results in Table II shows that the A2C+RoE-policy is able to reach high scores in these variations event though it was never trained on them.

behavior. Important future work will test how RoE scales to hundreds or even thousands of events. A promising testbed for such experimentation is StarCraft, for which events can easily be defined as the production of each unit and building type, as well as killing different opposing unit types. We believe that reinforcement learning methods that are guided by intrinsic motivation are key to solving these challenging environments.

The A2C baseline reached the best performance in the original *Deathmatch*. However, it can be argued whether it learned to actually play Doom, or just learned to follow a fixed sequence of actions that lead to the same behavior every time. While it can be useful to find a niche behavior with high performance, learning a rich and versatile behavior has particular relevance for video games. Here, behaviors that explore the game's features could potentially help for automatic game testing and also lead to more human-like behaviors for NPCs.

Regarding our implementation of the RoE approach, future work will also explore other variations in determining the episodic mean occurrence of events, such as discounting the mean occurrences over time. With this modification, event occurrences older than N episodes (the event buffer only holds N event occurrences) would still effect the intrinsic reward.

It is important to note that since we save the best model based on the mean extrinsic reward across all worker threads, increasing the number of threads should make the evaluation less noisy by reducing the chances of accidentally overriding the best model with a worse performing one. This hypothesis still needs to be confirmed, but the number of threads was already increased from 4 to 16 in the longer *Deathmatch* scenario to speed up learning.

VII. Conclusion

We introduced *Rarity of Events* (RoE), a simple reinforcement learning approach that determines reward based on the temporal rarity of pre-defined events. This approach was able to reach high-performing scores in five challenging VizDoom scenarios with sparse and/or delayed rewards. Compared to a traditional A2C baseline, the results are significantly better in four of the five scenarios. Importantly, the presented approach is able to not only receive a high final reward, but also discovers versatile behavior that can adapt to critical changes in the environment, which is challenging for the baseline A2C approach. In our experiments, the extrinsically motivated baseline either fails in these environments or learns a behavior that is unable to adapt to changes in the environments it has been trained on. In the future, the presented RoE approach could allow more complex scenarios to be solved, for which it is infeasible to learn from extrinsic rewards without manual reward shaping and curriculum learning.

VIII. Acknowledgements

We thank OpenAI for publishing accessible implementations of A2C and GitHub user p-kar for the integration of A2C to VizDoom[4]. We would also like to show our gratitude to the members of the Game Innovation Lab at New York University Tandon School of Engineering for their feedback and inspiring ideas. This work was financially supported by the Elite Research travel grant from The Danish Ministry for Higher Education and Science.

References

[1] S. Alvernaz and J. Togelius. Autoencoder-augmented neuroevolution for visual doom playing. In *Computational Intelligence and Games (CIG), 2017 IEEE Conference on*, pages 1–8. IEEE, 2017.

[2] M. Andrychowicz, D. Crow, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5055–5065, 2017.

[3] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.

[4] https://github.com/p-kar/a2c-acktr-vizdoom

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009.

[5] D. E. Berlyne. Conflict, arousal, and curiosity. 1960.

[6] S. Bhatti, A. Desmaison, O. Miksik, N. Nardelli, N. Siddharth, and P. H. Torr. Playing doom with slam-augmented deep reinforcement learning. *arXiv preprint arXiv:1612.00380*, 2016.

[7] M. M. Botvinick, Y. Niv, and A. C. Barto. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.

[8] D. S. Chaplot and G. Lample. Arnold: An autonomous agent to play fps games. In *AAAI*, pages 5085–5086, 2017.

[9] Y. Chebotar, K. Hausman, M. Zhang, G. Sukhatme, S. Schaal, and S. Levine. Combining model-based and model-free updates for trajectory-centric reinforcement learning. *arXiv preprint arXiv:1703.03078*, 2017.

[10] G. Cuccu and F. Gomez. When novelty is not enough. In *European Conference on the Applications of Evolutionary Computation*, pages 234–243. Springer, 2011.

[11] A. Dosovitskiy and V. Koltun. Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*, 2016.

[12] L. Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957. ISBN 9780804709118.

[13] D. Floreano, P. Dürr, and C. Mattiussi. Neuroevolution: from architectures to learning. *Evolutionary Intelligence*, 1(1):47–62, 2008.

[14] G. Gordon and E. Ahissar. Hierarchical curiosity loops and active sensing. *Neural Networks*, 32:119–129, 2012.

[15] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine. Continuous deep Q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.

[16] J. M. HUNT. Intrinsic motivation and its role in psychological development. *Nebraska symposium on motivation*, 13:189–282, 1965. URL https://ci.nii.ac.jp/naid/20001159493/en/.

[17] N. Justesen, P. Bontrager, J. Togelius, and S. Risi. Deep learning for video game playing. *arXiv preprint arXiv:1708.07902*, 2017.

[18] F. Kaplan and P.-Y. Oudeyer. Intrinsically motivated machines. In *50 years of artificial intelligence*, pages 303–314. Springer, 2007.

[19] M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jaśkowski. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*, pages 1–8. IEEE, 2016.

[20] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pages 3675–3683, 2016.

[21] T. D. Kulkarni, A. Saeedi, S. Gautam, and S. J. Gershman. Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*, 2016.

[22] G. Lample and D. S. Chaplot. Playing FPS games with deep reinforcement learning. In *AAAI*, pages 2140–2146, 2017.

[23] A. D. Laud. Theory and application of reward shaping in reinforcement learning. Technical report, 2004.

[24] J. Lehman and K. O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

[25] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems*, pages 206–214, 2012.

[26] M. L. Maher, K. E. Merrick, and R. Saunders. Achieving creative behavior using curious learning agents. In *AAAI spring symposium: Creative intelligent systems*, volume 8, pages 40–46, 2008.

[27] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[29] 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*, pages 1928–1937, 2016.

[30] J.-B. Mouret and J. Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

[31] A. Y. Ng. *Shaping and policy search in reinforcement learning*. PhD thesis, University of California, Berkeley, 2003.

[32] P.-Y. Oudeyer. Computational theories of curiosity-driven learning. *arXiv preprint arXiv:1802.10546*, 2018.

[33] P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1:6, 2009.

[34] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, volume 2017, 2017.

[35] J. K. Pugh, L. B. Soros, and K. O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

[36] S. Risi and J. Togelius. Neuroevolution in games: State of the art and open challenges. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(1):25–41, 2017.

[37] S. Risi, C. E. Hughes, and K. O. Stanley. Evolving plastic neural networks with novelty search. *Adaptive Behavior*, 18(6): 470–491, 2010.

[38] R. M. Ryan and E. L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[39] T. Salimans, J. Ho, X. Chen, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[40] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[41] J. Schmidhuber. Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Frontiers in psychology*, 4:313, 2013.

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

[43] F. P. Such, V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, and J. Clune. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv preprint arXiv:1712.06567*, 2017.

[44] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, et al. Starcraft II: a new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[45] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[46] Y. Wu and Y. Tian. Training agent for first-person shooter game with actor-critic curriculum learning. 2016.

[47] Y. Wu, E. Mansimov, R. B. Grosse, S. Liao, and J. Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In *Advances in neural information processing systems*, pages 5285–5294, 2017.