УДК 004.855 DOI: 10.18698/0236-3933-2023-4-61-84

МУЛЬТИАГЕНТНОЕ ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ИСПОЛЬЗОВАНИЕМ КОЛЛЕКТИВНОЙ ВНУТРЕННЕЙ МОТИВАЦИИ

В.Э. Большаков bolshakovv@bmstu.ru
С.А. Сакулин sakulin@bmstu.ru
А.Н. Алфимцев alfim@bmstu.ru

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Одной из серьезных проблем в обучении с подкреплением являются редкие вознаграждения от среды. Для решения этой задачи необходимы эффективные методы исследования среды. При создании таких методов исследования используются модели внутренней мотивации. Большинство задач реального мира характеризуются наличием только редких вознаграждений, однако помимо этого существуют мультиагентные среды, в которых обычные методы внутренней мотивации не дают удовлетворительных результатов. В настоящее время востребованы прикладные задачи на стыке этих двух проблем — мультиагентные среды с редкими вознаграждениями. Для решения подобных задач предложен метод CIMA, комбинирующий в себе алгоритмы мультиагентного обучения с моделями внутренней мотивации, использующий как внешнее вознаграждение от среды, так и внутреннее коллективное вознаграждение кооперативной мультиагентной системы. При этом в методе CIMA в качестве базового алгоритма обучения с подкреплением может быть использован любой нейросетевой алгоритм мультиагентного обучения. Эксперименты проведены в специально подготовленной мультиагентной среде с редкими вознаграждениям на базе SMAC, а эффективность предложенного метода обоснована в результате сравнительного анализа с современными методами мультиагентной внутренней мотивации

Ключевые слова

Мультиагентное обучение с подкреплением, глубокое обучение, внутреннее вознаграждение

Поступила 13.02.2023
Принята 19.06.2023

Исследования, выполненные Сакулиным С.А. и Алфимцевым А.Н., поддержаны грантом РНФ (№ 22-21-00711)

Введение. В обучении с подкреплением агент взаимодействует со средой и получает от нее сигнал вознаграждения [1]. Действуя методом проб и ошибок, со временем агент учится максимизировать ожидаемое будущее вознаграждение. Такое вознаграждение называется внешним, получаемым от среды, и, как правило, определяется в соответствии с решаемой задачей, а сама функция вознаграждения задается с помощью экспертных знаний и в явном виде. Прорывные результаты машинного обучения были получены в виртуальных мирах компьютерных игр [2–4] и в задачах управления объектами реального мира [5]. Однако существуют актуальные технические задачи, в которых такой метод машинного обучения практически неприменим.

В задачах обучения с подкреплением, где задействовано множество одновременно обучающихся агентов, больше не работают теоретические гарантии достижения оптимальной стратегии действий, а применение классических методов одноагентного обучения уже не дает успешных результатов. Агенты делают среду взаимодействия нестационарной, недетерминированной и непредсказуемой, мешая друг другу обучаться [6]. При этом ряд задач из реального мира удобно и естественно представляются именно в виде кооперативных мультиагентных систем: управление трафиком, координация автономных устройств, распределение ресурсов [5, 7].

В настоящее время самым успешным решением проблемы мультиагентного обучения является использование архитектуры исполнитель–критик с помощью схемы централизованного обучения с децентрализованным исполнением CTDE (Centralized Training with Decentralized Execution) [8]. В ходе обучения централизованный критик получает полную информацию о среде и агентах, что позволяет более точно оценивать действия с учетом влияния агентов друг на друга. Когда обучение закончено, в режиме эксплуатации централизованный критик больше не используется, а агенты выполняют действия самостоятельно, согласно выученным стратегиям.

Класс задач, в которых традиционные методы обучения с подкреплением недостаточно эффективны — это среды с редкими вознаграждениями [9]. При взаимодействии с такими средами агенту сложнее оценивать успешность своих действий и учиться максимизировать ожидаемое будущее вознаграждение. Например, в среде с редкими вознаграждениями агенту часто нужно выполнить длинную цепочку действий, прежде чем получить само вознаграждение, и понять, насколько его действия были эффективны. Такая ситуация осложняется тем, что для исследования сре-

ды агент в начале обучения действует случайно, что чрезвычайно понижает его шансы на выполнение успешных длинных цепочек действий.

Заметных успехов в решении проблемы редких вознаграждений удалось достичь, используя внутреннее вознаграждение для агентов [10]. Для того чтобы не полагаться на редкое внешнее вознаграждение, агенты дополнительно используют внутреннюю мотивацию, которая позволяет автономно получать новые декларативные знания о среде в процессе обучения. С помощью такого метода возможно создать более общую функцию вознаграждения, которая будет меньше зависеть от прикладной задачи обучения, человеческого фактора (эксперта по знаниям) и будет способна решать проблемы, возникающие в средах с редкими вознаграждениями. Таким общим сигналом вознаграждения для агента может стать сумма внешнего и внутреннего вознаграждений, что делает возможным для агента получать онлайн-подкрепление в реальном масштабе времени.

В настоящей работе исследуются методы, объединяющие алгоритмы мультиагентного обучения с подкреплением и модели внутренней мотивации. На основе сравнительного экспериментального анализа шести моделей внутренней мотивации разрабатывается метод CIMA (Collective Intrinsic Motivation of Agents), который предлагает расширить концепцию внутренней мотивации до мультиагентного уровня и тем самым позволяет добиться наиболее эффективного обучения множества агентов в среде с редкими вознаграждениями. При этом в методе CIMA в качестве базового алгоритма обучения с подкреплением может использоваться любой нейросетевой алгоритм мультиагентного обучения. Для проведения экспериментов в настоящей работе специально разработана оригинальная мультиагентная среда для тестирования уровня внутренней мотивации агентов. Мультиагентная среда основана на компьютерной стратегической игре StarCraft II, доступ к которой осуществляется посредством интерфейса мультиагентного машинного обучения SMAC [11].

Материалы и методы решения задач, принятые допущения. Одним из методов применения внутренней мотивации в обучении с подкреплением является получение знаний о среде, с которой агент взаимодействует, т. е. ее исследование [12]. В классическом обучении с подкреплением агент, исследуя окружающую среду сначала действует полностью или ε-случайно, и со временем, корректируя свое поведение или строя модель среды и учитывая полученные знания, начинает действовать все более близким к оптимальному образом. На практике часто бывает, что простым случайным исследованием среды агент не получает почти никакой

новой информации о ней [13]. Рассмотрим три фундаментальных принципа, использующих внутреннюю мотивацию и помогающих агенту собирать информацию о среде в условиях редких вознаграждений: новизна состояния, несоответствие другим состояниям, ошибка предсказания.

В процессе взаимодействия со средой агент будет совершать доступные действия и попадать в различные ее состояния. Новизну состояния тогда можно выразить через частоту попадания агента в это состояние — чем реже агент в нем был, тем большее значение принимает новизна. Именно это значение используется в качестве внутреннего вознаграждения для агента [14–16].

Такой метод в чистом виде пригоден для использования только в средах с дискретными состояниями, например, мир–сетка, поскольку тогда можно точно подсчитать число попаданий в каждое состояние. Однако в большинстве сложных сред состояния непредставимы в дискретном виде и являются непрерывными значениями, а агент может не попасть в одно и то же состояние дважды. В этом случае прибегают к так называемому псевдосчету — попытке определенным образом представить пространство состояний с помощью хэширования или построения вложенного пространства, что позволяет обобщать состояние с группой соседних схожих состояний [17].

