DOI 10.31891/2307-5732-2022-315-6-111-114
УДК 004.93, 004.8

КРИВЕНЧУК Юрій
Національний університет "Львівська політехніка"
https://orcid.org/0000-0002-2504-5833
e-mail: yurii.p.kryvenchuk@lpnu.ua
ПЕТРЕНКО Дмитро
Національний університет "Львівська політехніка"
https://orcid.org/0000-0003-3720-9038
e-mail: dmytro.o.petrenko@lpnu.ua

СИСТЕМА СТАБІЛІЗАЦІЇ ПОЛОЖЕННЯ ДРОНУ З ВИКОРИСТАННЯМ НАВЧАННЯ З ПІДКРІПЛЕННЯМ

В роботі наведено результати дослідження теми використання алгоритмів навчання з підкріпленням у системах керування дронами, з метою покращення якості та збільшення швидкодії такого типу систем, їх поширення та впровадження у відповідні сфери в Україні. Виділено та описано такі основні етапи: огляд систем навчання з підкріпленням, визначення основних параметрів, за якими проводитиметься навчання, порівняння результатів, отриманих на різних мережах. Після проведення аналізу результатів було виявлено, що створення системи стабілізації положення дрону з використанням навчання з підкріпленням є актуальним та доцільним завданням на сьогодні, а найбільш ефективним інструментом для цього є використання навчання з підкріпленням в поєднанні з глибинними нейронними мережами.

Ключові слова: керування дроном, навчання з підкріпленням, глибинні нейронні мережі.

KRYVENCHUK Yurii, PETRENKO Dmytro
Lviv Polytechnic National University

CREATION OF DRONE STABILIZATION SYSTEM USING REINFOCEMENT LEARNING

The paper presents the results of research on the topic of using reinforcement learning algorithms in drone control systems, with the aim of improving the quality and increasing the speed of this type of systems, their distribution and implementation in the relevant areas in Ukraine. The following main stages are highlighted and described: review of training systems with reinforcement, determination of the main parameters according to which training will be conducted, comparison of results obtained on different networks. After analyzing the results, it was found that creating a drone position stabilization system using reinforcement learning is a relevant and appropriate task today, and the most effective tool for this is the use of reinforcement learning in combination with deep neural networks. Drone settings contain many parameters. Selecting these parameters and learning how to control the drone takes a lot of time. Drone pilots usually rely on their own experience and intuition when flying. This article examines the use of deep reinforcement learning to assist the pilot in typical or complex situations, as well as to extend the life of drones and avoid out-of-state situations. A general model represents an algorithm with input parameters equal to those required to represent the possible states and output parameters of the system sufficient to describe the possible actions. The algorithm automatically selects different models according to different parameters. It is determined that the algorithm can successfully start work with a low-efficiency model template and show good model performance and adjust the parameters of the number of layers, policy, entropy ratio, etc. This shows the potential for further application of these algorithms for designing drones. The result obtained during the execution of this work was a system that allows to simplify the process of choosing a deep learning algorithm with reinforcement in any created simulation environment for an agent of any complexity simulated in the Unreal Engine 4 game engine. The drone setup master must correctly formulate the task that the drone must perform, determine the main requirements for performance and the main possible bad options for performance. As a result of training, the drone will be able to stabilize itself from different positions, which will help to avoid emergency situations. This work can be widely applied in modern realities.

Keywords: reinforcement learning, quadcopters, drones, deep learning.

Постановка проблеми

На теперішній час дрони відіграють значну роль як у повсякденному житті, так і у вирішенні вузькопрофільних завдань. Часом налаштування дрону займає значний час та потребує професіоналів зі знанням алгоритмів налаштування. Натомість використовуючи систему автоматичного навчання підбору параметрів кожен дрон буде налаштовано згідно своїх унікальних характеристик та задач, за які він буде відповідати. Це дозволить автоматизувати та пришвидшити процес налаштування дрону, а від оператора вимагатиметься лише правильна постановка завдання для дрона.

Дану розробку доцільно буде використовувати у сучасних реаліях, коли часто необхідно, щоб дрон міг нерухомо стабілізуватись на одному місці для подальшої роботи.

Розроблену систему буде використано для створення системи автоматичного навчання дрона вирішенню поставленої задачі. Така система зможе у короткий термін визначити ключові параметри та їх вплив на роботу дрона і оптимізувати алгоритми роботи дрона для найоптимальнішого розв'язання задачі.

Аналіз останніх джерел

У роботі Джеміна Хванбо [1] навчання дрона відбувається за допомогою двох глибинних нейронних мереж у процесі навчання з підкріпленням: мережа значень та мережа політики. Як вхідні параметри використовуються значення матриці обертання а також вектори лінійних та кутових швидкостей. У даній реалізації було використано метод глибоко детермінізованої політики оптимізації DDPG. В якості політики оптимізації було використано метод природного градієнтного спуску.

