DOI: https://doi.org/10.36910/6775-2524-0560-2021-44-06
УДК 622.276
Мальцев Антон Юрійович, к. ф.-м. н., доцент
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ЩОДО ЗАСТОСУВАННЯ ГЛИБОКОГО НАВЧАННЯ З ПІДКРІПЛЕННЯМ У СУЧАСНИХ СИСТЕМАХ

Мальцев А. Ю. Щодо застосування глибокого навчання з підкріпленням у сучасних системах. У статті розкрито принципи застосування глибокого навчання з підкріпленням у сучасних системах. Підкреслено, що у функції навчання з підкріпленням входить адаптація немарківської моделі прийняття рішень до ситуації, що склалася за рахунок аналізу передісторії процесу прийняття рішень, внаслідок чого підвищується якість прийнятих рішень. Описано принцип реалізації навчання з підкрі.пленням та схематично розкрито схему взаємодії агента з навколишнім середовищем. Для детального опису запропоновано використання 2D-задачі балансування полюсів, яку покладено в основу математичного аспекту. Наголошено, що у сучасних системах найбільш часто використовується дві схеми навчання з підкріпленням це метод часових різниць та метод Монте-Карло. Здійснено математичне обґрунтування кожного методу окремо та запропоновано архітектуру глибокої Q-мережі. Описано модельні та безмодельні методи, підкреслено, що модельні методи засновані на моделях навчання з підкріпленням, що змушують агента намагатися зрозуміти світ і створити модель для його подання. Безмодельні методи намагаються захопити дві функції, функцію переходу від станів і функцію винагороди, з цієї моделі агент має посилання і може планувати відповідно. Проте, зазначається, що немає необхідності вивчати модель, і агент може замість цього вивчати політику безпосередньо, використовуючи такі алгоритми, як Q-навчання або градієнт політики. Глибока Q-мережа, використовує згорткову нейронну мережу для прямої інтерпретації графічного представлення вхідного стану з навколишнім середовищем. Обґрунтовано, що глибоку Q-мережу можна розглядати як параметризовану мережу політики, яка постійно навчається для наближення оптимальної політики, а, математично, глибока Q-мережа використовує рівняння Беллмана для мінімізації функції втрат, що є ефективним для зниження часу. Однак використання нейронної мережі для наближення функції значення виявилося нестабільним і може призвести до розбіжностей через зміщення, що походить від корелятивних вибірок.
Ключові слова: штучний інтелект, машинне навчання, глибоке навчання з підкріпленням, система.

Maltsev Anton. On the application of deep learning with reinforcement in modern systems The article reveals the principles of application of deep learning with reinforcement in modern systems. It is emphasized that the function of reinforced learning includes the adaptation of the non-Markov model of decision-making to the situation that has developed due to the analysis of the prehistory of the decision-making process, which improves the quality of decisions. The principle of realization of training with reinforcement is described and the scheme of interaction of the agent with environment is schematically opened. For a detailed description, the use of a 2D pole balancing problem is proposed, which is the basis of the mathematical aspect. It is emphasized that in modern systems two schemes of reinforcement are most often used: the method of time differences and the method of Monte Carlo. The mathematical substantiation of each method is carried out separately and the architecture of a deep Q-network is offered. Model and non-model methods are described, it is emphasized that model methods are based on models of training with reinforcement, forcing the agent to try to understand the world and create a model for its presentation. Non-model methods try to capture two functions, the transition function and the reward function, from this model the agent has a link and can plan accordingly. However, it is noted that there is no need to study the model, and the agent can instead study the policy directly, using algorithms such as Q-learning or policy gradient. Deep Q-network, uses a convolutional neural network to directly interpret the graphical representation of the input state with the environment. It is substantiated that the deep Q-network can be considered as a parameterized policy network, which is constantly trained to approximate the optimal policy, and, mathematically, the deep Q-network uses the Bellman equation to minimize the loss function, which is effective in reducing time. However, the use of the neural network to approximate the value function proved to be unstable and could lead to discrepancies due to the bias resulting from correlative samples.
Key words: artificial intelligence, machine learning, deep learning with reinforcement, system.

Постановка проблеми. Протягом останніх кількох років глибоке навчання з підкріпленням стало основним підходом до вирішення багатьох завдань штучного інтелекту в різних сферах життя сучасної людини.