Методом получения внутреннего вознаграждения также является дистилляция случайной нейросети [15]. Данный метод частично комбинирует такие принципы внутренней мотивации, как новизна состояния и ошибка предсказания, поскольку в каждом состоянии предсказывающая нейронная сеть обучается восстанавливать выход случайной фиксированной нейросети. По мере обучения такое предсказание будет становиться точнее для часто посещаемых состояний, а ошибка предсказания для редких состояний по-прежнему будет велика. Данный метод расширили для использования в мультиагентной среде [18]. В этом случае общее наблюдение всех агентов подавалось в обе нейросети, а внутреннее вознаграждение, пропорциональное ошибке предсказания, назначалось всем агентам.

Еще одним принципом внутренней мотивации является сравнение вновь полученного из среды состояния с другими обычно посещаемыми состояниями. Причем различают как внутриэпизодную новизну, учитывающую состояния, которые агент посетил в течение одного текущего эпизода, так и межэпизодную новизну, учитывающую состояния, которые агент посещал за все время обучения [19, 20].

В ряде методов, реализующих данный принцип, для хранения состояний используется буфер. Каждое новое состояние определенным образом сравнивается с содержимым этого буфера и, если оно достаточно непохоже, агент получает вознаграждение за попадание в него. Существуют разные способы хранения состояний в таком буфере, а также стратегии выбора максимально непохожего следующего состояния для посещения агентом. Недостатком такого метода является уменьшение эффективности буфера при увеличении пространства состояний среды [21]. Преодолеть эту проблему призвана группа методов, суть которых заключается в выделении ранее посещенных состояний в некоторое распределение. После чего можно использовать расстояние Кульбака — Лейблера между текущим состоянием и прошлыми в качестве внутреннего вознаграждения [22].

Рассмотрим такой источник внутренней мотивации, как ошибка предсказания. С его помощью можно попытаться направить агента в состояния, переход в которые сложно предугадать, зная прошлое состояние и выполненное в нем действие. Предполагается, что агент использует наблюдения среды для формирования собственного представления о состоянии, в котором находится. Если он попадает туда, где никак не планировал оказаться, выполнив свое предыдущее действие, значит, его предсказание было неверным. Это можно назвать ошибкой предсказания и формализовать в виде некоторого расстояния между реальным и предсказанным состояниями, тогда в качестве внутреннего вознаграждения можно использовать значение ошибки предсказания агента [23].

Модель среды, которую постепенно строит агент, можно представить нейросетью. Если обучать эту нейросеть по ходу исследования среды, то со временем она будет хорошо предсказывать попадание в уже знакомые и многократно посещенные ранее состояния. Действительно и обратное — переход в редкие или вовсе новые состояния такая нейросеть будет предсказывать плохо, а значит, и внутреннее вознаграждение будет большим, пропорционально ошибке предсказания [24].

Методы, реализующие данные фундаментальные принципы внутренней мотивации, их нейросетевые архитектуры, а также архитектуры обучения с подкреплением, над которыми они надстроены, приведены в табл. 1. В настоящей работе для каждого принципа будет предложена и экспериментально апробирована его оригинальная мультиагентная реализация: мультиагентный метод новизны состояний, мультиагентный метод несоответствия другим состояниям, метод CIMA. Кроме того, каж-

дый метод будет выполнен в виде двух типов реализаций: индивидуального и коллективного. Полученные шесть методов будут сравниваться между собой и с state-of-the-art мультиагентным методом LIIR (Learning Individual Intrinsic Reward), который не вошел в табл. 1, так как основан на другом подходе [25].

Таблица 1

Основные принципы, архитектуры и методы реализации внутренней мотивации в обучении с подкреплением

Принцип IM	Метод IM	Архитектура IM	Архитектура RL
Новизна состояния	TRPO-AE-hash [17]	Autoencoder	TRPO
	DDQN-PC [27]	CTS	DQN
	DQN-PixelCNN [28]	PixelCNN	DQN
	DQN-SR [16]	Autoencoder	DQN
Несоответствие	IE [19]	Autoencoder-RNN-MLP	DQN
	ECO [20]	Siamese NN (2×CNN- MLP)	PPO
	EX^2 [21]	Autoencoder-CNN- MLP	TRPO
	CB [22]	CNN-MLP	PPO
	VSIMR [29]	VAE	A2C
Ошибка предсказания	Dynamic-AE [30]	Autoencoder	DQN
	ICM [24]	MLP -CNN	A3C
	EMI [23]	VAE	TRPO

Создатели метода LIIR уделяют особое внимание проблеме присвоения индивидуального вознаграждения агентам при условии получения общего командного вознаграждения и используют метод актор-критик. Авторы предлагают строить функцию внутреннего вознаграждения для каждого агента, которая по-разному их стимулирует на каждом временном шаге. Причем внутренне вознаграждение отдельного агента используется для обучения собственного вспомогательного критика при формировании индивидуальной политики. Сама функция внутреннего вознаграждения параметризована и обновляется так, чтобы максимизировать ожидаемое накопленное командное вознаграждение.

Неопределенность, связанная с неспособностью единичного агента получить точное состояние среды, осложняло применение классической модели марковского процесса принятия решений для обучения с подкреплением. Для преодоления неопределенности состояния среды разработана математическая модель частично наблюдаемого марковского процесса принятия решений POMDP (Partially Observable Markov Decision Process), которая на мультиагентный случай расширяется в виде модели децентрализованного POMDP (Dec-POMDP), за счет введения совместных действий и наблюдений агентов [26].

Модель Dec-POMDP представляет собой кортеж: $\left(N, S, \mathbb{A}, T, R, \mathbb{O}, \Omega, h, b^0\right)$, где N, S, $\mathbb{A}$ и $\mathbb{O}$ — множества агентов, состояний, совместных действий и совместных наблюдений; T, R и Ω — функции переходов, вознаграждения и наблюдений; h — временной горизонт задачи; b^0 — начальное распределение состояний. Множество совместных действий $\mathbb{A}$ состоит из $\mathbb{A} = \times_{i \in n} A_i$, где A_i — множество действий i-го агента. В каждый момент времени t i-й агент выполняет некоторое действие a_i, которое приводит к формированию совместного действия мультиагентной системы $a = \left(a_1, a_2, ..., a_n\right)$. Результат совместного действия a на среду определяется функцией переходов T как вероятность $T\left(s' \mid s, a\right)$ перехода из состояния s в состояние s'.

Множество совместных наблюдений $\mathbb{O}$ состоит из $\mathbb{O} = \times_{i \in n} \mathbb{O}_i$, где $\mathbb{O}_i$ — множество наблюдений, доступных i-му агенту. В каждый момент времени t i-й агент получает некоторое наблюдение o_i, которое образует совместное наблюдение $o = \left(o_1, o_2, ..., o_n\right)$. Функция наблюдений Ω определяет вероятность восприятия агентами состояния среды как совместного наблюдения $\Omega\left(o \mid s', a\right)$.

