УДК 004.896 https://doi.org/10.26160/2541-8637-2023-11-7-13

МЕТОД ГЛУБОКОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ В ЗАДАЧАХ МАРШРУТИЗАЦИИ ДВИЖЕНИЯ МОБИЛЬНЫХ РОБОТОВ В СРЕДЕ С ПРЕПЯТСТВИЯМИ

Дудаков А.С.[1], Турсунов Т.Р.[1], Филимонов Н.Б.[1,2]
[1]Московский государственный университет им. М.В. Ломоносова;
[2]Институт проблем управления им. В.А. Трапезникова РАН, Москва

Ключевые слова: мобильный робот, задача маршрутизации движения, среда с препятствиями, методы машинного обучения, глубокое обучение с подкреплением.
Аннотация. Обсуждается задача маршрутизации движения мобильных роботов в среде с препятствиями с использованием метода глубокого обучения с подкреплением. Приведены основные положения и классификация метода глубокого обучения с подкреплением. Дан краткий обзор современного состояния задачи маршрутизации движения мобильных роботов в среде с препятствиями на основе метода глубокого обучения с подкреплением.

THE METHOD OF DEEP REINFORCEMENT LEARNING IN MOTION PLANNING PROBLEM OF MOBILE ROBOTS IN AN ENVIRONMENT WITH OBSTACLES

Dudakov A.S.[1], Tursunov T.R.[1], Filimonov N.B.[1,2]
[1]Lomonosov Moscow State University;
[2]Trapeznikov Institute of Control Problems of the RAS, Moscow

Keywords: mobile robot, motion planning problem, environment with obstacles, machine learning methods, deep reinforcement learning.
Abstract. The motion planning problem of mobile robots in an environment with obstacles using deep reinforcement learning is discussed. The main concepts and classification of deep reinforcement learning method are presented. A brief review of the current state of the motion planning problem of mobile robots in an environment with obstacles based on deep reinforcement learning is given.

Введение

Задача маршрутизации движения (МД) мобильных роботов (МР), включая беспилотные летательные аппараты (БПЛА, дрон, англ. UAV), является одной из ключевых в современной робототехнике. Данная задача заключается в планировании в рабочей зоне безопасной траектории движения МР из произвольно заданного начального состояния в заданное целевое конечное состояние с обходом встречных препятствий.

Для решения задачи МД МР весьма широкое распространение находят классические, традиционные методы на основе графов, дорожной карты, клеточной декомпозиции, диаграммы Вороного, потенциальных полей и др. При этом методы решения задач МД МР непрерывно совершенствуются и в последние годы все большую популярность приобретают эвристические и метаэвристические методы, вдохновленные природными явлениями и не имеющие строгого математического обоснования. Здесь следует выделить, например, методы имитации отжига, муравьиных колоний, «фейерверков», роя

частиц, реактивные методы, а также методы вычислительного интеллекта: нейросетевые методы, методы нечеткой логики, эволюционные методы, включая генетические алгоритмы и др. Настоящая работа посвящена краткому обзору методов решения задачи МД МР на основе машинного обучения и, в частности, глубокого обучения с подкреплением. Данные методы лишены недостатков, присущих классическим методам и обладают несомненными преимуществами, гарантируя высокую надежность и скорость поиска решения, а также способность автоматически обучаться с принятием решений.

Основные положения метода глубокого обучения с подкреплением

Машинное обучение (англ. Machine Learning, ML) – это раздел теории искусственного интеллекта, изучающий методы решения задач путем обучения на основе готовых решений множества сходных задач. Среди методов машинного обучения наибольшее распространение получили методы глубокого обучения и обучения с подкреплением.

Метод глубокого обучения (англ. Deep Learning, DL) – способ машинного обучения на основе нейронных сетей, использующий многослойную систему нелинейных фильтров для извлечения признаков с преобразованиями [1]. Здесь каждый последующий слой, получая данные предыдущего слоя, формирует иерархическую структуру с уровнями абстракции от низкого до высокого, в которой признаки каждого слоя могут обучаться с учителем или без учителя.

Метод обучения с подкреплением (англ. Reinforcement Learning, RL) – способ машинного обучения путем взаимодействия системы с внешней средой. Здесь система, принимающая решения, именуется агентом, цель которого – максимизировать некоторый критерий, именуемый «наградой», получаемой в процессе обучения в виде сигнала обратной связи от среды. Стратегия действия агента формируется автоматически на основе получаемых сведений о результатах действий.