В публікації [2] автори зосереджуються на автономному керуванні дроном протягом заданого маршруту. Юньлонг Сонг та Матс Штайнвег створили безмодельну систему з нейронною мережею. Вони

формулют задачу планування траєкторії оптимального часу в системі навчання з підкріпленням. З цією метою вони моделюють завдання за допомогою марковського процесу прийняття рішень з нескінченним горизонтом (MDP). У цій роботі вони представили заснований на навчанні метод навчання політики нейронної мережі, яка може генерувати майже оптимальні за часом траєкторії через кілька воріт для квадрокоптерів. Вони продемонстрували сильні сторони такого підходу, включаючи продуктивність, майже оптимальну за часом, здатність обробляти великі зміни доріжок, а також масштабованість і можливість узагальнення для вирішення великомасштабних макетів випадкових доріжок, зберігаючи ефективність обчислень. Автори підтвердили згенеровану траєкторію за допомогою фізичного квадрокоптера та досягли агресивного польоту на швидкостях до 60 км/год. Ці висновки свідчать про те, що глибока RL має потенціал для створення адаптивних оптимальних за часом траєкторій для квадрокоптерів і заслуговує на подальше дослідження.

У роботі [3] Вікторія Дж. Ходж, Річард Хокінс і Роб Александр створили алгоритм, що ґрунтується на алгоритмі глибокого підкріплення Proximal Policy Optimization (PPO) із поступовим навчанням навчальної програми для покращення навчання та рекурентним рівнем LSTM, щоб дозволити агенту запам'ятати, де він був, і повернутися назад, коли він застряг. Алгоритми навчання з глибоким підкріпленням здатні керувати досвідом навчання для реальних проблем, що робить їх ідеальними для їхнього завдання. Вони навмисно налаштували наш алгоритм так, щоб він був адаптованим і потенційно міг працювати в складних і динамічних середовищах.

У публікації [4] автори запропонували механізм передачі даних на основі навчання з підкріпленням (RL) для досягнення надійного зв'язку дронів у мережі дронів із стільниковим зв'язком. Використовуючи структуру Q-навчання, вони запропонували гнучкий спосіб прийняття рішень механізму передачі для заданої траєкторії польоту. Також показали, як мережа може мати компроміс між кількістю передач і потужністю отриманого сигналу, регулюючи відповідні ваги цих величин у функції винагороди. Результати моделювання показали, що запропонований підхід може значно зменшити кількість передач, зберігаючи надійний зв'язок, порівняно з базовою схемою передач, у якій дрон завжди підключається до найсильнішої комірки. Є кілька потенційних напрямків для майбутніх досліджень. По-перше, існуюча структура розглядає мобільність дронів у 2D. Природним розширенням буде забезпечення мобільності 3D-дронів. По-друге, район випробувань і маршрути польотів, які розглядаються в цій роботі, досить обмежені.

У документі [5] проведено загальний огляд можливостей використання алгоритмів навчання з підкріпленням для керування та роботи безпілотних літальних апаратів (БПЛА) розглядаються три основні проблеми, з якими стикаються БПЛА: (I) планування шляху, (II) навігація та (III) контроль. Кожен із цих елементів містить багато підзавдань, які потребують високого рівня контролю, щоб функціонувати належним чином. Алгоритми навчання з підкріпленням використовуються для допомоги в навігації в невідомих середовищах, які не мають математичної моделі, придатної для їх опису. Спочатку у статті розглядаються кілька алгоритмів навчання з підкріпленням, пов'язаних з БПЛА та його поведінкою при навчанні. Потім обговорюється планування шляху дронів, навігація та керування за допомогою згаданих підходів до навчання з підкріпленням.

У дослідженні [6] автори порівняли навчання з підкріпленням для дискретних і безперервних просторів дії в уникненні перешкод за допомогою дрона. Застосований ними метод із використанням мережі сегментації для простору безперервної дії навчається за допомогою мережі актор-критик із парадигми RL. Значною перевагою цього є, звичайно, те, що ручне маркування не потрібне, що економить робочу силу та час. Продуктивність моделі сегментації на основі U-net також значно покращена. Питання для RL з точки зору проблем автономної навігації полягало в тому, як можна мінімізувати розрив між реальним і навчальним середовищем. За допомогою серії експериментів вони демонструють, що розроблена та навчена модель здійснила успішні польоти не лише в навченому середовищі, але й у деяких змінених середовищах. Дане дослідження може бути першою спробою, коли пілот-людина здійснив перегони безпілотника за допомогою алгоритму та оцінки продуктивності між ними.