Используя модель Dec-POMDP множество агентов обучается оптимизировать свои действия путем максимизации ожидаемого кумулятивного вознаграждения $R_t = \sum_{t=0}^{h-1} \gamma^{t-1} r_t\left(o_t, a_t\right)$ с параметром дисконтирования $\gamma \in \left[0, 1\right]$ с некоторым временным горизонтом h и с учетом начального распределения состояний среды $b^0 : S \to \left[0, 1\right]$ в момент времени $t = 0$. При этом поведение i-го агента определяется стохастической стратегией $\pi_i\left(a_i | o_i\right) : \mathbb{O}_i \times A_i \to \left[0, 1\right]$. Цель обучения i-го агента — найти оптимальную стохастическую стратегию π_i^*, которая максимизирует суммарное воз-

награждение агента. Целевая функция *i*-го агента может быть задана в виде $J(\theta_i) = \mathbb{E}_{\pi_{\theta_i}}[R_t]$, где θ_i — вектор параметров стохастической стратегии π_i, который в глубоком обучении является весами нейронной сети. Для максимизации целевой функции в процессе машинного обучения параметры стохастической стратегии θ_i изменяются в направлении ее градиента $\nabla_{\theta_i} J(\theta_i)$.

Благодаря теореме о градиенте стратегии [31] разработано несколько эффективных методов, использующих вычисление градиентов стохастических стратегий, отличающихся только способом определения значений функции ценности $Q_{\pi_{\theta_i}}(o_i, a_i)$. Однако стохастическая стратегия $\pi_{\theta_i}(a_i | o_i)$ оставалась зависимой как от множества наблюдений агента $\mathbb{O}_i$, так и от множества действий A_i, но не учитывала наблюдения и действия других агентов, а потому оценки градиента имели высокую дисперсию. Поэтому была предложена детерминированная стратегия $\mu_{\theta_i}(a_i | o_i) : \mathbb{O}_i \to A_i$, зависящая только от множества наблюдений агента $\mathbb{O}_i$ [32]. В безмодельном офлайн-методе глубокого детерминированного градиента стратегий DDPG (Deep Deterministic Policy Gradient) стратегия μ_θ и функция ценности $Q^{\mu\theta}$ аппроксимировались глубокими нейронными сетями, используя архитектуру исполнитель–критик.

В качестве основы для мультиагентного обучения с подкреплением в настоящей работе выбран state-of-the-art мультиагентный метод MADDPG (Multi-Agent DDPG) [33], который расширил метод DDPG на мультиагентный случай, применил CTDE и продемонстрировал одни из лучших результатов нейросетевого обучения в SMAC. Таким образом, для *i*-го агента из множества одновременно обучающихся агентов *n* детерминированный градиент стратегии μ_{θ_i}, обозначенный как μ_i, может быть вычислен по формуле

$$\nabla_{\theta_i} J(\theta_{\mu_i}) = \mathbb{E}_{\mu_i}\left[\nabla_{\theta_i}\mu_i(a_i | o_i) Q_i^{\mu}((o_1, ..., o_n), (a_1, ..., a_n))|_{a_i = \mu_i(o_i)}\right].$$

Здесь $Q_i^{\mu}((o_1, ..., o_n), (a_1, ..., a_n))$ — совместная функция ценности, зависящая от выполненных действий всех агентов $(a_1, ..., a_n)$ и всех локальных наблюдений $(o_1, ..., o_n)$, полученных каждым агентом отдельно.

В рассматриваемых далее мультиагентных методах внутренней мотивации общее вознаграждение агента $r_t(o_t, a_t)$ состоит из внешнего вознаграждения от среды и внутреннего вознаграждения для каждого

агента. Для i-го агента $r_i = r^{ex} + r_i^{in}$, где r^{ex} является общим для всех агентов внешним вознаграждением; r_i^{in} — внутреннее вознаграждение i-го агента в индивидуальных сценариях или же общее для всех агентов внутреннее вознаграждение в коллективных сценариях. При использовании метода MADDPG в чистом виде внутренние вознаграждения агентов равны нулю.

Результаты. Поскольку рассматриваемую в рамках StarCraft II среду взаимодействия можно представить в виде двумерного мира–сетки, то для внутренней мотивации, на основе принципа новизны состояний, использовалось разбиение состояний на дискретные ячейки. В каждый момент времени t проводился подсчет числа попаданий в определенные области двумерного мира–сетки s_g. Для хранения этой информации применялся двумерный массив SN. Внутреннее вознаграждение для i-го агента назначается следующим образом:

$$r_i^{in}\left(s_g\right) = \left(1 - \frac{SN_t^i\left(s_g\right) - \min\limits_{s_g}\left(SN_t^i\right)}{\max\limits_{s_g}\left(SN_t^i\right)}\right)^2,$$

где $SN_t^i\left(s_g\right)$ — число попаданий i-го агента в его текущее состояние в двумерном мире–сетке s_g (соответствующее некоторому состоянию среды s) на момент времени t, $\min\limits_{s_g}\left(SN_t^i\right)$ и $\max\limits_{s_g}\left(SN_t^i\right)$ — это соответственно минимальное и максимальное число попаданий i-го агента среди всех состояний двумерного мира–сетки на момент времени t.

Два варианта реализации данного мультиагентного метода новизны состояний MASN (Multi-Agent State Novelty Method) — индивидуальный (MASNi) и коллективный (MASNc) — отличаются тем, что в первом варианте подсчет посещений каждого состояния был индивидуальным для каждого агента, т. е. для n агентов $SN = \left\{SN^{\,i}, ..., SN^{\,n}\right\}$, во втором варианте — общим для всех, т. е. все агенты обновляют значения в SN.

Внутренняя мотивация на основе принципа несоответствия другим состояниям удобно реализуется с помощью вариационного автокодировщика VAE (Variational Autoencoder) [34]. При использовании VAE для проецирования пространства состояний в вероятностное скрытое представление, которое отражает внутреннюю структуру среды, можно естественным образом получить некоторую меру несоответствия. Эта мера определяется тем, насколько апостериорное распределение

скрытого представления отклоняется от априорного предположения. Измерить это можно с помощью расстояния Кульбака — Лейблера $KL\left(p\left(z|\, s\right)||\, p\left(z\right)\right)$, где для состояния среды s $p\left(z\right)$ — априорное распределение скрытого представления состояния среды z, $p\left(z|s\right)$ — апостериорное. Такая мера используется в качестве внутреннего вознаграждения в состоянии s:

$$r^{in}\left(s\right)=KL\left(p\left(z|s\right)||\, p\left(z\right)\right).$$

Точно определять апостериорное распределение сложно и нецелесообразно, поэтому его можно аппроксимировать вариационным распределением $q_{\varphi}(z\,|\,s)$, параметризованного с помощью параметра нейросетей φ. Это вполне естественно можно выполнить с помощью вариационного автокодировщика VAE, который к тому же уменьшает размерность входного состояния s при переводе его в скрытое представление z. Вариационный автокодировщик минимизирует следующую функцию потерь:

$$\mathcal{L}=\left(\theta,\,\varphi\right)=\mathbb{E}_{q_{\varphi}(z\,|\,x)}\left[\log p_{\theta}\left(x\,|\,z\right)\right]-KL\left(q_{\varphi}\left(z\,|\,x\right)||\, p\left(x\right)\right),$$

где первая часть выражения является ошибкой восстановления выходного вектора по входному, вторая часть используется для приближения апостериорного распределения к априорному; θ и φ — параметры нейросетей кодировщика и декодировщика; x — входной вектор, состоящий из кортежа совместных наблюдений и действий вида $\left(o_1,\,...,\,o_n,\,a_1,\,...,\,a_n\right)$.

