

ЛАУРА  
ГРЕССЕР

ВАН  
ЛУН КЕНГ

# Глубокое обучение с подкреплением



ТЕОРИЯ  
И ПРАКТИКА  
НА ЯЗЫКЕ PYTHON



ЛАУРА ГРЕССЕР, ВАН ЛУН КЕНГ

# Глубокое обучение с подкреплением

ТЕОРИЯ И ПРАКТИКА  
НА ЯЗЫКЕ PYTHON



Санкт-Петербург · Москва · Минск

2022

ББК 32.813 + 32.973.23-018

УДК 004.89

Г91

## Грессер Лаура, Кенг Ван Лун

Г91 Глубокое обучение с подкреплением: теория и практика на языке Python. — СПб.: Питер, 2022. — 416 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-4461-1699-7

Глубокое обучение с подкреплением (глубокое RL) сочетает в себе два подхода к машинному обучению. В ходе такого обучения виртуальные агенты учатся решать последовательные задачи о принятии решений. За последнее десятилетие было много неординарных достижений в этой области — от однопользовательских и многопользовательских игр, таких как го и видеоигры Atari и Dota 2, до робототехники.

Эта книга — введение в глубокое обучение с подкреплением, уникально комбинирующее теорию и практику. Авторы начинают повествование с базовых сведений, затем подробно объясняют теорию алгоритмов глубокого RL, демонстрируют их реализации на примере программной библиотеки SLM Lab и напоследок описывают практические аспекты использования глубокого RL.

Руководство идеально подойдет как для студентов, изучающих компьютерные науки, так и для разработчиков программного обеспечения, которые знакомы с основными принципами машинного обучения и знают Python.

**16+** (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.813 + 32.973.23-018

УДК 004.89

Права на издание получены по соглашению с Pearson Education Inc. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги. Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

ISBN 978-0135172384 англ.

ISBN 978-5-4461-1699-7

© 2020 Pearson Education, Inc.

© Перевод на русский язык ООО Издательство «Питер», 2022

© Издание на русском языке, оформление ООО Издательство «Питер», 2022

© Серия «Библиотека программиста», 2022

# Краткое содержание

Предисловие .....	16
Введение .....	18
Благодарности .....	22
Об авторах .....	23
От издательства .....	24
<b>Глава 1. Введение в обучение с подкреплением .....</b>	<b>25</b>

## **Часть I. Алгоритмы, основанные на стратегиях и полезностях**

<b>Глава 2. REINFORCE .....</b>	<b>52</b>
<b>Глава 3. SARSA .....</b>	<b>81</b>
<b>Глава 4. Глубокие Q-сети .....</b>	<b>112</b>
<b>Глава 5. Улучшение DQN .....</b>	<b>136</b>

## **Часть II. Комбинированные методы**

<b>Глава 6. Метод актора-критика с преимуществом (A2C) .....</b>	<b>168</b>
<b>Глава 7. Оптимизация ближайшей стратегии .....</b>	<b>198</b>
<b>Глава 8. Методы параллелизации .....</b>	<b>228</b>
<b>Глава 9. Сравнительный анализ алгоритмов .....</b>	<b>239</b>

### Часть III. Практика

Глава 10. Начало работы с глубоким RL.....	242
Глава 11. SLM Lab.....	274
Глава 12. Архитектура сетей.....	286
Глава 13. Аппаратное обеспечение.....	311

### Часть IV. Проектирование сред

Глава 14. Состояния .....	328
Глава 15. Действия.....	358
Глава 16. Вознаграждения.....	374
Глава 17. Функция переходов.....	383
Заключение .....	389

### Приложения

Приложение А. История глубокого обучения с подкреплением .....	394
Приложение Б. Примеры сред .....	397
Список используемых источников.....	405

