Comparison of Reinforcement Learning Algorithms in Problems of Acquiring Locomotion Skills in 3D Space

D. A. Kozlov[1]
djoade100@gmail.com

V. V. Myasnikov[1]
vmyas@geosamara.ru

Introduction

In this work, we study the influence of the composition of a set of environmental observations on the solution of the problem of acquiring movement skills by an agent in three-dimensional space. We found that some redundant data can slow down the learning process and interfere with problem solving. Experiments on the effects of rules are conducted in the Unity game engine's environment using the ML-Agents package. The algorithm under study, Soft Actor-Critic, was chosen because it is one of the best ways to learn how to move in three-dimensional space.

Environment

The SimplestBipedal agent is depicted in Figure 1. During this task, the reward is delivered in proportion to the pace and direction of the agent's movement as it approaches the goal..

A pair of legs, each of which may be moved in two joints, are attached to this agent. At the same moment, there is mobility in two planes in the upper joint, and only one plane of mobility in the lower joint at the same time.

The robot is located on a large platform, the size of which exceeds the size of the agent by more than a hundred times. At the beginning of the episode, the agent is placed in the middle of this square area, and the object that is the goal that the agent needs to reach is placed in a random point of this area. If the agent reaches the target, then he receives a reward, and the target disappears and appears at another random point on the site. Now the agent needs to turn to the target and continue moving.

We designed this agent expressly for the experiment since it is a simple physical model that solves a complex problem with a huge number of degrees of freedom while maintaining a low level of complexity. In all tests, this agent receives information about the surroundings, including whether any limbs are now touching the surface, the distance to the target, the direction to the target, and the height of the body above the surface, among other things. Every other piece of information differs from experiment to experiment.

There was a total of nine experiments carried out. There, the sets of observations were organized in a variety of ways to test the theory that the more complete the information supplied to the agent, the more quickly the task will be solved. The data used as observations conveyed to the agent in each experiment were chosen in accordance with Table 1 as the observations transmitted to the agent.

Reinforcement Learning

Reinforcement learning is a type of machine learning in which a system learns by interacting with its environment. The agent, interacting with the environment, receives a reward depending on his actions. The task of the agent is to maximize this reward. In this case, labeled datasets are not needed as supervised learning methods, but a well-formulated reward function is required.

Investigated algorithm

Soft Actor Critic (SAC) is an algorithm that optimizes the probabilistic aspect of a decision-making process, serving as a link between stochastic policy optimization and DDPG type decision making processes. It is not a direct successor to TD3, but it has some characteristics with it.

TABLE I. Performed Experiments

Information transmitted to the agent	Experiment number								
	0	1	2	3	4	5	6	7	8
Translations parts in global coordinates	+	+	+	+	+	+		+	+
Rotate parts in global coordinates	+	+	+	+	+		+	+	+
Translations parts in local coordinates			+	+	+	+	+		
Rotate parts in local coordinates	+		+	+	+		+	+	
Speed of parts in global coordinates	+			+	+	+	+		
Velocity of parts in local coordinates				+	+			+	
Joint positions and angles					+				+
Force applied at the joints					+				+

