

Краткое содержание

| | |
|--|-----|
| Об авторе | 16 |
| О редакторах | 17 |
| Предисловие к русскоязычному изданию | 18 |
| Предисловие | 20 |
| Глава 1. Что такое обучение с подкреплением | 25 |
| Глава 2. OpenAI Gym | 49 |
| Глава 3. Глубокое обучение с помощью PyTorch | 70 |
| Глава 4. Метод кросс-энтропии | 96 |
| Глава 5. Динамическое программирование и уравнение Беллмана | 115 |
| Глава 6. Глубокие Q-сети | 133 |
| Глава 7. Расширения для DQN | 166 |
| Глава 8. Торговля акциями с использованием обучения с подкреплением | 221 |
| Глава 9. Градиенты по стратегиям | 243 |
| Глава 10. Метод актора-критика | 264 |
| Глава 11. Асинхронный метод актора-критика | 281 |
| Глава 12. Тренировка чат-ботов с помощью обучения с подкреплением | 298 |
| Глава 13. Веб-навигация | 343 |
| Глава 14. Непрерывное пространство действий | 385 |
| Глава 15. Доверительные области — TRPO, PPO и ACKTR | 412 |
| Глава 16. Оптимизация методом черного ящика в RL | 427 |
| Глава 17. Методы, основанные на моделях среды: воображение | 449 |
| Глава 18. AlphaGo Zero | 471 |
| Заключение | 492 |

Оглавление

| | |
|--|----|
| Краткое содержание | 5 |
| Об авторе | 16 |
| О редакторах | 17 |
| Предисловие к русскоязычному изданию | 18 |
| Предисловие | 20 |
| Для кого эта книга | 21 |
| Структура издания | 21 |
| Извлеките максимум из этой книги | 22 |
| Скачивание кода для примеров | 23 |
| Скачивание цветных изображений | 23 |
| Условные обозначения | 23 |
| От издательства | 24 |
| Глава 1. Что такое обучение с подкреплением | 25 |
| Обучение с учителем, без учителя и с подкреплением | 26 |
| Зависимости и отношения в обучении с подкреплением | 29 |
| Вознаграждение | 30 |
| Агент | 31 |
| Среда | 32 |
| Действия | 32 |
| Наблюдения | 33 |
| Марковские процессы принятия решений | 36 |
| Марковский процесс | 36 |
| Марковский процесс с вознаграждением | 41 |
| Марковский процесс принятия решений | 44 |
| Резюме | 48 |

| | |
|--|----|
| Глава 2. OpenAI Gym | 49 |
| Структура агента | 49 |
| Аппаратные и программные требования | 51 |
| OpenAI Gym API | 53 |
| Пространство действий | 53 |
| Пространство наблюдений | 54 |
| Среда | 56 |
| Создание среды | 57 |
| Сеанс CartPole | 59 |
| Агент CartPole, действующий случайным образом | 61 |
| Дополнительный функционал Gym — обертки и мониторы | 63 |
| Обертки | 63 |
| Класс Monitor | 65 |
| Резюме | 69 |
| Глава 3. Глубокое обучение с помощью PyTorch | 70 |
| Тензоры | 70 |
| Создание тензоров | 71 |
| Скалярные тензоры | 73 |
| Операции над тензорами | 74 |
| Тензоры с поддержкой графического процессора | 74 |
| Градиенты | 75 |
| Базовые элементы нейронных сетей | 79 |
| Пользовательские слои | 81 |
| Последние связующие элементы — функции потерь и оптимизаторы | 84 |
| Функции потерь | 84 |
| Оптимизаторы | 85 |
| Мониторинг с TensorBoard | 87 |
| Основы TensorBoard | 88 |
| Вывод графиков | 89 |
| Пример: GAN на изображениях Atari | 90 |
| Резюме | 95 |
| Глава 4. Метод кросс-энтропии | 96 |
| Классификация методов глубокого обучения | 96 |
| Практическое применение метода кросс-энтропии | 98 |