Мультиагентный метод MADOS (Multi-Agent Discrepancy of Other States Method) несоответствия другим состояниям также был реализован в двух вариантах. В индивидуальном варианте (MADOSi) каждый агент имеет собственный вариационный автокодировщик, который на вход получает наблюдения и действия лишь этого агента, а внутренние вознаграждения были индивидуальными. В коллективном варианте (MADOSc) вариационный автокодировщик был общим для всех агентов, т. е. централизованным. В этом случае входным вектором являются уже наблюдения и последние действия всех агентов, а внутреннее вознаграждение становится коллективным.

В качестве внутреннего вознаграждения в третьем принципе внутренней мотивации используется ошибка предсказания агентом следующего состояния, в которое он попадет после выполнения некоторого действия.

Для реализации подобной внутренней мотивации был выбран автокодировщик, входом для которого служило наблюдение агента и выполненное действие. Для моделирования индивидуальных и коллективных внутренних вознаграждений вновь было реализовано два варианта: отдельная нейросеть автокодировщик для каждого агента MAPEi (Multi-Agent Prediction Error Method) и одна общая нейросеть CIMA для всех агентов.

Внутреннее вознаграждение для *i*-го агента в состоянии s вычисляется как ошибка предсказания следующего наблюдения $o_i'^{(true)}$:

$$r^{in}(s) = \left|\left| \left(o_i'^{(pred)} - o_i'^{(true)} \right) \right|\right|_2^2.$$

Детали обучения агентов с помощью метода CIMA приведены в алгоритме 1.

Алгоритм 1. Коллективная внутренняя мотивация агентов

1. Инициализировать:

$\alpha_{ac} \leftarrow 0,001$; $\alpha_{cr} \leftarrow 0,001$; $\alpha_{ae} \leftarrow 0,001$; $\tau \leftarrow 0,01$; $\gamma \leftarrow 0,99$;

$Ns \leftarrow 2 \cdot 10^6$; $i \leftarrow 1, ..., n$; $\mathcal{D} \leftarrow 0$; $Bs \leftarrow 2048$; $j \leftarrow 1, ..., Bs$;

$\mathcal{N} \leftarrow \xi$; $\theta \leftarrow \xi$; $\theta' \leftarrow \xi$; $\eta \leftarrow \xi$; $\eta' \leftarrow \xi$.

2. Цикл по шагам от $t = 1$ до Ns шагов обучения.

3. Получить совместное наблюдение $o = (o_1, ..., o_n)$.

4. Для каждого *i*-го агента выбрать действие $a_i = \mu_{\theta_i}(o_i) + \mathcal{N}_t$ с учетом текущей стратегии μ_{θ_i} и случайного шума $\mathcal{N}_t$.

5. Выполнить действия $a = (a_1, ..., a_n)$, получить внешнее вознаграждение r^{ex} и новое совместное наблюдение $o' = (o_1', ..., o_n')$.

6. Получить коллективное внутреннее вознаграждение $r^{in} = \left|\left| \psi_\eta(o, a) - o' \right|\right|_2^2$ с учетом текущего состояния автокодировщика ψ_η и вычислить общее вознаграждение $r = r^{ex} + r^{in}$.

7. Сохранить (o, a, r, o') в буфер воспроизведения $\mathcal{D}$.

8. Цикл по агентам от $i = 1$ до n агентов.

9. Выбрать случайную выборку (o^j, a^j, r^j, o'^j) размером Bs из буфера $\mathcal{D}$.

10. Используя целевую нейронную сеть критика, вычислить:

$$y^j = r_i^j + \gamma Q_i^{\mu'}(o'^j, a_1', ..., a_n')\big|_{a_k' = \mu_k'(o_k^j)}.$$

11. Обновить основную нейронную сеть критика со скоростью обучения α_{cr}, минимизируя функцию потерь:

$$\mathcal{L}(\theta_i) = \frac{1}{Bs}\sum_j \left(y^i - Q_i^{\mu}\left(o^j, a_1^j, ..., a_n^j\right)\right)^2.$$

12. Обновить основную нейронную сеть исполнителя со скоростью обучения α_{ac}, используя градиент:

$$\nabla_{\theta_i} J \approx \frac{1}{Bs}\sum_j \nabla_{\theta_i}\mu_i\left(o_i^j\right)\nabla_{a_i}Q_i^{\mu}\left(o^j, a_1^j, ..., a_i, ...a_n^j\right)\Big|_{a_i=\mu_i\left(o_i^j\right)}.$$

13. Обновить основную нейронную сеть автокодировщика со скоростью обучения α_{ae}, минимизируя функцию потерь:

$$\mathcal{L}(\eta) = \frac{1}{Bs}\sum_j \left(\psi_\eta\left(o^j, a_1^j, ..., a_n^j\right) - o'^j\right)^2.$$

14. Конец цикла.

15. Обновить параметры целевых нейросетей с помощью техники мягкой замены, используя коэффициент τ:

$$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\,\theta_i'.$$

16. Конец цикла.

В алгоритме введены обозначения: Ns — число шагов обучения; Bs — размер мини-выборки из буфера воспроизведения $\mathcal{D}$, при этом случайными значениями ξ инициализируются исследовательский шум $\mathcal{N}$ и веса основных и целевых нейронных сетей θ, θ', а также нейросети автокодировщика η. Указанные численные значения для инициализации алгоритма подобраны экспериментальным путем. Алгоритм использует схему обучения CTDE, относится как к классу алгоритмов вычисления стратегий, так и к классу алгоритмов вычисления ценности состояний–действий, наследует от MADDPG использование случайного шума $\mathcal{N}$ вместо ε (жадного выбора действий) и технику мягкой замены весов. На рис. 1 приведена общая архитектура метода CIMA, в которой централизованные критики для обучения используют внутреннее r^{ex} и внешнее r^{in} вознаграждения, а децентрализованные исполнители совместными действиями a и совместными наблюдениями o оказывают влияние на обучение нейросетевого модуля внутренней мотивации.

Для проведения сравнительных экспериментов по мультиагентному обучению с подкреплением выбрана популярная программная библиотека SMAC, предоставляющая возможность децентрализованного управле-

ния множеством агентов в среде стратегической компьютерной игры StarCraft II. Каждая союзная управляемая сущность контролируется независимым обучающимся агентом, имеющим доступ лишь к локальным наблюдениям среды. Управляемые сущности противодействующей команды контролирует внутриигровой искусственный интеллект StarCraft II.