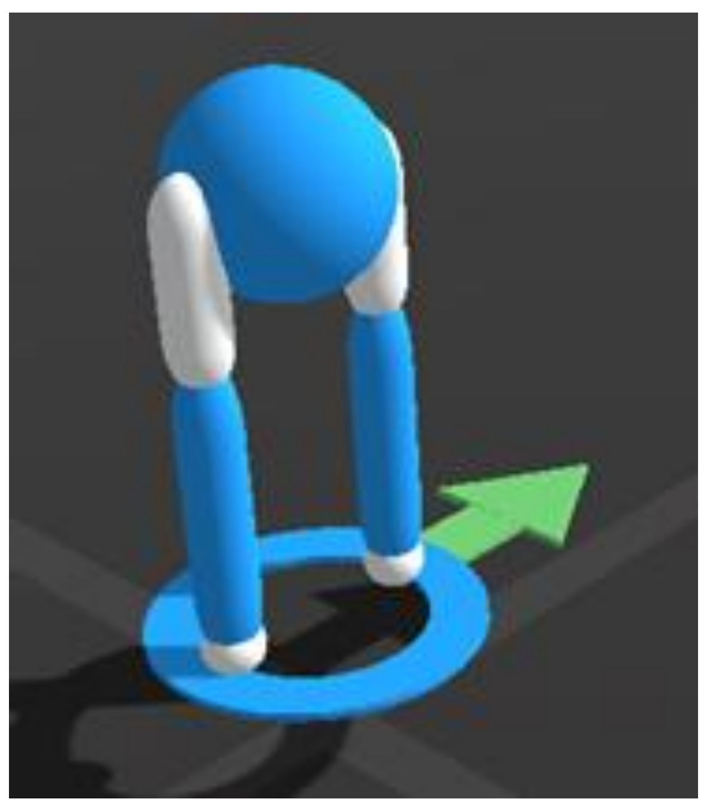

Fig 1. Appearance of Agent SimplestBipedal

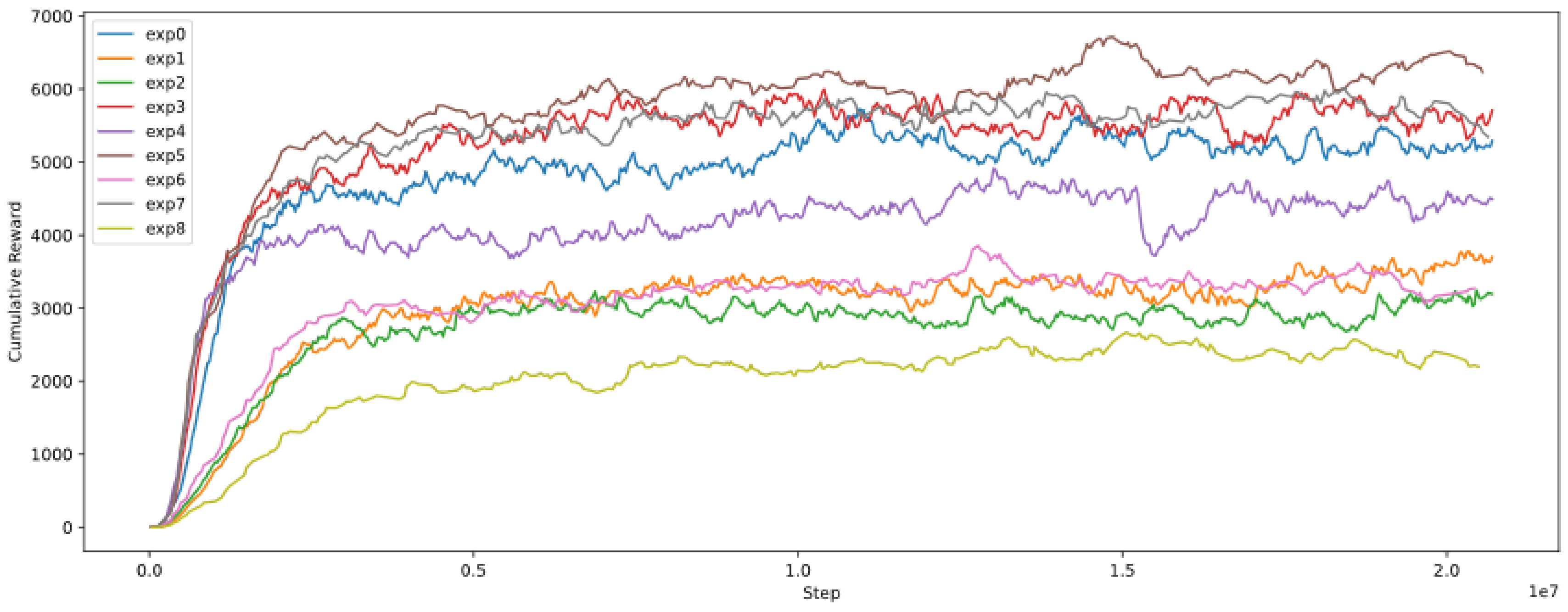

Fig 2. The cumulative reward values used in the experiments are related to the learning processes

Conclusion

According to the findings of this study, the contribution of each component to the set of transmitted environmental observations is not immediately apparent at the beginning of the training process. In this instance, it is critical to conduct preliminary research into how certain data influences the results before beginning training.

Even though the application problem is the same, several reinforcement learning approaches can be used to solve it, and their performances will change solely because of the differences in the sets of environmental observations communicated by the agent. When you initially look at a collection of observations, it is possible that you will not see any evident patterns in the way they have been assembled.

In this context, it may be conceivable to propose improvements to reinforcement learning algorithms that are concerned with the selection of specific observations for transmission to the training system. It is feasible, for example, to conduct a statistical study of the contribution of various characteristics to the learning outcome and to draw conclusions from the results. If we undertake a particular number of training episodes with all distinct sets of observations, we can then rectify their composition and examine the link between a successful solution to the problem and the contribution of each individual parameter at the following step.

Acknowledgement

The work was supported by the Russian Science Foundation grant No. 21-11-00321, https://rscf.ru/en/project/21-11-00321.

[1]SAMARA NATIONAL RESEARCH UNIVERSITY
34, Moskovskoye shosse, Samara, 443086, Russia
www.ssau.ru