Навчання з підкріпленням зазвичай розглядається як загальна формалізація завдання прийняття рішень і глибоко пов'язане з динамічним програмуванням, оптимальним управлінням та теорією ігор. Проте його постановка проблеми майже не робить припущень щодо загальної моделі чи її структури і зазвичай передбачає, що середовище надається агенту у вигляді чорної скриньки. Це дозволяє застосовувати навчання з підкріпленням практично у всіх налаштуваннях та змушує розроблені алгоритми адаптуватися до багатьох видів викликів. Зазвичай повідомляється, що останні алгоритми навчання з підкріпленням можна переносити з одного завдання на інше без будь-яких змін для конкретних завдань і мало або зовсім без налаштування гіперпараметрів.

Оскільки об'єктом є стратегія, навчання з підкріпленням вважається підфайлом машинного навчання. Але замість того, щоб вчитися на основі даних, як це встановлено в класичних контрольованих та неконтрольованих моделях навчання, агент вчиться на досвіді взаємодії з

навколишнім середовищем. Будучи більш «природною» моделлю навчання, ця методика висуває нові вимоги, властиві лише навчанню з підкріпленням, такі як необхідність інтеграції та проблема затримки.

Аналіз останніх досліджень і публікацій. Навчання з підкріпленням, в даний час, набирає все більших обертів і стає все більш актуальним. На поточний момент існує великий обсяг літератури з питань застосування глибокого навчання з підкріпленням у різних сферах життя. Дані питання розглядали як вітчизняні, так і зарубіжні дослідники.

С. В. Білашенко, Н. Н. Шаповалова та О. Г. Рибальченко [1] здійснили дослідження архітектури глибокої згорткової нейронної мережі для розпізнавання зображень. В ході числового експерименту обґрунтували підбір оптимальних гіперпараметрів експлуатації моделі: швидкості навчання, кількості шарів у мережі, кількості нейронів у прихованому шарі.

Відносно без модельного навчання варто відзначити роботу В. В. Півошенко, М. С. Кулика, Ю. Ю. Іванова та А. С. Васюри [2]. Авторами розглянуто сучасний метод машинного навчання, який має назву навчання з підкріпленням. У задачах, які розв'язуються на основі взаємодії, найчастіше непрактично намагатися отримувати приклади необхідної поведінки інтелектуального програмного агента, які були б одночасно коректними та доречними для всіх ситуацій, оскільки наявні умови невизначеності, що виникають через неповноту інформації про навколишнє середовище та можливі дії інших ботів або людей.

В. М. Синєглазов та А. Т. Кот [3] визначити принципи розробки гібридних нейронних мереж ансамблевої структури. О. І. Чумаченко [4] окреслила особливості структурно-параметричного синтезу гібридних нейронних мереж.

Із зарубіжних авторів варто відзначити такі роботи як: Eoh, Gyuho & Park, Tae-Hyoung [5], J. Dornheim, N. Link, and P. Gumbsch [6], Kayakökü, Hakan & Guzel, Mehmet & Bostanci, Gazi Erkan & Medeni, Ihsan & Mishra, Deepti [7], Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D. [8], Vesal, S., Malakarjun Patil, S., Ravikumar, N., Maier, A. K. [9], Garau-Luis, Juan & Crawley, Edward & Cameron, Bruce. [10], Frikha, Mohamed & Gammar, Sonia & Lahmadi, Abdelkader & Andrey, Laurent [11], W. Haskell, and W. Huang [12], M. Rahman and H. Rashid [13] та інші.

Проте, враховуючи описані наукові набутки, за темою, питання розкриття принципів застосування глибокого навчання з підкріпленням у сучасних системах залишається відкритим та потребує детального опрацювання.

Постановка завдання. Розкрити принципи застосування глибокого навчання з підкріпленням у сучасних системах.

Викладення основного матеріалу дослідження. Навчання з підкріпленням – це навчання проб та помилок:

1) шляхом безпосередньої взаємодії з оточенням;
2) самонавчання на протязі всього часу;
3) досягнення мети визначення.

Зокрема, навчання з підкріпленням визначає будь-яку особу, яка приймає рішення, як агента, а все, що знаходиться поза агентом, як середовище. Взаємодія між агентом та середовищем описується за допомогою трьох суттєвих елементів: стану s, дії a та винагороди r, як показано на рис. 1.

