- формирований в условиях работы в зоне радиоактивного загрязнения, оптимизации комплекса защитных мероприятий);
- визуальное определение полей загрязнения (расчётных и экспериментальных) на трёхмерных моделях элементов промышленной и городской застройки;
- в динамике представление данных по переносу загрязняющих примесей в приземном слое воздуха, а также по полям загрязнения;
- представление информации, характеризующей дозы по маршрутам полей радиоактивного загрязнения и т. п.;
- подготовка данных для оптимизации работы средств робототехники;
- создание имитационных программ для отработки плана действий аварийных служб и подразделений, участвующих в ликвидации последствий чрезвычайных ситуаций [1].

В целом, имитационное 3D моделирование дает большие преимущества: сбор и хранение информации о зданиях, сооружениях различного назначения, состояния путей эвакуации. Позволяет быстро оценить ситуацию на 3D модели, обеспечить: ликвидацию аварии, минимизировать последствия, быструю и слаженную работу аварийных служб и подразделений.

Для работы с имитационной системой «НЕОЛАНТ» необходима подготовка специалистов. Система формирует у обучаемых комплексное, системное, структурное понимание динамических процессов, происходящих в зоне чрезвычайной ситуации, алгоритм действий для ликвидации аварии и минимизации ее последствий. Кроме того, актуальность имитационной системы заключается в том, что подобные, чрезвычайные ситуации невозможно произвести из-за их опасности, значительных материальных затрат на натурное моделирование.

Освоение данного программного продукта поможет обучаемым приобрести необходимые профессиональные компетенции.

К недостаткам можно отнести необходимость создания достоверной информационной модели, что требует привлечения усилий разнопрофильных специалистов [2].

1. http://www.neolant.ru/projects/news_detail.php?ID=204 (дата обращения 06.07.2020)
2. Донец С.А., Санникова С.М. Применение 3D визуализации для проведения пожарно-тактических расчетов с учетом современных информационных технологий // Пожарная безопасность: проблемы и перспективы. - 2017. Т. 1. № 1 (8). С. 58-60.

Доржиев А.С.

Обучение с подкреплением для беспилотных летательных аппаратов

*Финансовый университет при Правительстве Российской Федерации
(Россия, Москва)*

doi:10.18411/sciencepublic-12-07-2020-02

idsp: sciencepublic-12-07-2020-02

*Научный руководитель
Никитин П.В.*

Беспилотные летательные аппараты (БПЛА), такие как квадрокоптеры, являются активно развивающейся сферой авиастроения. Простота обслуживания и быстроедействие по сравнению с пилотируемой и наземной техникой, маневренность, небольшие габариты, способность нести в полете полезные грузы, а также возможность установки камер для съемки позволили решать широкий спектр задач, начиная от

строго военных задач патрулирования и мониторинга территорий до оперативных мобильных дефибрилляторов для оказания медицинской помощи [1].

Современный квадрокоптер – это беспилотный летательный аппарат (БПЛА) с четырьмя моторами, одна пара которых вращается по часовой стрелке, вторая – против [2]. В связи с этим пропадает необходимость в хвостовом винте, который имеется у вертолета. Управление движением осуществляется путём изменения скорости вращения винтов.

Технологии искусственного интеллекта, машинное обучение, большие данные также имеют богатую сферу применения: а именно: скоринговые, прогнозные и рекомендательные системы, распознавание объектов, поиск аномалий и т.д. [3, 4]. В ряде задач для создания подобных программных продуктов используются нейронные сети. Они позволяют реализовывать более сложные модели, которые, например, сами добывают и анализируют данные, а затем, обучаясь на них, корректируют свои дальнейшие действия. Такие модели называются самообучающимися и реализуются методами обучения с подкреплением.

Обучение с подкреплением (ОсП или RL – Reinforcement Learning) – это раздел машинного обучения, при котором агент взаимодействует с окружающей средой, тем самым получает некоторый опыт, на основе которого корректирует свои дальнейшие действия. На рис. 1 приведены схема и основные термины ОсП:

Рис. 1. Схема итерации обучения с подкреплением

Состояние – это фиксированное или динамическое положение агента. Множество различных действий агента называется «пространством действий» или действием. Среда – это место, где агент существует и с которым взаимодействует. Награда представляет собой функцию, которая служит для анализа действий агента, настройки алгоритма обучения и его оптимизации. Она зависит от текущего состояния мира и выбранного действия, а также стратегии, которая является совокупностью правил, по которым агент выбирает следующее действие.

ОсП может применяться в ряде задач коптеров, таких как: полеты по траектории, преследование цели, дроны-регулирующие и т.д.

В качестве примера рассмотрим применение ОсП для того, чтобы научить коптер взлетать. Введем агента, который управляет дроном, устанавливая случайным образом обороты в секунду на каждом из его четырех моторов. Задача взлета сводится к перемещению из начальной точки (0, 0, 0) в целевую точку (0, 0, 20) в координатах (x, y, z). Агент награждается в случае, если коптер взлетел и не разбился. Штрафуется в случае крушения при попытке взлета.

Для решения данной задачи используется алгоритм Deep Deterministic Policy Gradient (DDPG), который применяется для случая, когда управляющие команды представляют собой значения из непрерывных пространств [5]. В нашей задаче возможных значений оборотов двигателя, которые выбирает агент, теоретически бесконечно много, поэтому нужно, чтобы алгоритм выдавал конкретные значения, которые нам необходимо применить в текущей ситуации.