Модель взаимодействия агента и среды может быть описана марковским процессом принятия решений (англ. Markov decision process, MDP) и представлена кортежем (S, A, R, P), где S – конечное множество состояний среды, A – конечное множество действий агента (алфавит доступных действий), R – значение приобретенного вознаграждения, P – вероятность перехода между состояниями.

Если агент в момент времени t совершает некоторое действие $a_t \in A$, то он переходит из состояния $s_t \in S$ в каждый временной шаг в состояние $s_{t+1} \in S$ с вероятностью P и получает некоторое вознаграждение $r_t \in R$ от среды. Цель обучения с подкреплением заключается в максимизации суммарного вознаграждения среды

$$R_t \to \max,\ R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

где $\gamma = \text{const}$ – коэффициент дисконтирования, ускоряющий процесс обучения.

Обозначим через π стратегию агента, обеспечивающую отображение пространства S в пространство A. Введем в рассмотрение две функции

полезности (ФП) $V_\pi(s)$ и $Q_\pi(s,a)$, оценивающие ожидаемые награды стратегии π и удовлетворяющие уравнениям Беллмана:

$$V_\pi(s) = E_\pi[R_t \mid s_t = s] = \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V_\pi(s')],$$

$$Q_\pi(s,a) = E_\pi[R_t \mid s_t = s, a_t = a] = \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma \sum_{a'} Q_\pi(s',a')],$$

где $P^a_{ss'} = P(s_{t+1} = s' \mid s_t = s, a_t = a)$, $R^a_{ss'} = E[r_{t+1} \mid s_t = s, s_{t+1} = s', a_t = a]$.

Обучение с подкреплением реализуется в совместном поиске на каждом шаге текущей оценки ФП и текущей стратегии действий. Для оптимизации агентом своей стратегии π были разработаны следующие два метода обучения: метод Монте-Карло и метод темпорально-разностного обучения (англ. Temporal-Difference learning, TD). В 1989 г. Воткинс (C.J.C.H. Watkins) предложил алгоритм *Q*-обучения [2], объединяющий модель MDP с методом TD-обучения.

Интеграция технологии глубоких нейронных сетей с методом RL привела к созданию *метода глубокого обучения с подкреплением* (англ. Deep Reinforcement Learning, DRL). Современные методы DRL разделяются на value-based методы (VB-методы) и policy-based методы (PB-методы). VB-методы DRL вычисляют стратегию π агента путем итеративного обновления ФП. При этом искомая стратегия агента определяется оптимально достижимым значением ФП. PB-методы DRL путем аппроксимации стратегии π формируют сеть стратегий и выбирают действия агента, которые оптимизируют методом градиента параметры стратегии с максимизацией вознаграждения.

В задачах МД МР с использованием метода DRL робот рассматривается как интеллектуальный подвижный агент, имеющий на борту: сенсорную систему датчиков внутренней и внешней информации (камеры, лидары); интеллектуальную систему навигации и управления движением; двигательную систему и коммуникационную систему для информационного взаимодействия с другими роботами-агентами.

Особенностью системы навигации МР, реализующей метод DRL в задаче МД, является наличие трех ключевых элементов: пространство состояний *S*, пространство действий *A* и функция вознаграждения *R*. Параметры пространства состояний включают начальную и целевую точки МР и препятствия. Состояние препятствий характеризуется их положением, размером и скоростью в случае динамических препятствий. При этом различают действия в виде дискретного перемещения (движение вперед, назад, поворот налево, направо и т.д.) и действия в виде непрерывного изменения скорости МР (линейной и угловой) и двигателей его привода (левого и правого).

Для повышения эффективности обучения в большинстве исследований используются следующие методы формирования вознаграждения:

– положительное вознаграждение за достижение цели и движение к ней;

– отрицательное вознаграждение, во-первых, за столкновение с препятствием или слишком близкое приближение к нему, во-вторых, за каждый временной шаг для побуждения робота двигаться быстрее на пути к цели.

Маршрутизация МР VB-методами DRL

VB-метод DRL оптимизирует стратегию π путем максимизации ФП $Q(s,a)$, значение которой для каждого состояния МР может быть получено на основе алгоритма *Q*-обучения. Это один из самых эффективных алгоритмов обучения с подкреплением, который использует жадную стратегию для выбора действий агента. Значение ФП $Q(s,a)$ на шаге t обновляется согласно следующей рекуррентной формуле:

$$Q^*(s_t,a_t) = Q(s_t,a_t) + \alpha(r_t + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a)),$$

где Q^* – оптимальное значение ФП, r_t – текущее вознаграждение, α – обучающий фактор.

