**RL в ИБ**

Лекция Семёна Козлова из курса DLcourse <https://www.youtube.com/watch?v=_x0ASf9jV9U>

Оригинальная статья Andrej Karpathy http://karpathy.github.io/2016/05/31/rl/

Перевод статьи Andrej Karpathy <https://habr.com/ru/post/439674/>

1. **Титульный слайд (название модуля, автор)**

Приветствие. Занятие «Обучение с подкреплением для информационной безопасности» в модуле «Искусственный интеллект в информационной безопасности».

1. **Программа модуля**

Кратко – про что будет доклад (две части – кратко про основы RL и про его применение в ИБ). На практическом занятии: анализ спама (классификация)

1. **Обучение с подкреплением**

Что такое обучение с подкреплением в целом. Пояснить, кто такой «агент» и какие у него есть возможности.

<http://ai.lector.ru/?go=lection07>

1. **Среда**

Что такое марковский процесс и как среда видит агента (с выбором действий).

<https://spinningup.openai.com/en/latest/spinningup/rl_intro.html>

1. **Состояние**

Состояние с точки зрения среды и с точки зрения агента. Рисунок.

1. **Действие**

Определение действия в RL, классификация пространств действий. Действие с точки зрения агента и среды. Примеры действий в понятных играх (задачах). Вероятность перехода в состояние после действия.

1. **Награда**

Определение награды в RL. Методы оценки действий – немедленный и отсроченный. Показать на рисунке.

1. **Траектории (эпизоды)**

Описание траектории в RL. Объяснение способа их формирования. Табличное представление траекторий и шагов в них.

1. **Политики**

Описание политики в RL. Задача политики. Формулы. Коэффициент дисконтирования (жадные и e-жадные политики). Будущие награды.

1. **Функции значения**

Определение. Пояснить формулы для политики в целом и для оптимальной политики. Объяснить, как выбирается оптимальное действие.

1. **Policy Gradients: общая идея**

Объяснение формулы функции политики. Объяснение формулы ожидаемой награды. Демонстрация изменения политика при обучении. Демонстрация анимации с пояснением того, что такое градиентный спуск. Объяснение формирования вероятности траектории.

1. **Policy Gradients: немного математики**

Напоминание вероятность прохождения стратегии. Показать трюк с градиентом и логарифмом. Показать, как перейти к логарифму вероятности прохождения траектории.

<https://spinningup.openai.com/en/latest/spinningup/rl_intro3.html>

1. **Policy Gradients: в поисках градиента**

Напоминание про то, что ищем (градиент политики) и про то, что доказано (логарифму вероятности прохождения траектории). Подстановка второго в первое, оценка по выборочному среднему. Возможность оценить по сыгранным траекториям. Демонстрация того, как считается изменение параметров политики.

1. **Policy Gradients: Reward-to-Go**

Рассказ о том, как улучшить базовую политику и работать с наградой за каждый шаг в траектории. Демонстрация на рисунке, к каким положительным изменениям это ведёт (уменьшение шума и пр.)

1. **Policy Gradients: baseline**

Лемма EGLP. Использование baseline для дополнительной мотивации нейронной сети (101>100, но 1>>0).

1. **Policy Gradients: хватит математики, как это работает?**

Описание алгоритма обучения модели RL с применением нейронной сети для оптимизации параметров.

1. **Policy Gradients: actor-critic**

Расширенный (адаптивный) вариант baseline. Дополнительный выходной нейрон.

1. **Выводы по Policy Gradients**

Описание сильных и слабых сторон, подведение промежуточного итога.

1. **Q-learning**

Решение проблемы неообходимости переигрывания эпизодов, характерной для классического Policy Gradients. Q-функция – принципиально важное нововведение: выбирается не оптимальное действие, а награда, которую можно получить за каждое действие.

1. **Уравнение Беллмана**

Демонстрация и объяснение формулы ожидаемой награды, расписанной на «следующий и остальные» шаги. Демонстрация неизменности при оптимальной политике.

1. **Q-learning: идея**

Описание возможности «заглядывать в будущее» с помощью «старой версии себя». Уточнённая награда. Демонстрация на схеме.

1. **Q-learning: алгоритм**

Описание алгоритма обучения Q-learning алгоритма – схема и таблица на слайде.