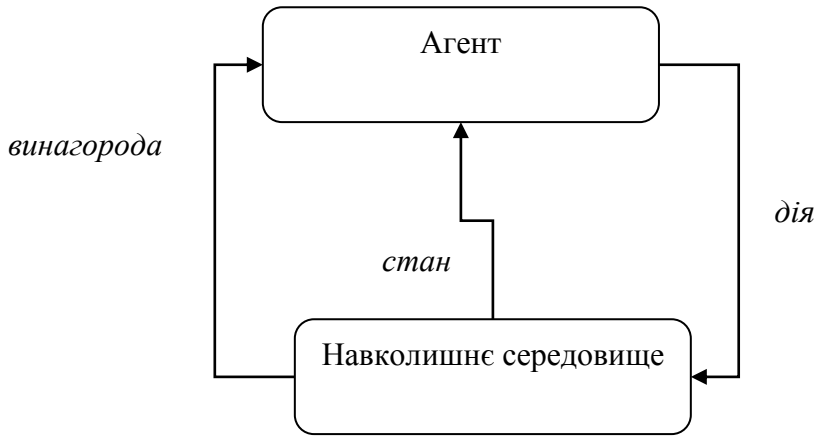

Рис. 1. Схема взаємодії агента з навколишнім середовищем

Стан навколишнього середовища на етапі t позначається як s_t. Тим самим агент перевіряє s_t і виконує відповідну дію за адресою. Потім середовище змінює свій стан s_t на $s_t + 1$ і надає агенту винагороду за зворотний зв'язок $r_t + 1$. Наприклад, рисунок 2 ілюструє одну з найдавніших проблем навчання з підкріпленням, 2D-задачу балансування полюсів. У цій задачі стан середовища на часовому кроці t може бути представлений набором $s_t = [x_c, v_c, \alpha_p, \omega_p]\, t$, де x_c позначає x-координату візка в декартовій системі координат Oxy, v_c представляє швидкість візка вздовж колії, α_p-кут, що створюється полюсом та віссю Oy, а ω_p вказує кутову швидкість полюса навколо центру I.