Алгоритм планирования движения МР на основе *Q*-обучения использует в качестве входных данных информацию с лидара или камеры глубины, на основе чего формирует дискретное значение ФП $Q(s,a)$, которое хранится и итеративно обновляется. В работе [3] предложено новое пространство состояний, позволяющее уменьшить объем информации о значениях ФП и повысить скорость решения задачи МД. В работе [4] построено отображение между начальными и конечными значениями ФП. Следует отметить, что предварительные знания об окружающей среде интегрируются в систему для ускорения обучения и повышения эффективности стратегий.

Для планирования движения МР на основе визуального восприятия, *V*. Mnih и др. в работе [5] впервые объединили сверточную нейронную сеть с алгоритмом *Q*-обучения с подкреплением и предложили модель Deep *Q*-Network (DQN), которая использует изображение или видеоданные в качестве входной информации. Далее строится сеть значений, вычисляется и запоминается ФП. Здесь число нейронов в выходном слое сети определяет количество действий, совершаемых агентом (движение вперед, назад, поворот налево, направо и т.д.). В работе [6] предложена модификация метода DQN – метод DDQN (Double Deep *Q*-Network), использующий структуру двойной сети: текущая *Q*-сеть оптимизирует действие агента, а целевая *Q*-сеть оценивает данное действие.

В работе [7] для обучения движения МР внутри помещения со статическими препятствиями, карта глубины принята в качестве входной информации Q-сети, а действия и скорость робота – выходной. В работе [8] для решения задачи МД БПЛА реализован метод DQN с использованием камеры глубины в качестве элемента сенсорной системы. Следует отметить работу [9], в которой МР обучен автоматическому навигационному поиску в неизвестной среде. Здесь навигационная система МР использует глубокую нейронную сеть для обучения стратегиям исследования среды на основе локальных карт. В работе [10] метод DQN использован для маршрутизации полета квадрокоптера на основе данных динамики дрона и модели масштабируемой сенсорной системы.

Маршрутизация МР PB-методами DRL

Весьма серьезным недостатком VB-методов является их неприменимость для непрерывных задач МД. Данный недостаток удается обойти PB-методами DRL с использованием градиентной стратегии π [11]. Основными алгоритмами PB-методов DRL являются алгоритмы: DDPG, TRPO, PPO и A3C. Все эти

алгоритмы основаны на концепции “Актор-Критик” (англ. Actor-Critic, АК), согласно которой Актор отвечает за генерацию действий и взаимодействие с окружающей средой, а Критик использует алгоритм DQN для вычисления ФП на каждом шаге. По данным значениям ФП осуществляется оценка работы Актора и коррекция его действий на следующем шаге. В рамках концепции АК используется следующие аппроксимации ФП и стратегии:

$$\hat{v}(s,\omega) \approx v_\pi(s); \hat{q}(s,a,\omega) \approx q_\pi(s,a),$$

$$\pi_\theta(s,a) = P(a \mid s,\theta) \approx \pi(a \mid s),$$

а также следующее параметрическое обновление стратегии:

$$\theta = \theta + \alpha \log \pi_\theta(s_t, a_t) v_t,$$

где v_t – оптимальное значение состояния, вычисленное Критиком.

Актор использует v_t для обновления параметра стратегии θ и выбора действий, которые Критик использует для обновления параметра ω Q-сети. В результате Критик помогает Актору вычислять значение состояния v_t.

В работе [12] на основе концепции АК с использованием глубоких нейронных сетей предложен алгоритм Depth Deterministic Policy Gradient (DDPG). Здесь детерминированная стратегия μ определяет выбор действий агента $a_t = \mu(s_t \mid \theta^\mu)$, θ^μ – параметры сети стратегии. Целевая функция $J(\theta^\mu)$ вычисляется как математическое ожидание суммарной награды:

$$J(\theta^\mu) = E_{\theta^\mu}[r_1 + \gamma r_2 + \gamma^2 r_3 + ...].$$

В результате, целевое назначение алгоритма DDPG – максимизация целевой функции $J(\theta^\mu)$ и минимизация ФП Q.

