

УДК 004.85  
ББК 32.971.3  
Л93

Лю Ю. (Х.)  
Л93 Обучение с подкреплением на PyTorch: сборник рецептов / пер. с англ.  
А. А. Слинкина. – М.: ДМК Пресс, 2020. – 282 с.: ил.

**ISBN 978-5-97060-853-1**

Библиотека PyTorch выходит на передовые позиции в качестве средства обучения с подкреплением (ОП) благодаря эффективности и простоте ее использования. Эта книга организована как справочник по работе с PyTorch, охватывающий широкий круг тем – от самых азов (настройка рабочей среды) до практических задач (рассмотрение ОП на конкретных примерах).

Вы научитесь использовать алгоритм «многоруких бандитов» и аппроксимацию функций; узнаете, как победить в играх Atari с помощью глубоких Q-сетей и как эффективно реализовать метод градиента стратегии; увидите, как применить метод ОП к игре в блэкджек, к окружающим средам в сеточном мире, к оптимизации рекламы в интернете и к игре Flappy Bird.

Издание предназначено для специалистов по искусственному интеллекту, которым требуется помощь в решении задач ОП. Для изучения материала необходимо знакомство с концепциями машинного обучения; опыт работы с библиотекой PyTorch необязателен, но желателен.

УДК 004.85  
ББК 32.971.3

First published in the English language under the title 'PyTorch 1.x Reinforcement Learning Cookbook Russian language edition copyright © 2020 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-83855-196-4 (англ.)  
ISBN 978-5-97060-853-1 (рус.)

Copyright © Packt Publishing 2019  
© Оформление, издание, перевод,  
ДМК Пресс, 2020

# Содержание

<b>Об авторе</b> .....	12
<b>О рецензентах</b> .....	13
<b>Предисловие</b> .....	14
 <b>Глава 1. Приступаем к обучению с подкреплением и PyTorch</b> .....	19
Подготовка среды разработки .....	19
Как это делается .....	20
Как это работает .....	21
Это еще не все .....	21
Установка OpenAI Gym .....	22
Как это делается .....	23
Как это работает .....	23
Это еще не все .....	23
Окружающие среды Atari .....	24
Как это делается .....	24
Как это работает .....	27
Это еще не все .....	28
Окружающая среда CartPole .....	29
Как это делается .....	30
Как это работает .....	32
Это еще не все .....	32
Основы PyTorch .....	33
Как это делается .....	33
Это еще не все .....	36
Реализация и оценивание стратегии случайного поиска .....	36
Как это делается .....	36
Как это работает .....	39
Это еще не все .....	39
Алгоритм восхождения на вершину .....	41
Как это делается .....	42

Как это работает .....	46
Это еще не все .....	46
Алгоритм градиента стратегии .....	47
Как это делается .....	48
Как это работает .....	51
Это еще не все .....	52

## Глава 2. Марковские процессы принятия решений и динамическое программирование .....53

Технические требования .....	53
Создание марковской цепи .....	54
Как это делается .....	54
Как это работает .....	55
Это еще не все .....	57
Создание МППР .....	57
Как это делается .....	58
Как это работает .....	59
Это еще не все .....	60
Оценивание стратегии .....	60
Как это делается .....	61
Как это работает .....	62
Это еще не все .....	63
Имитация окружающей среды FrozenLake .....	66
Подготовка .....	66
Как это делается .....	66
Как это работает .....	68
Это еще не все .....	69
Решение МППР с помощью алгоритма итерации по ценности .....	70
Как это делается .....	70
Как это работает .....	72
Это еще не все .....	73
Решение МППР с помощью алгоритма итерации по стратегиям .....	74
Как это делается .....	75
Как это работает .....	77
Это еще не все .....	77
Игра с подбрасыванием монеты .....	78
Как это делается .....	79
Как это работает .....	83
Это еще не все .....	85

<b>Глава 3. Применение методов Монте-Карло для численного оценивания</b> .....	87
Вычисление $\pi$ методом Монте-Карло .....	88
Как это делается.....	88
Как это работает .....	89
Это еще не все .....	90
Оценивание стратегии методом Монте-Карло .....	92
Как это делается.....	92
Как это работает .....	94
Это еще не все .....	94
Предсказание методом Монте-Карло в игре блэкджек .....	95
Как это делается.....	96
Как это работает .....	98
Это еще не все .....	99
Управление методом Монте-Карло с единой стратегией .....	101
Как это делается.....	102
Как это работает .....	104
Это еще не все .....	106
Разработка управления методом Монте-Карло с $\varepsilon$ -жадной стратегией .....	108
Как это делается.....	108
Как это работает .....	111
Управление методом Монте-Карло с разделенной стратегией .....	111
Как это делается.....	112
Как это работает .....	114
Это еще не все .....	115
Разработка управления методом Монте-Карло со взвешенной выборкой по значимости .....	116
Как это делается.....	116
Как это работает .....	117
Это еще не все .....	118
 <b>Глава 4. TD-обучение и Q-обучение</b> .....	119
Подготовка окружающей среды Cliff Walking.....	119
Подготовка .....	120
Как это делается.....	120
Как это работает .....	122
Реализация алгоритма Q-обучения.....	122
Как это делается.....	123
Как это работает .....	124
Это еще не все .....	125
Подготовка окружающей среды Windy Gridworld .....	127
Как это делается.....	128
Как это работает .....	132