# Оглавление

Предисловие .....	16
Введение.....	18
Благодарности.....	22
Об авторах .....	23
От издательства .....	24
О научном редакторе русскоязычного издания .....	24
<b>Глава 1. Введение в обучение с подкреплением .....</b>	<b>25</b>
1.1. Обучение с подкреплением .....	25
1.2. Обучение с подкреплением как МППР .....	31
1.3. Обучаемые функции в обучении с подкреплением.....	35
1.4. Алгоритмы глубокого обучения с подкреплением .....	37
1.4.1. Алгоритмы, основанные на стратегии .....	38
1.4.2. Алгоритмы, основанные на полезности .....	39
1.4.3. Алгоритмы, основанные на модели среды .....	40
1.4.4. Комбинированные методы .....	41
1.4.5. Алгоритмы, которые обсуждаются в этой книге.....	42
1.4.6. Алгоритмы по актуальному и отложенному опыту.....	43
1.4.7. Краткий обзор методов .....	44
1.5. Глубокое обучение для обучения с подкреплением .....	44
1.6. Обучение с подкреплением и обучение с учителем .....	47
1.6.1. Отсутствие оракула.....	47
1.6.2. Разреженность обратной связи .....	48
1.6.3. Генерация данных .....	49
1.7. Резюме .....	49

## Часть I. Алгоритмы, основанные на стратегиях и полезностях

<b>Глава 2. REINFORCE</b> .....	52
2.1. Стратегия .....	53
2.2. Целевая функция .....	53
2.3. Градиент стратегии .....	54
2.3.1. Вывод формулы для градиента по стратегиям .....	55
2.4. Выборка методом Монте-Карло .....	58
2.5. Алгоритм REINFORCE .....	59
2.5.1. Усовершенствование метода REINFORCE .....	60
2.6. Реализация REINFORCE .....	61
2.6.1. Минимальная реализация REINFORCE .....	61
2.6.2. Построение стратегий с помощью PyTorch .....	64
2.6.3. Выборка действий .....	67
2.6.4. Расчет потерь, обусловленных стратегией .....	67
2.6.5. Цикл обучения в REINFORCE .....	68
2.6.6. Класс Memoгу для хранения примеров при обучении по актуальному опыту .....	69
2.7. Обучение агента в REINFORCE .....	72
2.8. Результаты экспериментов .....	75
2.8.1. Эксперимент по оценке влияния коэффициента дисконтирования $\gamma$ .....	76
2.8.2. Эксперимент по оценке влияния базового значения .....	78
2.9. Резюме .....	79
2.10. Рекомендуемая литература .....	80
2.11. Историческая справка .....	80
<b>Глава 3. SARSA</b> .....	81
3.1. Q-функция и V-функция .....	82
3.2. Метод временных различий .....	85
3.2.1. Принцип метода временных различий .....	88
3.3. Выбор действий в SARSA .....	95
3.3.1. Исследование и использование .....	96
3.4. Алгоритм SARSA .....	97
3.4.1. Алгоритмы обучения по актуальному опыту .....	98
3.5. Реализация SARSA .....	99
3.5.1. $\epsilon$ -жадная функция выбора действий .....	99

3.5.2. Расчет Q-функции потерь.....	100
3.5.3. Цикл обучения в SARSA.....	102
3.5.4. Память для хранения пакетов прецедентов при обучении по актуальному опыту.....	103
3.6. Обучение агента SARSA.....	105
3.7. Результаты экспериментов.....	108
3.7.1. Эксперимент по определению влияния скорости обучения .....	108
3.8. Резюме .....	109
3.9. Рекомендуемая литература.....	110
3.10. Историческая справка .....	111
<b>Глава 4. Глубокие Q-сети.....</b>	<b>112</b>
4.1. Настройка Q-функции в DQN.....	113
4.2. Выбор действий в DQN .....	115
4.2.1. Стратегия Больцмана.....	118
4.3. Хранение прецедентов в памяти.....	121
4.4. Алгоритм DQN .....	122
4.5. Реализация DQN.....	124
4.5.1. Расчет Q-функции потерь.....	124
4.5.2. Цикл обучения DQN.....	125
4.5.3. Память прецедентов.....	126
4.6. Обучение агента DQN.....	129
4.7. Результаты экспериментов.....	132
4.7.1 Эксперимент по определению влияния архитектуры сети.....	132
4.8. Резюме .....	134
4.9. Рекомендуемая литература.....	134
4.10. Историческая справка .....	135
<b>Глава 5. Улучшение DQN.....</b>	<b>136</b>
5.1. Прогнозные сети .....	137
5.2. Двойная DQN.....	139
5.3. Приоритизированная память прецедентов .....	143
5.3.1. Выборка по значимости .....	145
5.4. Реализация улучшенной DQN.....	146
5.4.1. Инициализация сети.....	147
5.4.2. Расчет Q-функции потерь.....	147
5.4.3. Обновление прогнозной сети .....	148