У дисертації [7] досліджуються різні алгоритми навчання з підкріпленням і знаходять алгоритм RL, який підходить для цього завдання. Простір дії задачі безперервний, оскільки всі чотири моторні команди безперервні. Там, де Q-навчання не здатне обробляти безперервні простори дій, DDPG і його варіантний алгоритм TD3 здатні обробляти ці простори дій. TD3 — це вдосконалена версія DDPG з трьома додатковими функціями щодо DDPG:

1. Обрізане подвійне Q-навчання
2. Затримка оновлень політики
3. Згладжування цільової політики

Перший прийом запобігає надмірним оцінкам у Q-функції. Другий прийом допомагає стабільності в тренуваннях. Третій трюк ускладнює використання помилок у Q-функції, додаючи шум до дій.

Структура функції винагороди вибирається щільною, а не розрідженою. Безперервний характер поточних завдань (відстеження точки шляху) робить цей варіант кращим. Крім того, під час навчання для виконання завдань, для яких цільові стани навряд чи відбудуться «випадково», рідкісні винагороди не є кращими. Для завдання наведення функція винагороди суттєво штрафує за помилку позиції та додає невелику від'ємну винагороду за похідну за часом команд двигуна. Для завдання контролю позиції реалізовано два типи функції винагороди: тип 1 штрафує як помилку позиції, так і позицію, тоді як тип 2 штрафує лише помилку позиції.

У статті [8] було представлено метод розробки агентів глибинного навчання з підкріпленням для безперервного детального керування безпілотником для отримання високоякісних кадрів фронтального виду людини. За допомогою HPID було розроблено реалістичне середовище моделювання. Це середовище було розширено за допомогою 3D-моделей разом із технікою вирівнювання обличчя та деформації, щоб дозволити симулювати ефекти безперервних команд керування, подолавши обмеження, пов'язані з обмеженою кількістю поз обличчя, які містяться в цьому наборі даних. Також був запропонований відповідний підхід до формування винагороди для підвищення стабільності використовуваного безперервного методу RL. Окрім виконання безперервного контролю, було продемонстровано, що CDC можна також ефективно поєднувати з середовищами моделювання, які підтримують лише дискретні команди керування, покращуючи точність керування, навіть у цьому випадку. Нарешті, запропонований підхід порівнювали як з агентом RL, який виконує дискретне керування, так і з традиційним контролером, який безпосередньо використовує вихід глибокої моделі, яка виконує оцінку пози. Експериментально доведено, що запропонований підхід покращує контроль.

Автори у своїй роботі [9] досліджували проблему планування траєкторії для групи DBS у непередбачуваних динамічних середовищах. У розглянутій системі DBS спільно дрони літають навколо розглянутого середовища, щоб надавати наземним користувачам послугу зв'язку висхідної лінії зв'язку на вимогу. Вони сформулювали досліджувану задачу в оптимізаційній постановці та запропонували алгоритм VD-RL для її вирішення. Запропонований алгоритм VD-RL змушує DBS самостійно оновлювати свої індивідуальні стратегії для досягнення максимальної командної корисності DBS, поділяючи лише свою корисність і цінність іншим DBS. Щоб покращити швидкість конвергенції алгоритму VD-RL у невидимих середовищах, вони також запропонували метод метанавчання для оптимізації ініціалізацій у рішенні VD-RL. Результати моделювання показують, що запропонований алгоритм VD-RL з механізмом метанавчання перевершує традиційні алгоритми MARL.

У статті [10] розглядаються алгоритми машинного навчання з доповненням та метаевристичні алгоритми. Результатом дослідження стала можливість комбінувати різні алгоритми DRL з генетичним алгоритмом і автоматично вибирати найкращі моделі DRL для вирішення. Під час експерименту популяція з 30 агентів CartPole була проаналізована у середовищі віртуального тренажерного залу. Результатом експерименту став вибір одного алгоритму DRL із вибірки з деякими відмінностями в гіперпараметрах моделі.

Метою роботи є створення системи що буде здатна стабілізувати дрон на основі алгоритмів глибинного навчання з підкріпленням.

Виклад основного матеріалу

Етапи процесу апроксимації віку

У навчанні з підкріпленням створюється метод винагороди за бажану поведінку та покарання за негативну поведінку. Цей метод призначає позитивні значення бажаним діям, щоб заохотити агента, і негативні значення для небажаної поведінки. Це програмує агента прагнути довгострокової та максимальної загальної винагороди для досягнення оптимального рішення.

Час, необхідний для того, щоб навчання проходило належним чином за допомогою цього методу, може обмежити його корисність і затрачати комп'ютерні ресурси. У міру того, як середовище навчання стає все більш складним, збільшуються вимоги до часу та обчислювальних ресурсів.

Замість того, щоб посилатися на конкретний алгоритм, область навчання з підкріпленням складається з кількох алгоритмів, які використовують дещо різні підходи. Відмінності в основному зумовлені їхніми стратегіями вивчення навколишнього середовища [3].