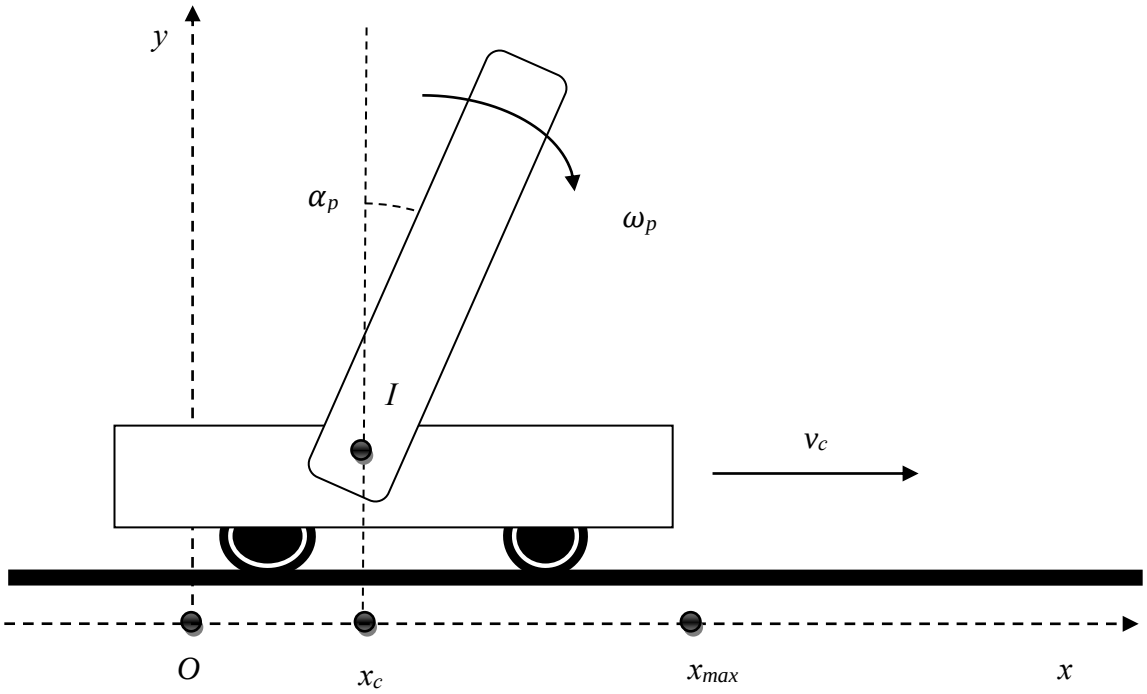

Рис. 2. 2D-задача балансування полюсів

Агент може виконати дві можливі дії на кожному часовому кроці t: прикласти одиницю сили ($|\vec{F}| = 1$) до візка вздовж осі Ox зліва направо при $a_t = \vec{F}$ або справа наліво при $a_t = -\vec{F}$. Агенту надається винагорода за зворотний зв'язок $r_{t+1} = +1$ за кожну дію, яка може утримувати полюс вертикально, і $r_{t+1} = 0$ в іншому випадку. Тому мета агента – якомога довше утримувати полюс вертикально і в кінцевому підсумку максимізувати накопичену винагороду за зворотний зв'язок.

Як правило, взаємодія між агентом та середовищем може бути представлена низкою станів, дій та винагород: $s_0, a_0, r_1, s_1, a_1, \ldots, r_n, s_n$. Хоча n може наближатися до нескінченності, n на практиці часто обмежується, визначаючи кінцевий стан $s_n = s_T$. У цьому випадку ряд станів, дій та винагород від початкового стану до термінального стану називається епізодом. Наприклад, у завданні балансування полюсів є можливість визначити термінальний стан так, ніби $|\alpha_p| > 10^{\circ}$ або $|x_c| > X_{max}$.

Наступним кроком є оформлення рішення агента шляхом визначення концепції політики. Політика π – це функція відображення з будь-якого сприйнятого стану s на дію, прийняту з цього стану. Політика є детермінованою, якщо ймовірність вибору дії a з $s: p\,(a \mid s) = 1$ для всіх станів s. На відміну від цього, політика є стохастичною, якщо існує стан s, такий що $p\,(a \mid s) < 1$. У будь-якому випадку є можливість визначити політику π як розподіл ймовірностей дій-кандидатів, які будуть обрані з певного стану:

$$\pi = \psi(s) = \left\{ p(a_i|s) \,\middle|\, \forall a_i \in \Delta_\pi \Lambda \sum_i p(a_i|s) = 1 \right\}$$

де Δ_π представляє всі дії – (простір дій) політики π. Для наочності вважаємо, що простір дій дискретний, оскільки безперервний випадок можна прямо вивести за допомогою інтегральних позначень. Крім того, припускаємо, що наступний стан s_{t+1} та винагорода за зворотний зв'язок r_{t+1} повністю визначаються поточною парою стану-дії (s_t, a_t) незалежно від історії. Будь-яка проблема навчання з підкріпленням задовольняє цій умові. Отже, динаміка (модель) задачі навчання з

підкріпленням повністю уточнюється, задаючи всі ймовірності переходу $p\,(a_i \mid s)$. Детермінована політика π_d:

$$\pi_d = \psi_d(s) = \begin{cases} 1, a_i = a(s) \,\Lambda\, a(s) \in \Delta_{\pi_d} \\ 0, \forall a_i \in \Delta_{\pi_d} \,\Lambda\, a_i \neq\ a(s) \end{cases}$$

де $a(s)$ позначає дію, вжиту у стані s. Детермінована політика ефективна у практичному застосуванні, оскільки вона має передбачувану поведінку, що є вирішальним чинником для розробки ефективного алгоритму навчання з підкріпленням. На практиці є можливість вивести детерміновану політику π_d зі стохастичної політики π, використовуючи наступне правило:

$$R_1 : \pi \to \pi_d = \psi_d(s) = \begin{cases} 1, a_i = a_j \,\Lambda\, \mathrm{j} = arg\,max\,\pi(s, a_k) \\ 0, \forall a_i \in \Delta_\pi \,\Lambda\, a_i \neq\ a_j \end{cases}$$

де $\pi(s, a_k)$ позначає ймовірність здійснення дій $a_k \in \Delta_\pi$ у стані s за допомогою політики π та $\Delta_{\pi_d} = \Delta_\pi$.