В работе [13] для решения задачи МД МР использован алгоритм DDPG управления посадкой БПЛА на мобильную платформу, а в работе [14] представлен обучаемый планировщик траекторий, использующий модификацию алгоритма DDPG – асинхронный DDPG. В работе [15] алгоритм DDPG использован для задачи МД квадрокоптера путем связи управляющих команд напрямую с состоянием системы.

Для выбора подходящего шага в градиентной стратегии π с ее монотонным возрастанием в работе [16] предложен алгоритм Trust Region Policy Optimization (TRPO), в котором новая стратегия представляется суммой старой стратегии и остаточного члена. В работе [17] алгоритм TRPO использован для решения так называемой задачи социальной навигации МР, а в работе [18] – для управления квадрокоптером в высокоточной среде моделирования.

Для упрощения алгоритма TRPO в работе [19] предложена его модификация – алгоритм Proximal Policy Optimization (PPO). Используя преимущество монотонного возрастания стратегии PPO, в работе [20] предложен алгоритм appoNav (асинхронный PPO) для решения задачи визуальной навигации. Для задачи МД по лабиринту в помещении в работе [21] предложен алгоритм PPO, основанный на методе DDQN, который значительно сокращает время обучения.

Следует отметить еще один алгоритм, основанный на концепции АК – алгоритм Asynchronous Advantage Actor-Critic (A3C), предложенный V. Mnih и

др. [22]. Данный алгоритм создает несколько параллельных сред, для исследования которых используется группа Акторов. Алгоритм A3C показал высокую эффективность в таких задачах, как непрерывное управление роботизированной “рукой” [23], маршрутизация перемещения в лабиринте [24], навигация и управление четырехколесным роботом в 3D-среде [25]. В работе [26] предложен метод MK-A3C, основанный на асинхронных преимуществах памяти, для решения непрерывной задачи МД МР в сложной динамической среде. Здесь нейронная сеть используется для повышения способности робота как к временному обучению, так и к запоминанию. Благодаря оценке модели среды удается при встрече с препятствиями избежать проблемы ловушек.

Заключение

В работе представлен краткий обзор современного состояния задачи МД МР в среде с препятствиями на основе метода глубокого обучения с подкреплением. Результаты обзора можно систематизировать в виде таблицы, содержащей основные особенности следующих методов глубокого обучения с подкреплением: DQN, DDPG, TRPO, PPO, A3C.

Метод	Особенности
DQN	Простота реализации и эффективность. Неприменимость в непрерывных задачах МД МР.
DDPG	Применимость в непрерывных задачах МД МР.
TRPO	Вычислительная сложность и наличие накапливаемых ошибок. Проблема выбора оптимального шага в градиентной стратегии π.
PPO	Упрощенная версия реализации алгоритма TRPO.
A3C	Высокая эффективность благодаря распараллеливанию обучения нескольких агентов.

Список литературы

1. Deng L.,Yu D. Deep Learning: Methods and Applications // Foundations and Trends in Signal Processing. 2014, vol. 7, no. 3-4, pp. 1-199.
2. Watkins C.J.C.H. Learning from delayed rewards: PhD dissertation. – Cambridge, England: University of Cambridge, 1989. – 241 p.
3. Jaradat Kareem M.A., Al-Rousan M., Quadan L. Reinforcement based mobile robot navigation in dynamic environment // Robot. Comput.-Integr. Manuf. 2011, vol. 27, no. 1, pp. 135-149.
4. Song Y., Li Y.-B., Li C.-H., Zhang G.-F. An efficient initialization approach of Q-learning for mobile robots // Int. J. Control, Autom. Syst. 2012, vol. 10, no. 1, pp. 166-172.
5. Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., Graves A., Riedmiller M., Fidjeland A.K., Ostrovski G., Petersen S. Human-level control through deep reinforcement learning // Nature. 2015, vol. 518, no. 7540, pp. 529-533.
6. Van Hasselt H., Guez A., Silver D. Deep reinforcement learning with double Q-learning // Proc. 30th AAAI Conf. Artificial Intelligence, Phoenix, AZ, USA. 2016, pp. 2094-2100.
7. Tai L., Paolo G., Liu M. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation // Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS). Vancouver, Canada: IEEE. 2017, pp. 31-36.
8. Wang D., Li W., Liu X., Li N., Zhang C. UAV environmental perception and autonomous obstacle avoidance: A deep learning and depth camera combined solution // Comput. Electron. Agricult. 2020, vol. 175, no. 105523, pp. 1-31.