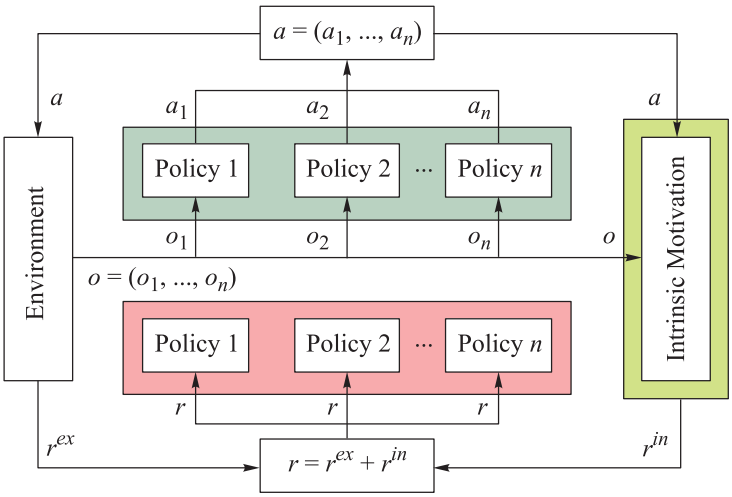


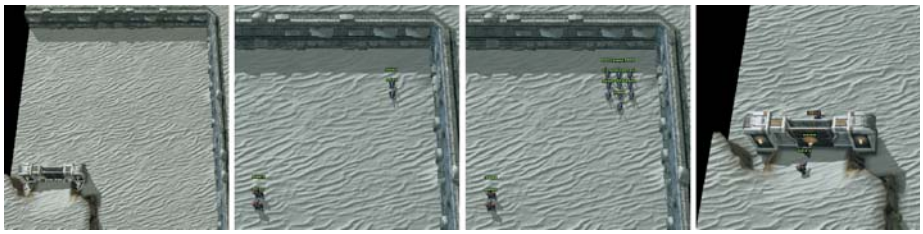
Рис. 1. Общая архитектура метода CIMA

Для наблюдения агентам доступны следующие показатели: очки здоровья, взаимное расположение и тип управляемой сущности в зоне видимости. Таким образом, среда является мультиагентной и частично наблюдаемой. Вознаграждение назначается агентам в двух случаях: при уничтожении противодействующей команды или при понижении очков здоровья сущностей противодействующей команды.

Для проведения экспериментов с коллективной внутренней мотивацией была разработана IM-среда, включающая в себя специальный элемент взаимодействия (рис. 2). При одновременном приближении всех союзных агентов к воротам они открываются и после попадания агентов внутрь безопасной зоны ворота закрываются, давая возможность спрятаться от сущностей противодействующей команды и впредь быть неуязвимыми. Это сильно повышает шансы на победу, однако требует последовательных кооперативных действий в условиях редкого вознаграждения.

Вознаграждения в данной среде являются редкими, потому что агенты получают наибольшее вознаграждение за выигрыш эпизода, но сама эта победа в ряде случаев требует длинной цепочки совместных действий. Агентам также дается незначительное вознаграждение за атаку по сущностям противодействующей команды, но оно суммарно на два порядка ниже, чем

вознаграждение за победу во всем эпизоде. Союзные агенты управляли дальнобойными сущностями типа морпех (*marine*), а в противодействующей команде были сущности ближнего боя типа зерглинг (*zergling*).



а *б* *в* *г*

Рис. 2. Сценарии взаимодействия агентов в мультиагентной IM-среде: общий вид среды (*а*), начало эпизода в сценарии «двое против двух» (*б*), начало эпизода в сценарии «двое против десяти» (*в*), победное завершение эпизода для союзных агентов (*г*)

В экспериментах с индивидуальным внутренним вознаграждением каждый агент получал вознаграждение на основе только своих наблюдений и действий, при коллективном вознаграждении агенты получали общее вознаграждение на основе наблюдений и действий всех союзников. В тех экспериментах, где число союзных сущностей превышало или было равно числу противодействующих сущностей, лучшей стратегией оказывалось прямое сражение, в экспериментах, где противодействующих сущностей было больше, наилучшей стратегией являлось совместное попадание агентов в укрытие, где им ничего не угрожало. Поэтому в качестве показательных экспериментов рассмотрим четыре экстремальных сценария (табл. 2).

Таблица 2

Основные сценарии проведения экспериментов при числе союзных агентов 2

Параметр основных сценариев	Число противодействующих агентов			
	2	2	10	10
Тип внутреннего вознаграждения	Индивидуальный	Коллективный	Индивидуальный	Коллективный

Во всех экспериментах в качестве оптимизатора параметров нейронной сети использовался Adam со скоростью обучения 0,001, обновление параметров целевых нейронных сетей осуществлялось по принципу мягкой за-

мены с параметром $\tau = 0{,}01$. Дисконтирующий множитель γ принимал значение 0,99. Размер буфера воспроизведения составлял 10^6, а обновление параметров нейросетей происходило каждый раз после добавления 100 новых записей в буфер. Размер мини-выборки равен 2048. Все нейронные сети состояли из полносвязных слоев, нейронные сети исполнителей — из двух MLP-слоев с функцией активации ReLU с 64 нейронами на слой, выходной слой с функцией активации Softmax. Для реализации нейронных сетей использовался фреймворк PyTorch, для методов MADDPG и LIIR — гиперпараметры, рекомендуемые их авторами.

Обсуждение полученных результатов. В ходе экспериментов оценивалось среднее вознаграждение, получаемое агентами в конце эпизода. Обучение проходило $2 \cdot 10^6$ шагов, после чего агенты участвовали еще в 100 эпизодах в тестовом режиме. В целях сравнения для каждого сценария была смоделирована оптимальная стратегия, созданная экспертом с учетом знаний о среде. Такая стратегия позволяла агентам достигать максимального возможного вознаграждения за эпизод (рис. 3).

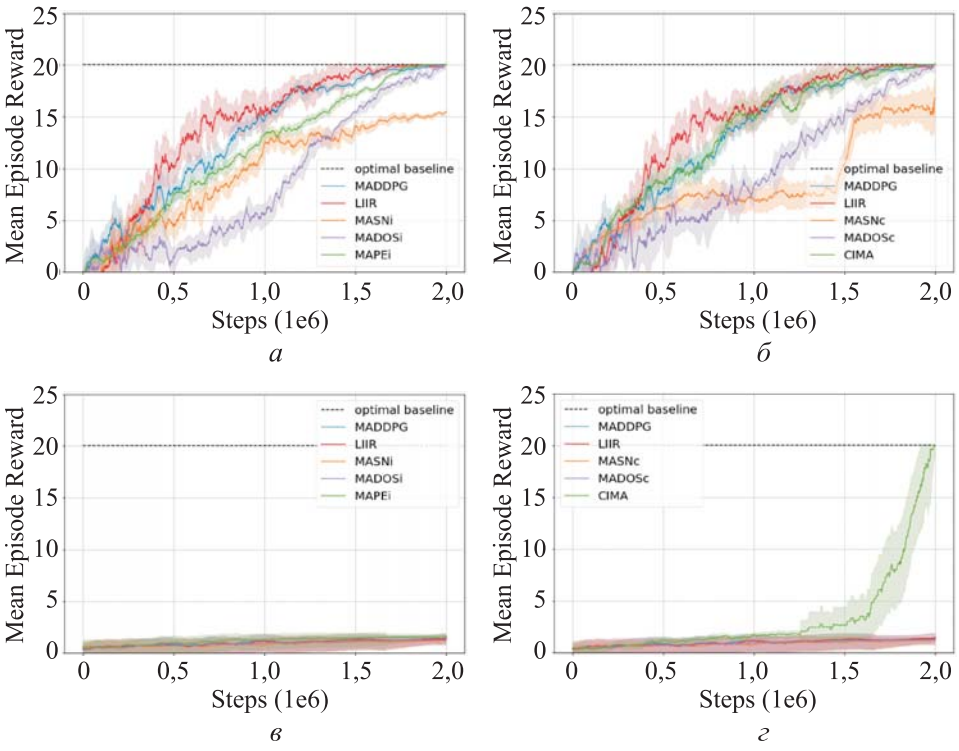


Рис. 3. Скриншоты результатов мультиагентного обучения в различных сценариях:

а, б — индивидуальные и коллективные вознаграждения, сценарии 2m Vs 2z;
в, г — индивидуальные и коллективные вознаграждения, сценарии 2m Vs 10z

В сценариях, где число союзных и противодействующих агентов одинаковое и внутреннего вознаграждения не требуется, почти все методы показали результаты, близкие к оптимальному. Однако при большом перевесе в пользу противодействующей команды по числу управляемых сущностей только у метода CIMA получается выучить оптимальную стратегию.