Спочатку агенту призначається довільна політика π_0. Він коригує політику π_0, щоб покращити себе, взаємодіючи з навколишнім середовищем у формі навчання проб та помилок. У цьому відношенні політика π_{t+1} краща за політику π_t і позначається як $\pi_{t+1} > \pi_t$. Тому, з часом, покращується ряд політик:

$$\pi_0 < \pi_1 < \cdots < \pi_t < \pi_{t+1} < \cdots < \pi^*$$

Цей процес, названий поліпшенням політики, повторюється до тих пір, поки агент не зможе знайти політику, кращу за оптимальну політику π^*. Однак за цим визначенням невідомо, як порівняти дві політики та вирішити, яка з них краща.

У сучасних системах найбільш часто використовується дві схеми навчання з підкріпленням це метод часових різниць та метод Монте-Карло.

Останній метод оцінює функцію значення шляхом багаторазового генерування епізодів та запису середнього прибутку в кожному стані або кожній парі стан-дія. Тому функція стан-значення обчислюється як:

$$\mathrm{V}_\pi^{MC}(s) = \lim_{i \to +\infty} \mathbb{E}\left[r^i(s_t) | s_t = s, \pi\right]$$

де $r^i(s_t)$ позначає спостереження повернення на стан s_t в епізоді i. Аналогічно, маємо функцію значення пари станів дії:

$$\mathrm{Q}_\pi^{MC}(s, a) = \lim_{i \to +\infty} \mathbb{E}\left[r^i(s_t, a_t) | s_t = s, a_t = a, \pi\right]$$

Метод Монте-Карло не вимагає ніяких знань щодо ймовірності переходу, тобто метод Монте-Карло не є модельним. Однак цей підхід зробив два суттєвих припущення для забезпечення збіжності:

1) кількість епізодів велика;

2) кожен стан і кожну дію необхідно відвідувати значну кількість разів.

Щоб зробити це «дослідження» можливим, використовуємо стратегію жадібності для вдосконалення політики:

$$R_3 : \pi \mapsto \pi' = \Psi'(s) = \begin{cases} 1 - \in + \dfrac{\in}{|\Delta_\pi(s)|}, a_i = a_j \,\Lambda\, j = arg\,max\, Q_\pi(s, a_k) \\ \dfrac{\in}{|\Delta_\pi(s)|}, \forall \Delta_\pi \in a_i \neq a_j \end{cases}$$

де $|\Delta_\pi(s)|$ позначає кількість дій-кандидатів, задіяних у стані s та $0 << 1$. Загалом, алгоритми Монте-Карло поділяються на дві групи: включені до політики та поза політикою. У методах на основі політики використовуємо політику π як для оцінки, так і для дослідження. Тому політика π має бути стохастичною або м'якою. На відміну від цього, офполітика використовує іншу політику $\pi' \neq \pi$ для створення епізодів, а отже, π може бути детермінованою. Метод політики є більш стабільним при роботі з безперервними станами і при використанні спільно з апроксиматором функцій (наприклад, нейронними мережами).

Подібно до методу Монте-Карло, метод часових різниць також є навчанням на основі досвіду (безмодельний метод). Однак, на відміну від Монте-Карло, метод часових різниць не чекає до кінця

епізоду, щоб оновитися. Він оновлює кожен крок епізоду, використовуючи одноетапне рівняння Беллмана[1], а отже, можливо, забезпечує більш швидку конвергенцію:

$$U_1: V^i(s_t) \leftarrow \alpha V^{i-1}(s_t) + (1-\alpha)\left(r_{t+1} + \gamma V^{i-1}(s_{t+1})\right)$$

де α – розмір кроку і $0 < \alpha < 1$. Метод часових різниць використовує попередні оцінені значення V^{i-1} для оновлення поточних значень V^i, який відомий як метод завантаження. Метод часових різниць також поділяється на дві категорії: контроль часових різниць на основі політики (Sarsa[2]) та контроль часових різниць поза політикою (навчання Q[3]). В Sarsa алгоритм оцінює функцію значення пари стану-дії на основі:

$$U_2: Q^i(s_t, a_t) \leftarrow \alpha Q^{i-1}(s_t, a_t) + (1-\alpha)\left(r_{t+1} + \gamma Q^{i-1}(s_{t+1}, a_{t+1})\right)$$