• Стан-дія-винагорода-стан-дія (SARSA). Цей алгоритм навчання з підкріпленням починається з надання агенту того, що відомо як політика. По суті, політика — це ймовірність, яка повідомляє їй про шанси певних дій, які призведуть до винагород або вигідних станів.

• Q-навчання. Цей підхід до навчання з підкріпленням використовує протилежний підхід. Агент не отримує жодної політики, а це означає, що його дослідження навколишнього середовища є більш самостійним.

• Глибокі Q-мережі. Ці алгоритми використовують нейронні мережі на додаток до методів навчання з підкріпленням. Вони використовують самостійне дослідження середовища навчання з підкріпленням. Майбутні дії засновані на випадковій вибірці минулих корисних дій, засвоєних нейронною мережею.

Загальна задача навчання з підкріпленням формалізується як стохастичний процес керування дискретним часом, де агент взаємодіє зі своїм середовищем таким чином: агент починає, у заданому стані у своєму середовищі $s_0 \in S$, збираючи початкове спостереження $\omega_0 \in \Omega$. На кожному кроці часу t агент повинен виконати дію $a_t \in A$. Як показано на рисунку 1, з цього випливає три наслідки:

• Агент отримує винагороду $rt \in R$.

• Стан переходить до $st+1 \in S$.

• Агент отримує спостереження $\omega t+1 \in \Omega$.

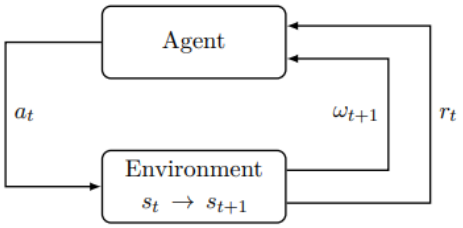

Рис. 1. Принцип роботи системи

Висновок

В даній статті було проведено дослідження можливостей використання алгоритмів глибинного навчання з підкріпленням для систем стабілізації дронів. Було визначено сильні та слабкі сторони таких алгоритмів. Основним завданням було створити систему, що могла б стабілізувати або повернути дрон у задане положення. Для цього спочатку було проведено комплексний аналіз проблеми та шляхів її вирішення. Оглянуто існуючі розробки у даній галузі. Проведено аналіз існуючих алгоритмів глибинного навчання з підкріпленням.

При дослідженні шляхів якими можна вирішити поставлену задачу було використано алгоритм DDPG так як він показав найбільшу ефективність. В подальшому планується провести більш комплексну порівняльну характеристику різних алгоритмів глибинного навчання з підкріпленням.

References

1. Jemin Hwangbo, Inkyu Sa, Roland Siegwart, Marco Hutter. Control of a Quadrotor with Reinforcement Learning. *IEEE Robotics and Automation Letters*, Volume: 2, Issue: 4, October 2017, doi: 10.1109/LRA.2017.2720851.

2. Song Y., Steinweg M., Kaufmann E., Scaramuzza D. Autonomous Drone Racing with Deep Reinforcement Learning. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, 2021. doi: 10.1109/IROS51168.2021.9636053.

3. HODGE Victoria J., HAWKINS Richard, ALEXANDER Rob. Deep reinforcement learning for drone navigation using sensor data. Neural Computing and Applications, 2021, 33.6: 2015-2033.

4. CHEN Yun et al. Efficient drone mobility support using reinforcement learning. 2020 IEEE wireless communications and networking conference (WCNC). IEEE, 2020. p. 1-6. doi: 10.1109/WCNC45663.2020.9120595.

5. Azar A. T., Koubaa A., Ali Mohamed N., Ibrahim H. A., Ibrahim Z. F., Kazim M., Casalino G. (2021). Drone deep reinforcement learning: A review. Electronics, 10(9), 999. doi: 10.3390/electronics10090999.

6. SHIN Sang-Yun; KANG Yong-Won; KIM Yong-Guk. Obstacle avoidance drone by deep reinforcement learning and its racing with human pilot. Applied sciences, 2019, 9.24: 5571, doi: 10.3390/app9245571.

7. Koning Tim. Low level quadcopter control using Reinforcement Learning: Developing a self-learning drone. (2020).

8. PASSALIS Nikolaos; TEFAS Anastasios. Continuous drone control using deep reinforcement learning for frontal view person shooting. Neural Computing and Applications, 2020, 32.9: 4227-4238. doi: 10.1007/s00521-019-04330-6.

9. Hu Y., Chen M., Saad W., Poor H. V., & Cui S. (2021). Distributed multi-agent meta learning for trajectory design in wireless drone networks. IEEE Journal on Selected Areas in Communications, 39(10), 3177-3192, doi: 10.1109/JSAC.2021.3088689.

10. Petrenko D. Selection of Deep Reinforcement Learning Using a Genetic Algorithm. *COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems,* 12 05 2022.