Реализация алгоритма SARSA.....	132
Как это делается.....	132
Как это работает .....	134
Это еще не все.....	134
Решение задачи о такси методом Q-обучения .....	136
Подготовка .....	137
Как это делается.....	137
Как это работает .....	140
Решение задачи о такси методом SARSA.....	142
Как это делается.....	142
Как это работает .....	143
Это еще не все.....	144
Реализация алгоритма двойного Q-обучения.....	146
Как это делается.....	146
Как это работает .....	148

## **Глава 5. Решение задачи о многоруком бандите..... 150**

Создание окружающей среды с многоруким бандитом .....	150
Как это делается.....	151
Как это работает .....	152
Решение задачи о многоруком бандите с помощью $\epsilon$ -жадной стратегии .....	153
Как это делается.....	154
Как это работает .....	155
Это еще не все.....	156
Решение задачи о многоруком бандите с помощью softmax-исследования .....	156
Как это делается.....	157
Как это работает .....	158
Решение задачи о многоруком бандите с помощью алгоритма верхней доверительной границы .....	159
Как это делается.....	160
Как это работает .....	161
Это еще не все.....	162
Решение задачи о рекламе в интернете с помощью алгоритма многорукого бандита .....	162
Как это делается.....	163
Как это работает .....	164
Решение задачи о многоруком бандите с помощью выборки Томпсона.....	165
Как это делается.....	166
Как это работает .....	171
Решение задачи о рекламе в интернете с помощью контекстуальных бандитов.....	172
Как это делается.....	173
Как это работает .....	175

<b>Глава 6. Масштабирование с помощью аппроксимации функций.....</b>	<b>177</b>
Подготовка окружающей среды Mountain Car .....	178
Подготовка .....	179
Как это делается.....	179
Как это работает .....	180
Оценивание Q-функций посредством аппроксимации методом градиентного спуска.....	180
Как это делается.....	181
Как это работает .....	184
Реализация Q-обучения с линейной аппроксимацией функций .....	185
Как это делается.....	185
Как это работает .....	187
Реализация SARSA с линейной аппроксимацией функций .....	188
Как это делается.....	189
Как это работает .....	190
Пакетная обработка с применением буфера воспроизведения опыта .....	191
Как это делается.....	192
Как это работает .....	194
Реализация Q-обучения с аппроксимацией функций нейронной сетью.....	195
Как это делается.....	195
Как это работает .....	197
Решение задачи о балансировании стержня с помощью аппроксимации функций .....	198
Как это делается.....	198
Как это работает .....	199
 <b>Глава 7. Глубокие Q-сети в действии .....</b>	<b>200</b>
Реализация глубоких Q-сетей.....	200
Как это делается.....	201
Как это работает .....	204
Улучшение DQN с помощью воспроизведения опыта.....	206
Как это делается.....	207
Как это работает .....	209
Реализация алгоритма Double DQN .....	210
Как это делается.....	211
Как это работает .....	214
Настройка гиперпараметров алгоритма Double DQN для среды CartPole .....	215
Как это делается.....	216
Как это работает .....	217
Реализация алгоритма Dueling DQN .....	218
Как это делается.....	219
Как это работает .....	220

Применение DQN к играм Atari.....	221
Как это делается.....	223
Как это работает .....	226
Использование сверточных нейронных сетей в играх Atari .....	227
Как это делается.....	227
Как это работает .....	230

## Глава 8. Реализация методов градиента стратегии и оптимизация стратегии.....

232

Реализация алгоритма REINFORCE .....	232
Как это делается.....	233
Как это работает .....	236
Реализация алгоритма REINFORCE с базой .....	238
Как это делается.....	238
Как это работает .....	241
Реализация алгоритма исполнитель–критик.....	242
Как это делается.....	243
Как это работает .....	246
Решение задачи о блуждании на краю обрыва с помощью алгоритма исполнитель–критик.....	248
Как это делается.....	248
Как это работает .....	251
Подготовка непрерывной окружающей среды Mountain Car.....	252
Как это делается.....	253
Как это работает .....	254
Решение непрерывной задачи о блуждании на краю обрыва методом A2C .....	254
Как это делается.....	254
Как это работает .....	257
Это еще не все .....	259
Решение задачи о балансировании стержня методом перекрестной энтропии .....	260
Как это делается.....	260
Как это работает .....	262

## Глава 9. Кульминационный проект – применение DQN к игре Flappy Bird .....

264

Подготовка игровой среды .....	264
Подготовка .....	265
Как это делается.....	265
Как это работает .....	269

Построение глубокой Q-сети для игры Flappy Bird .....	269
Как это делается.....	270
Как это работает .....	272
Обучение и настройка сети.....	273
Как это делается.....	273
Как это работает .....	275
Развертывание модели и игра .....	276
Как это делается.....	276
Как это работает .....	277
 <b>Предметный указатель .....</b>	 278