З іншого боку, Q-навчання використовує одноетапну оптимальність рівняння Беллмана для виконання оновлення, тобто Q-навчання безпосередньо апроксимує функцію значення оптимальної політики:

$$U_3: Q^i(s_t, a_t) \leftarrow \alpha Q^{i-1}(s_t, a_t) + (1-\alpha)\left(r_{t+1} + \gamma \max_{a^j{}_{t+1}} Q^{i-1}\left(s_{t+1}, a^j{}_{t+1}\right)\right)$$

Оператор max замінює детерміновану політику. Це чітко пояснює, чому Q-навчання не відповідає політиці.

На практиці методи Монте-Карло і часових різниць часто використовують структуру пам'яті таблиці (табличний метод) для збереження функції значення кожного стану або кожної пари стан-дія. Це робить їх неефективними через брак пам'яті при вирішенні складних задач, де кількість станів велика. Тому архітектура актора-критика (АК) розроблена таким чином, щоб подолати це обмеження. Зокрема, АК включає дві окремі структури пам'яті для агента: актора та критика. Структура актора використовується для вибору відповідної дії відповідно до спостережуваного стану та передачі до структури критики для оцінки. Критична структура використовує таку помилку часових різниць для вирішення майбутньої тенденції обраної дії:

$$\delta(a_t) = \beta\left(r_{t+1} + \gamma V(s_{t+1})\right) - (1-\beta)V(s_t)$$

де $0 < \beta < 1$; а якщо $\delta(a_t) > 0$, тенденція до вибору дії в майбутньому висока і навпаки. Крім того, АК може бути включеним до політики або поза політикою залежно від деталей впровадження.

Глибоке навчання з підкріпленням – це широкий термін, який вказує на комбінацію між глибоким навчанням та навчанням з підкріпленням. Глибока Q-мережа, використовує згорткову нейронну мережу для прямої інтерпретації графічного представлення вхідного стану s з навколишнім середовищем. Вихідні дані глибокої Q-мережі надають Q-значення всіх можливих дій $a \in \Delta_\tau$, зроблених у стані s, де Δ_τ позначає простір дій. Тому глибоку Q-мережу можна розглядати як мережу політики τ, параметризовану β, яка постійно навчається для наближення оптимальної політики. Математично глибока Q-мережа використовує рівняння Беллмана для мінімізації функції втрат $L(\beta)$:

$$L(\beta) = \mathbb{E}\left[\left(r + \gamma \max_{a'} Q(s', a'|\beta) - Q(s, a|\beta)\right)^2\right]$$

Однак використання нейронної мережі для наближення функції значення виявилося нестабільним і може призвести до розбіжностей через зміщення, що походить від корелятивних вибірок. Щоб зробити зразки некорельованими, є можливість використовувати цільову мережу τ', параметризовану β', яка оновлюється на кожних N кроках від мережі оцінки τ.

---

1 https://uk.wikipedia.org/wiki/%D0%A0%D1%96%D0%B2%D0%BD%D1%8F%D0%BD%D0%BD%D1%8F_%D0%91%D0%B5%D0%BB%D0%BB%D0%BC%D0%B0%D0%BD%D0%B0

2 https://uk.wikipedia.org/wiki/%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC_SARSA

3 https://uk.wikipedia.org/wiki/Q-%D0%BD%D0%B0%D0%B2%D1%87%D0%B0%D0%BD%D0%BD%D1%8F