5.4.4. DQN с прогнозными сетями .....	149
5.4.5. Двойная DQN.....	150
5.4.6. Приоритизированная память прецедентов.....	150
5.5. Обучение агента DQN играм Atari .....	156
5.6. Результаты экспериментов.....	161
5.6.1. Эксперимент по оценке применения двойной DQN с PER.....	162
5.7. Резюме .....	165
5.8. Рекомендуемая литература .....	165

## **Часть II. Комбинированные методы**

<b>Глава 6. Метод актора-критика с преимуществом (A2C) .....</b>	<b>168</b>
6.1. Актор .....	169
6.2. Критик .....	169
6.2.1. Функция преимущества .....	169
6.2.2. Настройка функции преимущества .....	174
6.3. Алгоритм A2C .....	175
6.4. Реализация A2C .....	178
6.4.1. Оценка преимущества .....	178
6.4.2. Расчет функции потерь для полезности и стратегии .....	181
6.4.3. Цикл обучения актора-критика .....	181
6.5. Архитектура сети .....	182
6.6. Обучение агента A2C.....	184
6.6.1. A2C с оценкой преимущества по отдаче за $n$ шагов в Pong .....	184
6.6.2. A2C с GAE в Pong.....	188
6.6.3. A2C по отдаче за $n$ шагов в BipedalWalker.....	188
6.7. Результаты экспериментов.....	191
6.7.1. Эксперимент по определению влияния отдачи за $n$ шагов .....	191
6.7.2. Эксперимент по выявлению влияния $\lambda$ в GAE .....	193
6.8. Резюме .....	195
6.9. Рекомендуемая литература .....	196
6.10. Историческая справка .....	196
<b>Глава 7. Оптимизация ближайшей стратегии .....</b>	<b>198</b>
7.1. Суррогатная целевая функция.....	199
7.1.1. Падение производительности.....	199
7.1.2. Преобразование целевой функции.....	201

7.2. Оптимизация ближайшей стратегии .....	208
7.3. Алгоритм PPO .....	212
7.4. Реализация PPO .....	214
7.4.1. Расчет функции потерь для стратегии в PPO .....	214
7.4.2. Цикл обучения PPO .....	215
7.5. Обучение агента PPO.....	217
7.5.1. PPO в Pong.....	217
7.5.2. PPO в BipedalWalker .....	220
7.6. Результаты экспериментов.....	223
7.6.1. Эксперимент по определению влияния $\lambda$ в GAE.....	223
7.6.2. Эксперимент по определению влияния переменной $\epsilon$ для усеченной функции потерь .....	225
7.7. Резюме .....	226
7.8. Рекомендуемая литература .....	227
<b>Глава 8. Методы параллелизации.....</b>	<b>228</b>
8.1. Синхронная параллелизация .....	229
8.2. Асинхронная параллелизация.....	230
8.2.1. Hogwild!.....	232
8.3. Обучение агента АЗС.....	234
8.4. Резюме .....	237
8.5. Рекомендуемая литература .....	238
<b>Глава 9. Сравнительный анализ алгоритмов.....</b>	<b>239</b>

### Часть III. Практика

<b>Глава 10. Начало работы с глубоким RL.....</b>	<b>242</b>
10.1. Приемы проектирования программ.....	242
10.1.1. Модульное тестирование.....	243
10.1.2. Качество кода.....	248
10.1.3. Рабочий процесс Git .....	250
10.2. Рекомендации по отладке.....	252
10.2.1. Признаки жизни.....	253
10.2.2. Диагностирование градиента стратегии .....	254
10.2.3. Диагностирование данных .....	254
10.2.4. Предварительная обработка .....	256