Наименее эффективным оказался метод MASNi. Метод не достигает высшего вознаграждения даже в относительно простых сценариях, однако его версия MASNc, использующая коллективное внутреннее вознаграждение, показывает себя лучше, чем версия с индивидуальными вознаграждениями. Можно также отметить не очень стабильное обучение методов MADOSi и MADOSc — в обоих сценариях 2m Vs 10z внутреннее вознаграждение от вариационного автокодировщика, по всей видимости, мешала обучению, но все же метод способен достигать оптимального вознаграждения в простых сценариях 2m Vs 2z.

Наиболее интересный результат показал метод CIMA. В сценарии 2m Vs 10z метод CIMA является единственным, при использовании которого агенты способны, коллективно исследуя среду, найти спасение в безопасной зоне и получить максимально возможное вознаграждение для данной среды (табл. 3). Это единственный подход, который после обучения достигает 100 % побед в тестовых эпизодах.

Таблица 3

Тестовые результаты средних вознаграждений для методов мультиагентного обучения в различных сценариях

Метод	Среднее вознаграждение в сценарии	
	2m Vs 2z	2m Vs 10z
Optimal baseline	20,04	20,04
MADDPG	20,04	1,60
LIIR	20,04	1,67
MASNi	15,68	1,29
MASNc	16,75	1,35
MADOSi	20,04	1,34
MADOSc	20,04	1,39
MAPEi	20,04	1,72
CIMA	**20,04**	**20,04**

Важно проанализировать посещаемые агентами состояния с точки зрения широты охвата их исследовательских действий. Для этого были построены тепловые карты, на которых отмечаются позиции агентов

в разные моменты времени. След агентов на тепловых картах с течением времени постепенно испаряется. Сравнение методов MADDPG и CIMA приведено на рис. 4. Можно отметить, что агенты, использующие метод MADDPG и стратегию исследования среды, основанную на случайном шуме, посещают только области, близкие к их стартовой позиции. Все становится даже хуже в сценарии 2m Vs 10z, поскольку большое число противодействующих агентов почти сразу уничтожает союзных агентов.

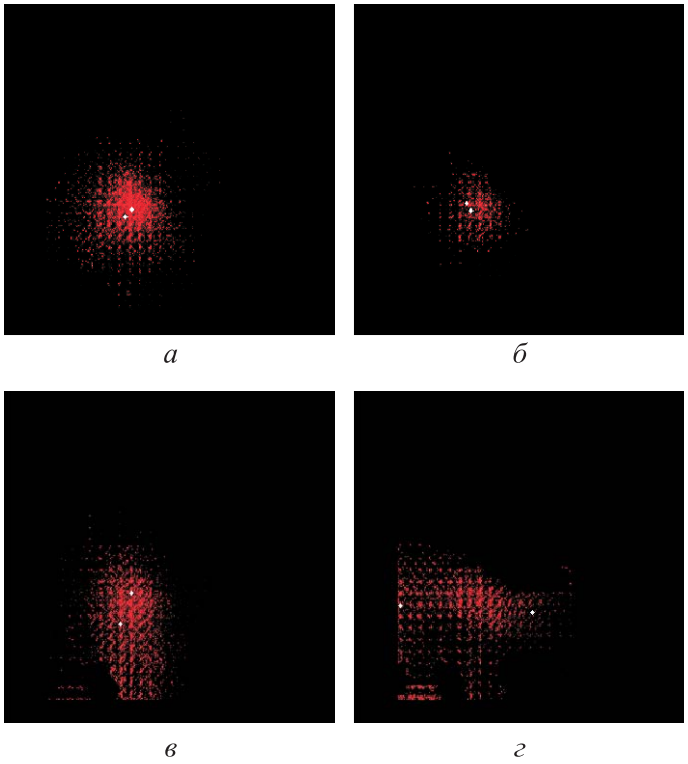


Рис. 4. Тепловые карты исследования среды:
а, *б* — метод MADDPG в сценариях 2m Vs 2z и 2m Vs 10z;
в, *г* — метод CIMA в сценариях 2m Vs 2z и 2m Vs 10z

В то же время агенты, использующие для обучения метод CIMA, посещают гораздо более обширные области на карте, что позволяет им находить безопасную зону и достигать максимальные показатели по вознаграждению и числу побед.

Таким образом, коллективная внутренняя мотивация работает лучше индивидуальной. В качестве объяснения основополагающей причины такого преимущества может выступить аналогия в виде большей эффективности CTDE обучения актор-критиков в сравнении с независимым мультиагентным обучением с подкреплением. В методах CTDE агенты

делятся информацией о среде (своими наблюдениями), отправляя в централизованного критика эти данные. В методе CIMA агенты в некотором роде получают коллективную мотивацию от ошибки предсказания будущих состояний, причем для такой новизны достаточно лишь одному агенту попасть в неизвестную область, и тогда весь входной вектор наблюдений будет отличаться от предыдущих, вызывая повышенное внутреннее вознаграждение от централизованного модуля новизны.

Заключение. В последние годы под внутренней мотивацией понималось несколько смежных технологий машинного обучения. В самом простом случае за метод внутренней мотивации принимался обычный reward shaping или наоборот внутреннюю мотивацию считали частным случаем meta learning. Для мультиагентного случая внутреннюю мотивацию могли использовать как вспомогательный инструмент для решения задачи разделения коллективного вознаграждения — credit assignment problem. При этом с точки зрения нейросетевой архитектуры реализация IM-методов прошла путь от дополнительного линейного слоя сети, схемы актор-критик до практически универсальных методов, готовых работать с любой изначальной архитектурой RL-обучения.

На основе сравнительного экспериментального анализа шести методов, построенных на базе трех фундаментальных принципов внутренней мотивации (новизна состояния, несоответствие другим состояниям, ошибка предсказания), был предложен метод CIMA для мультиагентного обучения с подкреплением в условиях редких вознаграждений от среды. Метод CIMA основан на идее коллективной внутренней мотивации агентов. Этот метод может быть использован вместе с известными алгоритмами обучения с подкреплением в качестве модуля для эффективного исследования среды, так как модель среды обучается на стадии централизованного обучения.

Метод CIMA позволяет агентам действовать коллективно, а не в одиночку и получать совместные бонусы от внутренней мотивации, находя и стимулируя наиболее эффективные действия агентов.

Специально разработанная мультиагентная среда для исследования уровня внутренней мотивации агентов на базе популярной программной библиотеки SMAC может стать еще одним дополнительным инструментом анализа и развития методов, архитектур и принципов внутренней мотивации мультиагентных систем.