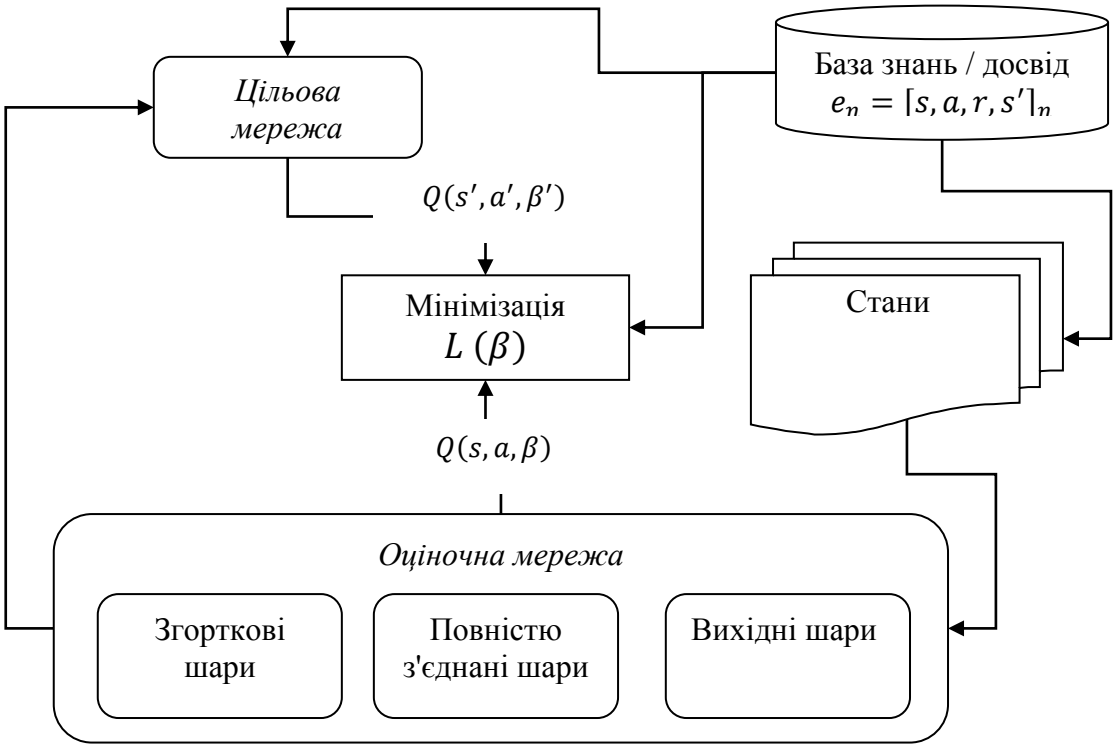

Рис. 3. Архітектура глибокої Q-мережі

Крім того, згенеровані зразки зберігаються в пам'яті повторного відтворення. Потім зразки вибираються випадковим чином та подаються у навчальний процес, як описано на рис. 3.

Безмодельний глибокий метод навчання з підкріпленням допомагає вирішити багато складних проблем як в одноагентних, так і в багатоагентних областях. Однак ця категорія методів вимагає величезної кількості зразків та тривалого часу навчання для досягнення хорошої продуктивності. Модельні методи демонструють ефективність з точки зору ефективності вибірки, переносимості та загальності у різних проблемах, використовуючи одномоментні та багатоагентні моделі.

Останні досягнення в галузі архітектури «людина в режимі циклу» можна об'єднати з глибоким навчанням з підкріпленням для інтеграції людей та автономних агентів щодо вирішення складних проблем. У звичайному режимі «людина в циклі» агенти протягом певного періоду автономно виконують покладені на них завдання, потім зупиняються і чекають людських команд, перш ніж продовжувати у такий спосіб, обмежений швидкістю. У режимі «людина в курсі» агенти виконують свої завдання автономно до завершення, а людина в ролі контролю або нагляду залишає за собою можливість втручатися в операції, які здійснюють агенти. Архітектура, заснована на системі «людина-на-циклі», може бути повністю автономною, якщо керівники дозволяють агентам виконувати завдання повністю самостійно.

Висновки і перспективи подальших досліджень. У роботі досліджено принципи застосування глибокого навчання з підкріпленням у сучасних системах. Глибоке навчання з підкріпленням значно полегшило автономність, що дозволяє розгортати багато додатків у робототехніці або автономних автомобілях. Однак найпоширенішим недоліком глибоких моделей навчання з підкріпленням є здатність взаємодіяти з людиною за допомогою технологій об'єднання людей і машин. У складних та суперечливих умовах існує гостра потреба в людському інтелекті, поєднаному з технологіями, тому що люди самі по собі не можуть утримати цей обсяг, а самі машини не можуть дати творчі відповіді при появі нових ситуацій.

Перспективи подальших досліджень ґрунтуються на масштабних спостереженнях з використанням модельних підходів або поєднання елементів модельного планування та безмодельної політики.