| | |
|--|------------|
| Метод кросс-энтропии в CartPole..... | 100 |
| Метод кросс-энтропии в FrozenLake | 108 |
| Теоретические основы метода кросс-энтропии | 113 |
| Резюме | 114 |
| Глава 5. Динамическое программирование и уравнение Беллмана | 115 |
| Ценность, состояние и оптимальность | 115 |
| Уравнение Беллмана для оптимального управления | 117 |
| Ценность действия | 120 |
| Метод итерации по ценностям | 122 |
| Итерация по ценностям на практике..... | 124 |
| Q-обучение для FrozenLake..... | 130 |
| Резюме | 132 |
| Глава 6. Глубокие Q-сети | 133 |
| Итерация по ценностям в реальности | 133 |
| Табличное Q-обучение | 135 |
| Глубокое Q-обучение..... | 139 |
| Взаимодействие со средой | 140 |
| Оптимизация SGD | 141 |
| Корреляция между переходами..... | 142 |
| Марковское свойство | 143 |
| Окончательный вид обучения DQN | 143 |
| DQN в Pong..... | 144 |
| Обертки | 145 |
| Модель DQN..... | 150 |
| Обучение | 152 |
| Запуск и выполнение | 161 |
| Модель в действии..... | 162 |
| Резюме | 165 |
| Глава 7. Расширения для DQN | 166 |
| Библиотека PyTorch Agent Net..... | 167 |
| Агент | 167 |
| Опыт агента..... | 169 |
| Буфер примеров | 170 |
| Обертки среды Gym | 170 |

| | |
|--|------------|
| Базовая DQN..... | 170 |
| N-шаговые DQN | 177 |
| Реализация | 180 |
| Двойные DQN | 182 |
| Реализация | 182 |
| Результаты | 185 |
| Зашумленные сети..... | 187 |
| Реализация | 187 |
| Результаты | 191 |
| Приоритизированный буфер примеров | 192 |
| Реализация | 193 |
| Результаты | 198 |
| Дуальная DQN | 198 |
| Реализация | 200 |
| Результаты | 201 |
| Категориальные DQN..... | 201 |
| Реализация | 204 |
| Результаты | 211 |
| Объединение всех методов..... | 213 |
| Реализация | 214 |
| Результаты | 218 |
| Резюме | 219 |
| Литература..... | 219 |
| Глава 8. Торговля акциями с использованием обучения с подкреплением | 221 |
| Торговля..... | 221 |
| Данные..... | 222 |
| Постановка задачи и ключевые решения | 223 |
| Торговая среда | 225 |
| Модели | 232 |
| Код обучения..... | 234 |
| Результаты | 234 |
| Полносвязная модель | 234 |
| Сверточная модель | 238 |
| Дальнейшие эксперименты | 241 |
| Резюме | 242 |

| | |
|--|-----|
| Глава 9. Градиенты по стратегиям | 243 |
| Ценности и стратегия | 243 |
| Почему стратегия? | 244 |
| Представление стратегии..... | 244 |
| Градиенты по стратегиям..... | 245 |
| Метод REINFORCE | 246 |
| Пример с CartPole..... | 247 |
| Результаты | 251 |
| Методы, основанные на стратегиях, в сравнении с методами, основанными на ценностях..... | 252 |
| Ограничения метода REINFORCE..... | 253 |
| Требование полных эпизодов | 253 |
| Высокая дисперсия градиентов | 254 |
| Исследование | 254 |
| Корреляция между примерами..... | 255 |
| Градиенты по стратегиям в CartPole..... | 255 |
| Результаты | 258 |
| Градиенты по стратегиям в Pong..... | 260 |
| Результаты | 261 |
| Резюме | 263 |
| | |
| Глава 10. Метод актора-критика | 264 |
| Понижение дисперсии | 264 |
| Дисперсия в CartPole..... | 266 |
| Актор-критик | 268 |
| A2C в Pong..... | 271 |
| Результаты A2C в Pong | 276 |
| Настройка гиперпараметров | 278 |
| Скорость обучения..... | 279 |
| β -энтропия..... | 279 |
| Количество сред | 280 |
| Размер обучающего набора | 280 |
| Резюме | 280 |

| | |
|--|-----|
| Глава 11. Асинхронный метод актора-критика | 281 |
| Корреляция и эффективность использования данных..... | 281 |
| Добавление еще одного A в A2C..... | 282 |
| Многопроцессорная обработка в Python..... | 285 |
| Параллелизм на уровне данных в A3C | 285 |
| Результаты | 291 |
| Параллелизм на уровне градиентов в A3C | 291 |
| Результаты | 296 |
| Резюме | 297 |
| Глава 12. Тренировка чат-ботов с помощью обучения с подкреплением | 298 |
| Обзор чат-ботов | 298 |
| Основы глубокого NLP | 301 |
| Рекуррентные нейронные сети..... | 301 |
| Эмбединги | 303 |
| Кодировщик-декодировщик | 304 |
| Обучение seq2seq..... | 305 |
| Обучение с использованием максимального правдоподобия | 305 |
| Оценка Bilingual evaluation understudy | 308 |
| RL в seq2seq | 308 |
| Самокритичное обучение на последовательностях..... | 310 |
| Пример чат-бота | 311 |
| Структура примера | 311 |
| Модули cornell.py и data.py | 312 |
| Оценка BLEU и utils.py..... | 314 |
| Модель | 314 |
| Обучение: перекрестная энтропия..... | 321 |
| Выполнение обучения..... | 325 |
| Проверка данных | 327 |
| Тестирование обученной модели | 328 |
| Обучение: SCST | 330 |
| Обучение SCST | 337 |
| Результаты | 338 |
| Бот для Telegram..... | 339 |
| Резюме | 342 |