На входе алгоритм принимает текущее состояние агента S , которое обладает свойством марковости, то есть включает в себя всю информацию, необходимую для предсказания следующих состояний системы. На основе состояния алгоритм выдает информацию о действии A , которое необходимо сделать, чтобы обеспечить оптимальное поведение, которое ведет к максимизации награды. Награда R — это вещественное число, которое используется для обучения алгоритма. Награда выдается в момент, когда агент, находясь в состоянии S , совершает действие A и переходит в состояние S' . Для обучения эти алгоритмы используют информацию о том, что они делали в прошлом и какой при этом был результат. Эти элементы опыта состоят из 4х компонентов (S, A, R, S') [6]. Такие четверки мы складываем в массив опыта и далее используем для обучения.

Обучение реализуется с помощью двух нейросетей: $\mu(S)$ — актер (действует), который возвращает нам вектор A , компоненты которого являются значениями соответствующих управляющих сигналов. Вторая нейросеть $Q(S, A)$ — критик. Она возвращает оценку q действия A в состоянии S . Критик используется для обучения актора и обучается с помощью уравнения Беллмана.

Q-функция актора:

$$Q^{\mu}(s_t, a_t) = E_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma Q^{\mu}(s_{t+1}, \mu(s_{t+1}))]$$

L-функция:

$$L(\theta^Q) = E_{s_t \sim \rho^{\beta}, a_t \sim \beta, r_t \sim E}[(Q(s_t, a_t | \theta^Q) - y_t)^2]$$

Уравнение Беллмана:

$$y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1}) | \theta^Q)$$

Обучение критика происходит с помощью градиента от оценки критика $\nabla Q(S, \mu(S))$ по весам актора:

$$\begin{aligned} \nabla_{\theta^{\mu}} J &\approx E_{s_t \sim \rho^{\beta}} \left[\nabla_{\theta^{\mu}} Q(s, a | \theta^Q) \Big|_{s=s_t, a=\mu(s_t | \theta^{\mu})} \right] \\ &= E_{s_t \sim \rho^{\beta}} \left[\nabla_a Q(s, a | \theta^Q) \Big|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^{\mu}} \mu(s | \theta^{\mu}) \Big|_{s=s_t} \right] \end{aligned}$$

Таким образом, алгоритм состоит из 5 шагов:

- 1) Выбираем партию из N элементов опыта (S, A, R, S')
- 2) Рассчитываем функции y и L
- 3) Обучаем критика с помощью вычисленной L
- 4) Берем градиент по весам актора $\nabla Q(S, \mu(S))$ и с помощью него корректируем веса нейронной сети актора
- 5) Повторяем шаги 1-4 до желаемого результата

Результат обучения и симуляции в Jupyter Notebook представлены на рис.2, где по вертикальной оси – награда, по горизонтальной оси – эпизоды обучения.

Рис. 2. График награды

По графику награды можно заметить, что примерно после 40 попытки агент понимает, как нужно задействовать моторы коптера, чтобы взлететь, о чем свидетельствует уровень награды, который в последующие эпизоды колеблется около максимального значения.

Таким образом, в данной статье разобрано одно из применений RL для задач коптеров. Подобных задач, стоящих на стыке техники, транспорта и ОсП, много: беспилотные автомобили, шагающие роботы, интеллектуальные манипуляторы и т.д. [7]. Поэтому, дальнейшие исследования в рамках этих направлений могут привести к созданию умных машин, способных обучаться, ставить и решать задачи самостоятельно. Это позволит автоматизировать промышленность, городской транспорт, а также расширит технические возможности для космических исследований.

1. Хамзатов М. Особенности развития беспилотной авиации в современных условиях [Электронный ресурс]. – режим доступа: URL: <http://uav.ru/articles/hmm.pdf> (дата обращения: 07.07.2020)
2. Квадрокоптер для любительской съемки [Электронный ресурс]. – режим доступа: URL: <http://infocopter.ru/kvadrokoptery-dlya-lyubitelskoj-semki-obzor-modelej/> (дата обращения: 10.06.2020)
3. Сюрмаков Р. Применение машинного обучения и Data Science в промышленности [Электронный ресурс]. – режим доступа: URL: <https://vc.ru/ml/79368-primenenie-mashinnogo-obucheniya-i-data-science-v-promyshlennosti> (дата обращения: 07.07.2020)
4. Никитин П.В., Баркова А.С. Математические методы для определения оценки резерва в страховании. Часть 1. Метод цепной лестницы // Вестник евразийской науки. 2019. Т. 11. № 6. С. 43.
5. Курама В., Алла С. Обучение с подкреплением на языке Python [Электронный ресурс]. – режим доступа: URL: <https://habr.com/ru/company/piter/blog/434738/> (дата обращения: 26.06.2020)
6. Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra. Continuous control with deep reinforcement learning. URL: <https://arxiv.org/abs/1509.02971> (дата обращения 25.06.2020)
7. Обучение с подкреплением / Р.С. Саттон, Э.Г. Барто; пер. с англ. – М.: БИНОМ. Лаборатория знаний, 2011. – 399 с.

Моисеев И.Ю.

**Тестирование метода совместного использования параметров в мультиагентном
глубоком обучении с подкреплением**

*ФГБОУ ВО «Финансовый университет при Правительстве Российской
Федерации»
(Россия, Москва)*

doi:10.18411/sciencepublic-12-07-2020-03

idsp: sciencepublic-12-07-2020-03

*Научный руководитель
Никитин П.В.*

Введение

Мультиагентное обучение с подкреплением является активно развивающейся ветвью машинного обучения, и имеет множество прикладных областей, от имитации рынка до системы управления дроном. [1-2]

Большинство моделей обучают агентов по независимым политикам. В результате каждый агент может действовать независимо от других агентов и центрального контроллера. Обучение подобных агентов сталкивается с проблемами из-за нестационарности среды.

Нестационарная среда – среда, которая изменяется в процессе обучения. Нестационарность сильно усложняет обучение, ведь знания агента о среде со временем устаревают.