Список бібліографічного опису.

1. Білашенко С. В. Розпізнавання зображень за допомогою згорткових нейронних мереж з використанням бібліотеки Keras / С. В. Білашенко, Н. Н. Шаповалова, О. Г. Рибальченко // Гірничий вісник : науково-технічний збірник. Кривий Ріг, 2018. Вип. 103. С. 148–154. – Бібліогр.: 12 назв. – DOI: 10.31721/2306-5435-2018-1-103-148-154.
2. Півошенко В. В. Аналіз та експериментальне дослідження методу безмодельного навчання з підкріпленням / В. В. Півошенко, М. С. Кулик, Ю. Ю. Іванов, А. С. Васюра // Вісник Вінницького політехнічного інституту. 2019. № 3. С. 40-49. – Режим доступу: http://nbuv.gov.ua/UJRN/vvpi_2019_3_7.
3. Синєглазов В. М., Кот А. Т. Розробка гібридних нейронних мереж ансамблевої структури / В. М.Синєглазов, А. Т. Кот // 2021. – Режим доступу. – http://webcache.googleusercontent.com/search?q=cache:u3aSPdfsdlUJ:journals.uran.ua/eejet/article/download/225301/226945/518366+&cd=9&hl=uk&ct=clnk&gl=ro
4. Чумаченко О.І. Структурно-параметричний синтез гібридних нейронних мереж. – Кваліфікаційна наукова праця на правах рукопису. Дисертація на здобуття наукового ступеню доктора технічних наук за спеціальністю 05.13.23 – «Системи та засоби штучного інтелекту» – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2019. 663 с.

References.

1. Eoh, Gyuho & Park, Tae-Hyoung. (2021). Cooperative Object Transportation Using Curriculum-Based Deep Reinforcement Learning. Sensors. 21. 10.3390/s21144780. Accessed: August, 6, 2021.
2. J. Dornheim, N. Link, and P. Gumbsch, "Model-Free Adaptive Optimal Control of Sequential Manufacturing Processes Using Reinforcement Learning," arXiv.org, 2019. [Electronic resource]. Available: https://arxiv.org/abs/1809.06646v1. Accessed: August, 6, 2021.
3. Kayakökü, Hakan & Guzel, Mehmet & Bostanci, Gazi Erkan & Medeni, Ihsan & Mishra, Deepti. (2021). A Novel Behavioral Strategy for RoboCode Platform Based on Deep Q-Learning. Complexity. 2021. 1-14. 10.1155/2021/9963018. Accessed: August, 6, 2021.
4. Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D. (2017). Dermoscopic Image Segmentation via Multistage Fully Convolutional Networks. IEEE Transactions on Biomedical Engineering, 64 (9), 2065–2074. doi: https://doi.org/10.1109/tbme.2017.2712771 Accessed: August, 6, 2021.
5. Vesal, S., Malakarjun Patil, S., Ravikumar, N., Maier, A. K. (2018). A Multi-task Framework for Skin Lesion Detection and Segmentation. OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, 285–293. doi: https://doi.org/10.1007/978-3-030-01201-4_31 Accessed: August, 6, 2021.
6. Garau-Luis, Juan & Crawley, Edward & Cameron, Bruce. (2021). Evaluating the progress of Deep Reinforcement Learning in the real world: aligning domain-agnostic and domain-specific research. Accessed: August, 6, 2021.
7. Frikha, Mohamed & Gammar, Sonia & Lahmadi, Abdelkader & Andrey, Laurent. (2021). Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey. Computer Communications. 178. 98-113. 10.1016/j.comcom.2021.07.014. Accessed: August, 6, 2021.
8. W. Haskell, and W. Huang, "Stochastic Approximation for Risk-Aware Markov Decision Processes", Arxiv.org, 2018. [Electronic resource]. Available: https://arxiv.org/pdf/1805.04238.pdf. Accessed: August, 6, 2021.
9. M. Rahman and H. Rashid, "Implementation of Q Learning and Deep Q Network for Controlling a Self-Balancing Robot Model," ArXiv.org, 2018. [Electronic resource]. Available: https://arxiv.org/pdf/1807.08272.pdf . Accessed: August, 6, 2021.