9. Li H., Zhang Q., Zhao D. Deep reinforcement learning-based automatic exploration for navigation in unknown environment // IEEE Trans. Neural Netw. Learn. Syst. 2020, vol. 31, no. 6, pp. 2064-2076.
10. Kang K., Belkhale S., Kahn G., Abbeel P., Levine S. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight // Proc. Int. Conf. Robot. Autom. (ICRA). Montreal, QC, Canada: IEEE. 2019, pp. 6008-6014.
11. Xu J., Wei Z., Xia L., Lan Y., Yin D., Cheng X., Wen J.-R. Reinforcement learning to rank with pairwise policy gradient // Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. Virtual, China: Association Computing Machinery. 2020, pp. 509-518.
12. Lillicrap T.P., Hunt J.J., Pritzel A., Heess N., Erez T., Tassa Y., Silver D., Wierstra D. Continuous control with deep reinforcement learning // Comput. Sci. 2015, vol. 8, no. 6, pp. 1-14.
13. Rodriguez-Ramos A., Sampedro C., Bavle H. A deep reinforcement learning strategy for UAV autonomous landing on a moving platform // J. Intell. Robot. Syst. 2019, vol. 93, no. 1-2, pp. 351-366.
14. Tai L., Liu M. Towards cognitive exploration through deep reinforcement learning for mobile robots // arXiv:1610.01733. 2016, pp. 1-8.
15. Wang Y., Sun J., He H., Sun C. Deterministic policy gradient with integral compensator for robust quadrotor control // IEEE Trans. Syst., Man, Cybern. Syst. 2020, vol. 50, no. 10, pp. 3713-3725.
16. Schulman J., Levine S., Abbeel P., Jordan M., Moritz P. Trust region policy optimization // Proc. Int. Conf. Mach. Learn. Lile, France: International Machine Learning Society (IMLS). 2015, pp. 1889-1897.
17. Li M., Jiang R., Ge S.S., Lee T.H. Role playing learning for socially concomitant mobile robot navigation // CAAI Trans. Intell. Technol. 2018, vol. 3, no. 1, pp. 49-58.
18. Koch W., Mancuso R., West R., Bestavros A. Reinforcement learning for UAV attitude control // ACM Trans. Cyber-Phys. Syst. 2019, vol. 3, no. 2, pp. 1-21.
19. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithm // Mach. Learn. 2017, pp. 1-12.
20. Zeng F.Y., Wang C. Visual navigation with asynchronous proximal policy optimization in artificial agents // Journal of Robotics. 2020, vol. 2020, pp. 1-7.
21. Marchesini E., Farinelli A. Discrete deep reinforcement learning for mapless navigation // Proc. IEEE Int. Conf. Robot. Autom. (ICRA). Paris, France: IEEE. 2020, pp. 10688-10694.
22. Mnih V., Badia A.P., Mirza M., Graves A. Asynchronous methods for deep reinforcement learning // Proc. 33rd Int. Conf. Mach. Learn. 2016, vol. 48, pp. 1928-1937.
23. Yang S., Wang Q. Robotic Arm Motion Planning with Autonomous Obstacle Avoidance Based on Deep Reinforcement Learning // 41st Chinese Control Conference (CCC), Hefei, China. 2022, pp. 3692-3697.
24. Mirowski P., Pascanu R., Viola F., Soyer H., Ballard A.J., Banino A., Denil M., Goroshin R., Sifre L., Kavukcuoglu K., et al. Learning to navigate in complex environments // arXiv:1611.03673. 2017, pp. 1-16.
25. Zhang K., Niroui F., Ficocelli M., Nejat G. Robot navigation of environments with unknown rough terrain using deep reinforcement learning // Proc. IEEE Int. Symp. Saf., Secur., Rescue Robot. (SSRR). Philadelphia, PA, USA: IEEE. 2018, pp. 1-7.
26. Zeng J., Ju R., L. Qin L., Hu Y., Yin Q., Hu C. Navigation in unknown dynamic environments based on deep reinforcement learning // Sensors. 2019, vol. 19, no. 18, pp. 1-18.

Сведения об авторах:
Дудаков Александр Сергеевич – студент;
Турсунов Тимур Рустамович – студент;
Филимонов Николай Борисович – д.т.н., профессор, главный научный сотрудник, заместитель заведующего кафедрой.