1. **Q-learning vs Policy Gradients**

Сравнение двух алгоритмов с описанием достоинств и недостатков каждого.

1. **Классификация алгоритмов RL**

Таксономия алгортмов RL, разработанная сообществом OpenAI. Верхнеуровнево описать и

<https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html>

1. **Примеры применения RL в ИБ**

Описать основные примеры применения RL в ИБ.

Li, C., & Qiu, M. (2019). Reinforcement Learning for Cyber-Physical Systems: with Cybersecurity Case Studies (1st ed.). Chapman and Hall/CRC, главы 7-9

1. **Противодействие радиопомехам при передаче информации: существующие методы**

Принципиальная задача. Существующие решения – отстройка по частоте.

1. **Противодействие радиопомехам при передаче информации: применение RL**

Описание метода, представленного в статье ниже. Смысл игры – угадать по расположению занятых каналов на текущем шаге, какие каналы будут заняты на следующем шаге.

B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy. An anti-jamming stochastic game for cognitive radio networks. IEEE Journal on Selected Areas in Communications, 29(4):877–889, 2011.

1. **Мобильные периферийные вычисления: существующие методы ИБ**

Точки доступа, в связи с внедрением 5G, планируют хранить информацию, чтобы быстрее обслуживать пользователей, и экономить передаваемый трафик.

Потенциальные риски – потратить ресурсы на хранение невостребованных данных и дать доступ к конфиденциальной информации тем людям, у которых его не должно быть.

1. **Мобильные периферийные вычисления: применение RL для защиты**

Описание среды как множества состояний и состояний как сэмплов информации. Всё в соответствии со статьёй:

L. Xiao, X. Wan, C. Dai, X. Du, X. Chen, and M. Guizani. Security in mobile edge caching with reinforcement learning. arXiv preprint arXiv:1801.05915, 2018.

1. **Распределение задач между работниками: существующие методы**

Задача массового обслуживания. Распределение аналитиков по событиям (или событий по аналитикам – в зависимости от масштаба).

1. **Распределение задач между работниками: применение RL для защиты**

Идея – определить модель игры и возможности каждого аналитика, и обучить модель играть так, чтобы получать выигрыш в достаточно долгосрочной перспективе.

R. Ganesan, S. Jajodia, A. Shah, and H. Cam. Dynamic scheduling of cybersecurity analysts for minimizing risk using reinforce- ment learning. ACM Transactions on Intelligent Systems and Technology (TIST), 8(1):4, 2016

1. **Поиск безопасного пути передачи информации: задача**

Задача – найти безопасный путь соединения через доверенные узлы. Для простых случаев есть готовые решения, но для сложных сетей они вычислительно сложны. Пример алгоритма Дейкстры – в анимации на слайде.

1. **Поиск безопасного пути передачи информации: применение RL**

Схожесть с задачей объезда пробок. Задача в терминологии RL. Потребность в адаптивной модели и возможности RL в этой области.

1. R. Nannapaneni «Optimal path routing using reinforcement learning», Dell EMC

2. Li, C., & Qiu, M. (2019). Reinforcement Learning for Cyber-Physical Systems: with Cybersecurity Case Studies (1st ed.). Chapman and Hall/CRC

1. **Автоматизированное тестирование на проникновение: существующие методы**

Экспертов пока заменить не получается, но есть много ПО для облегчения им жизни. Но ПО работает по детерминированным алгоритмам, что не даёт возможности обнаруживать новые уязвимости.

1. **Автоматизированное тестирование на проникновение: применение RL**

RL в этой области может помочь обнаруживать новые уязвимости с помощью комбинирования известных подходов и способов. Существующий инструмент – DeepExplot, основанный на Metasploit.

https://www.vulnhub.com/

https://github.com/rapid7/metasploitable3

Мясников А.В. «Применение машинного обучения с подкреплением в задаче тестирования на проникновение» (2020).

УК РФ ст. 272. Неправомерный доступ к компьютерной информации, УК РФ ст. 273. Создание, <>.. программ <..> предназначенных для <..> нейтрализации СЗИ <..> до 7 лет лишения свободы

Выводы по лекции.

Спасибо за внимание.