10.2.5. Память .....	256
10.2.6. Алгоритмические функции .....	256
10.2.7. Нейронные сети .....	257
10.2.8. Упрощение алгоритма .....	260
10.2.9. Упрощение задачи .....	260
10.2.10. Гиперпараметры .....	261
10.2.11. Рабочий процесс в SLM Lab .....	261
10.3. Практические приемы в играх Atari .....	263
10.4. Справочник по глубокому обучению с подкреплением .....	266
10.4.1. Таблицы гиперпараметров .....	266
10.4.2. Сравнение производительности алгоритмов .....	269
10.5. Резюме .....	273
<b>Глава 11. SLM Lab .....</b>	<b>274</b>
11.1. Алгоритмы, реализованные в SLM Lab .....	274
11.2. Файл spec .....	277
11.2.1. Синтаксис поиска в spec .....	279
11.3. Запуск SLM Lab .....	282
11.3.1. Команды SLM Lab .....	283
11.4. Анализ результатов эксперимента .....	283
11.4.1. Обзор экспериментальных данных .....	283
11.5. Резюме .....	285
<b>Глава 12. Архитектура сетей .....</b>	<b>286</b>
12.1. Виды нейронных сетей .....	286
12.1.1. Многослойные перцептроны .....	287
12.1.2. Сверточные нейронные сети .....	289
12.1.3. Рекуррентные нейронные сети .....	291
12.2. Рекомендации по выбору семейства сетей .....	293
12.2.1. Сравнение МППР и частично наблюдаемых МППР .....	293
12.2.2. Выбор сетей для сред .....	296
12.3. Net API .....	300
12.3.1. Выведение размерностей входного и выходного слоев .....	302
12.3.2. Автоматическое создание сети .....	304
12.3.3. Шаг обучения .....	307
12.3.4. Предоставление базовых методов .....	308
12.4. Резюме .....	309
12.5. Рекомендуемая литература .....	309

<b>Глава 13. Аппаратное обеспечение</b> .....	311
13.1. Компьютер .....	311
13.2. Типы данных .....	317
13.3. Оптимизация типов данных в RL.....	320
13.4. Выбор аппаратного обеспечения.....	325
13.5. Резюме.....	326

## **Часть IV. Проектирование сред**

<b>Глава 14. Состояния</b> .....	328
14.1. Примеры состояний.....	328
14.2. Полнота состояния .....	336
14.3. Сложность состояния .....	337
14.4. Потеря информации о состоянии.....	343
14.4.1. Преобразование изображений в градации серого .....	343
14.4.2. Дискретизация.....	344
14.4.3. Конфликты хеширования.....	344
14.4.4. Потери метаинформации.....	345
14.5. Предварительная обработка .....	348
14.5.1. Стандартизация.....	349
14.5.2. Предварительная обработка изображений.....	351
14.5.3. Предварительная обработка временных данных .....	353
14.6. Резюме.....	357
<b>Глава 15. Действия</b> .....	358
15.1. Примеры действий.....	358
15.2. Полнота действий.....	361
15.3. Сложность действий.....	364
15.4. Резюме.....	369
15.5. Проектирование действий в повседневной жизни .....	369
<b>Глава 16. Вознаграждения</b> .....	374
16.1. Роль вознаграждений.....	374
16.2. Рекомендации по проектированию вознаграждений .....	376
16.3. Резюме.....	382
<b>Глава 17. Функция переходов</b> .....	383
17.1. Проверка осуществимости .....	383
17.2. Проверка реалистичности .....	386
17.3. Резюме .....	388

<i>Заключение</i> .....	389
Воспроизводимость .....	389
Отрыв от реальности.....	390
Метаобучение и многозадачное обучение .....	390
Многоагентные задачи.....	391
Эффективность выборки .....	391
Обобщение.....	391
Исследование и структурирование вознаграждений .....	392

## **Приложения**

<b>Приложение А.</b> История глубокого обучения с подкреплением .....	394
<b>Приложение Б.</b> Примеры сред .....	397
Б.1. Дискретные среды.....	398
Б.1.1. CartPole-v0 .....	398
Б.1.2. MountainCar-v0.....	399
Б.1.3. LunarLander-v2.....	400
Б.1.4. PongNoFrameskip-v4 .....	401
Б.1.5. BreakoutNoFrameskip-v4.....	402
Б.2. Непрерывные среды.....	402
Б.2.1. Pendulum-v0 .....	402
Б.2.2. BipedalWalker-v2.....	403
Список используемых источников .....	405