ЛИТЕРАТУРА

[1] Singh S., Lewis R.L., Barto A.G., et al. Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Auton. Mental Develop.*, 2010, vol. 2, no. 2, pp. 70–82. DOI: https://doi.org/10.1109/TAMD.2010.2051031

[2] Mnih V., Kavukcuoglu K., Silver D., et al. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, no. 7540, art. 529.
DOI: https://doi.org/10.1038/nature14236

[3] Silver D., Huang A., Maddison C., et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, vol. 529, pp. 484–489.
DOI: https://doi.org/10.1038/nature16961

[4] Vinyals O., Babuschkin I., Czarnecki W.M., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, vol. 575, pp. 350–354.
DOI: https://doi.org/10.1038/s41586-019-1724-z

[5] El-Sallab A.A., Abdou M., Perot E., et al. Deep reinforcement learning framework for autonomous driving. *arXiv:1704.02532*.
DOI: https://doi.org/10.48550/arXiv.1704.02532

[6] Yang Y. Many-agent reinforcement learning. London, University College, 2021.

[7] Wiering M. Multi-agent reinforcement learning for traffic light control. *ICML*, 2000, pp. 1151–1158.

[8] Zheng L., Cheng J., Wang J., et al. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *NeurIPS*, 2021, vol. 1, pp. 3757–3769.

[9] Bellemare M.G., Naddaf Y., Veness J., et al. The arcade learning environment: an evaluation platform for general agents (extended abstract). *IJCAI*, 2015, pp. 4148–4152.

[10] Arthur A., Matignon L., Hassas S. A survey on intrinsic motivation in reinforcement learning. *arXiv:1908.06976*. DOI: https://doi.org/10.48550/arXiv.1908.06976

[11] Samvelyan M., Rashid T., de Witt C.S., et al. The starcraft multi-agent challenge. *arXiv:1902.04043*. DOI: https://doi.org/10.48550/arXiv.1902.04043

[12] Efroni Y., Mannor S., Pirotta M. Exploration-exploitation in constrained MDPs. *arXiv:2003.02189*. DOI: https://doi.org/10.48550/arXiv.2003.02189

[13] Jiang J., Lu Z. The emergence of individuality. *PMLR*, 2021, vol. 139, pp. 4992–5001.

[14] Martin J., Sasikumar S.N., Everitt T., et al. Count-based exploration in feature space for reinforcement learning. *IJCAI*, 2017, pp. 2471–2478.
DOI: https://doi.org/10.24963/ijcai.2017/344

[15] Burda Y., Edwards H., Storkey A., et al. Exploration by random network distillation. *arXiv:1810.12894*. DOI: https://doi.org/10.48550/arXiv.1810.12894

[16] Machado M.C., Bellemare M.G., Bowling M. Count-based exploration with the successor representation. *arXiv:1807.11622*.
DOI: https://doi.org/10.48550/arXiv.1807.11622

[17] Tang H., Houthooft R., Foote D., et al. Exploration: a study of count-based exploration for deep reinforcement learning. *NIPS*, 2017, vol. 1, pp. 2754–2763.

[18] Charoenpitaks K., Limpiyakorn Y. Multi-agent reinforcement learning with clipping intrinsic motivation. *Int. J. Mach. Learn.*, 2022, vol. 12, no. 3, pp. 85–90. DOI: https://doi.org/10.18178/ijmlc.2022.12.3.1084

[19] Oh J., Guo X., Lee H., et al. Actionconditional video prediction using deep networks in Atari games. *NIPS*, 2015, pp. 2863–2871.

[20] Savinov N., Raichuk A., Marinier R., et al. Episodic curiosity through reachability. *arXiv:1810.02274*. DOI: https://doi.org/10.48550/arXiv.1810.02274

[21] Fu J., Co-Reyes J., Levine S. Ex2: exploration with exemplar models for deep reinforcement learning. *NIPS*, 2017, pp. 2577–2587.

[22] Kim Y., Nam W., Kim H., et al. Curiosity-bottleneck: exploration by distilling task-specific novelty. *PMLR*, 2019, pp. 3379–3388.

[23] Kim H., Kim J., Jeong Y., et al. EMI: exploration with mutual information. *ICML*, 2019, vol. 97, pp. 5837–5851.

[24] Pathak D., Agrawal P., Efros A.G., et al. Curiosity-driven exploration by self-supervised prediction. *IEEE CVPRW*, 2017, pp. 488–489. DOI: https://doi.org/10.1109/CVPRW.2017.70

[25] Du Y., Han L., Fang M., et al. Liir: learning individual intrinsic reward in multi-agent reinforcement learning. *NIPS*, 2019, pp. 4403–4414.

[26] Amato C., Konidaris G.D., Cruz G., et al. Planning for decentralized control of multiple robots under uncertainty. *IEEE ICRA*, 2015, pp. 1241–1248. DOI: https://doi.org/10.1109/ICRA.2015.7139350

[27] Bellemare M., Srinivasan S., Ostrovski G., et al. Unifying count-based exploration and intrinsic motivation. *NIPS*, 2016, pp. 1471–1479.

[28] Ostrovski G., Bellemare M.G., van den Oord A., et al. Countbased exploration with neural density models. *ICML*, 2017, pp. 2721–2730.

[29] Klissarov M., Islam R., Khetarpal K., et al. Variational state encoding as intrinsic motivation in reinforcement learning. *ICLR*, 2019, pp. 2–7.

[30] Stadie B.C., Levine S., Abbeel P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv:1507.00814*. DOI: https://doi.org/10.48550/arXiv.1507.00814

[31] Sutton R.S., Barto A.G. Reinforcement learning. An introduction. Cambridge, MIT Press, 2018.

[32] Lillicrap T.P., Hunt J.J., Pritzel A., et al. Continuous control with deep reinforcement learning. *arXiv:1509.02971*. DOI: https://doi.org/10.48550/arXiv.1509.02971

[33] Lowe R., Wu Y., Tamar A., et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *NIPS*, 2017, pp. 6382–6393.

[34] Kingma D.P., Welling M. Auto-encoding variational bayes. *arXiv:1312.6114*. DOI: https://doi.org/10.48550/arXiv.1312.6114

Большаков Владислав Эдуардович — ассистент кафедры «Информационные системы и телекоммуникации» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Сакулин Сергей Александрович — канд. техн. наук, доцент кафедры «Информационные системы и телекоммуникации» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Алфимцев Александр Николаевич — д-р техн. наук, заведующий кафедрой «Информационные системы и телекоммуникации» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Просьба ссылаться на эту статью следующим образом:

Большаков В.Э., Сакулин С.А., Алфимцев А.Н. Мультиагентное обучение с подкреплением с использованием коллективной внутренней мотивации. *Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение*, 2023, № 4 (145), с. 61–84.
DOI: https://doi.org/10.18698/0236-3933-2023-4-61-84

MULTI-AGENT REINFORCEMENT LEARNING USING THE COLLECTIVE INTRINSIC MOTIVATION

V.E. Bolshakov bolshakovv@bmstu.ru
S.A. Sakulin sakulin@bmstu.ru
A.N. Alfimtsev alfim@bmstu.ru

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

One of the serious problems facing the reinforcement learning is infrequency in the environment rewards. To solve this problem, effective methods for studying the environment are required. Using the intrinsic motivation principle is one of the approaches to create such research methods. Most real-world problems are characterized by only the infrequent rewards; however, there are additionally multi-agent environments, where the conventional methods of intrinsic motivation are not providing satisfactory results. Currently, applied problems are in demand at the intersection of these two problems, i.e., multi-agent environments with infrequent rewards. To solve such problems, the CIMA (Collective Intrinsic Motivation of Agents) method is proposed combining the multi-agent learning algorithms with the internal motiva-

Keywords

Multi-agent reinforcement learning, deep learning, intrinsic reward

tion models and using both external reward from the environment and the internal collective reward from the cooperative multi-agent system. Moreover, the CIMA method is able to use any neural network multi-agent learning algorithm as the basic reinforcement learning algorithm. Experiments were carried out in a specially prepared multi-agent environment with the infrequent rewards based on SMAC; the proposed method efficiency was justified by results of the comparative analysis with the modern methods of multi-agent internal motivation

Received 13.02.2023
Accepted 19.06.2023

The research carried out by Sakulin S.A. and Alfimtsev A.N. was supported by the RSF grant no. 22-21-00711

REFERENCES

[1] Singh S., Lewis R.L., Barto A.G., et al. Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Auton. Mental Develop.*, 2010, vol. 2, no. 2, pp. 70–82. DOI: https://doi.org/10.1109/TAMD.2010.2051031

[2] Mnih V., Kavukcuoglu K., Silver D., et al. Human-level control through deep reinforcement learning. *Nature*, 2015, vol. 518, no. 7540, art. 529.
DOI: https://doi.org/10.1038/nature14236

[3] Silver D., Huang A., Maddison C., et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, vol. 529, pp. 484–489.
DOI: https://doi.org/10.1038/nature16961

[4] Vinyals O., Babuschkin I., Czarnecki W.M., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, vol. 575, pp. 350–354.
DOI: https://doi.org/10.1038/s41586-019-1724-z

[5] El-Sallab A.A., Abdou M., Perot E., et al. Deep reinforcement learning framework for autonomous driving. *arXiv:1704.02532*.
DOI: https://doi.org/10.48550/arXiv.1704.02532

[6] Yang Y. Many-agent reinforcement learning. London, University College, 2021.

[7] Wiering M. Multi-agent reinforcement learning for traffic light control. *ICML*, 2000, pp. 1151–1158.

[8] Zheng L., Cheng J., Wang J., et al. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *NeurIPS*, 2021, vol. 1, pp. 3757–3769.

[9] Bellemare M.G., Naddaf Y., Veness J., et al. The arcade learning environment: an evaluation platform for general agents (extended abstract). *IJCAI*, 2015, pp. 4148–4152.

[10] Arthur A., Matignon L., Hassas S. A survey on intrinsic motivation in reinforcement learning. *arXiv:1908.06976*. DOI: https://doi.org/10.48550/arXiv.1908.06976

[11] Samvelyan M., Rashid T., de Witt C.S., et al. The starcraft multi-agent challenge. *arXiv:1902.04043*. DOI: https://doi.org/10.48550/arXiv.1902.04043

[12] Efroni Y., Mannor S., Pirotta M. Exploration-exploitation in constrained MDPs. *arXiv:2003.02189*. DOI: https://doi.org/10.48550/arXiv.2003.02189

[13] Jiang J., Lu Z. The emergence of individuality. *PMLR*, 2021, vol. 139, pp. 4992–5001.

[14] Martin J., Sasikumar S.N., Everitt T., et al. Count-based exploration in feature space for reinforcement learning. *IJCAI*, 2017, pp. 2471–2478.
DOI: https://doi.org/10.24963/ijcai.2017/344

[15] Burda Y., Edwards H., Storkey A., et al. Exploration by random network distillation. *arXiv:1810.12894*. DOI: https://doi.org/10.48550/arXiv.1810.12894

[16] Machado M.C., Bellemare M.G., Bowling M. Count-based exploration with the successor representation. *arXiv:1807.11622*.
DOI: https://doi.org/10.48550/arXiv.1807.11622

[17] Tang H., Houthooft R., Foote D., et al. Exploration: a study of count-based exploration for deep reinforcement learning. *NIPS*, 2017, vol. 1, pp. 2754–2763.

[18] Charoenpitaks K., Limpiyakorn Y. Multi-agent reinforcement learning with clipping intrinsic motivation. *Int. J. Mach. Learn.*, 2022, vol. 12, no. 3, pp. 85–90.
DOI: https://doi.org/10.18178/ijmlc.2022.12.3.1084

[19] Oh J., Guo X., Lee H., et al. Actionconditional video prediction using deep networks in Atari games. *NIPS*, 2015, pp. 2863–2871.

[20] Savinov N., Raichuk A., Marinier R., et al. Episodic curiosity through reachability. *arXiv:1810.02274*. DOI: https://doi.org/10.48550/arXiv.1810.02274

[21] Fu J., Co-Reyes J., Levine S. Ex2: exploration with exemplar models for deep reinforcement learning. *NIPS*, 2017, pp. 2577–2587.

[22] Kim Y., Nam W., Kim H., et al. Curiosity-bottleneck: exploration by distilling task-specific novelty. *PMLR*, 2019, pp. 3379–3388.

[23] Kim H., Kim J., Jeong Y., et al. EMI: exploration with mutual information. *ICML*, 2019, vol. 97, pp. 5837–5851.

[24] Pathak D., Agrawal P., Efros A.G., et al. Curiosity-driven exploration by self-supervised prediction. *IEEE CVPRW*, 2017, pp. 488–489.
DOI: https://doi.org/10.1109/CVPRW.2017.70

[25] Du Y., Han L., Fang M., et al. Liir: learning individual intrinsic reward in multi-agent reinforcement learning. *NIPS*, 2019, pp. 4403–4414.

[26] Amato C., Konidaris G.D., Cruz G., et al. Planning for decentralized control of multiple robots under uncertainty. *IEEE ICRA*, 2015, pp. 1241–1248.
DOI: https://doi.org/10.1109/ICRA.2015.7139350

[27] Bellemare M., Srinivasan S., Ostrovski G., et al. Unifying count-based exploration and intrinsic motivation. *NIPS*, 2016, pp. 1471–1479.

[28] Ostrovski G., Bellemare M.G., van den Oord A., et al. Countbased exploration with neural density models. *ICML*, 2017, pp. 2721–2730.

[29] Klissarov M., Islam R., Khetarpal K., et al. Variational state encoding as intrinsic motivation in reinforcement learning. *ICLR*, 2019, pp. 2–7.

[30] Stadie B.C., Levine S., Abbeel P. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv:1507.00814*.
DOI: https://doi.org/10.48550/arXiv.1507.00814

[31] Sutton R.S., Barto A.G. Reinforcement learning. An introduction. Cambridge, MIT Press, 2018.

[32] Lillicrap T.P., Hunt J.J., Pritzel A., et al. Continuous control with deep reinforcement learning. *arXiv:1509.02971*. DOI: https://doi.org/10.48550/arXiv.1509.02971

[33] Lowe R., Wu Y., Tamar A., et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *NIPS*, 2017, pp. 6382–6393.

[34] Kingma D.P., Welling M. Auto-encoding variational bayes. *arXiv:1312.6114*.
DOI: https://doi.org/10.48550/arXiv.1312.6114

Bolshakov V.E. — Assistant, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Sakulin S.A. — Cand. Sc. (Eng.), Assoc. Professor, Department of Information Systems and Telecommunications, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Alfimtsev A.N. — Dr. Sc. (Eng.), Head of the Department of Information Systems and Telecommunications, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Please cite this article in English as:

Bolshakov V.E., Sakulin S.A., Alfimtsev A.N. Multi-agent reinforcement learning using the collective intrinsic motivation. *Herald of the Bauman Moscow State Technical University, Series Instrument Engineering*, 2023, no. 4 (145), pp. 61–84 (in Russ.).
DOI: https://doi.org/10.18698/0236-3933-2023-4-61-84