| | |
|---|-----|
| Глава 13. Веб-навигация | 343 |
| Навигация в Интернете..... | 343 |
| Автоматизация браузеров и RL | 344 |
| Бенчмарк Mini World of Bits..... | 345 |
| Universe от OpenAI..... | 347 |
| Установка | 348 |
| Действия и наблюдения..... | 348 |
| Создание рабочей среды..... | 349 |
| Стабильность MiniWoB | 352 |
| Метод «одного клика» | 352 |
| Действия с сеткой..... | 352 |
| Разбор примера | 354 |
| Модель | 355 |
| Код обучения..... | 355 |
| Запуск контейнеров | 360 |
| Процесс обучения..... | 362 |
| Проверка полученной в результате обучения стратегии | 364 |
| Проблемы метода одного щелчка | 365 |
| Демонстрационные примеры, выполненные человеком | 367 |
| Запись демонстрационных примеров | 368 |
| Формат записи..... | 370 |
| Обучение с использованием демонстрационных примеров..... | 373 |
| Результаты | 374 |
| Игра в крестики-нолики | 375 |
| Добавление текстового описания | 377 |
| Результаты | 382 |
| Стоит попробовать | 383 |
| Резюме | 384 |
| Глава 14. Непрерывное пространство действий | 385 |
| Почему непрерывное пространство?..... | 385 |
| Пространство действий..... | 386 |
| Среды..... | 386 |
| Метод актора-критика (A2C) | 389 |
| Реализация | 390 |
| Результаты | 393 |

| | |
|--|------------|
| Использование моделей и видеозаписей..... | 394 |
| Градиенты по детерминированным стратегиям..... | 395 |
| Исследование | 397 |
| Реализация | 397 |
| Результаты | 402 |
| Запись видео | 403 |
| Дистрибутивные градиенты по стратегиям | 403 |
| Архитектура..... | 404 |
| Реализация | 405 |
| Результаты | 409 |
| Стоит попробовать | 410 |
| Резюме | 411 |
| Глава 15. Доверительные области — TRPO, PPO и ACKTR | 412 |
| Введение | 412 |
| Roboschool | 413 |
| Производительность A2C | 413 |
| Результаты | 415 |
| Запись видео | 416 |
| Проксимальная оптимизация стратегии | 416 |
| Реализация | 417 |
| Результаты | 421 |
| Оптимизация стратегии по доверительной области | 422 |
| Реализация | 422 |
| Результаты | 423 |
| A2C с использованием ACKTR | 424 |
| Реализация | 425 |
| Результаты | 425 |
| Резюме | 426 |
| Глава 16. Оптимизация методом черного ящика в RL | 427 |
| Методы черного ящика | 427 |
| Эволюционные стратегии | 428 |
| Эволюционные стратегии в CartPole..... | 429 |
| Результаты | 433 |

| | |
|---|-----|
| Эволюционные стратегии в HalfCheetah | 434 |
| Результаты | 439 |
| Генетические алгоритмы | 440 |
| Генетические алгоритмы в CartPole..... | 441 |
| Результаты | 443 |
| Модификации генетических алгоритмов | 444 |
| Глубокий ГА..... | 444 |
| Поиск новизны..... | 444 |
| Генетический алгоритм в Cheetah | 445 |
| Результаты | 447 |
| Резюме | 448 |
| Литература..... | 448 |
| Глава 17. Методы, основанные на моделях среды: воображение | 449 |
| Сравнение безмодельных методов и методов, основанных на моделях..... | 449 |
| Недостатки моделей | 451 |
| Агент, дополненный воображением | 452 |
| Модель среды | 454 |
| Стратегия развертывания | 454 |
| Кодировщик развертываний..... | 455 |
| Результаты статьи..... | 455 |
| I2A в Breakout из Atari..... | 455 |
| Базовый агент A2C..... | 456 |
| Обучение EM..... | 457 |
| Агент с воображением | 460 |
| Результаты эксперимента | 465 |
| Базовый агент..... | 465 |
| Обучение весов EM | 467 |
| Обучение с моделью I2A..... | 468 |
| Резюме | 470 |
| Литература..... | 470 |
| Глава 18. AlphaGo Zero | 471 |
| Настольные игры | 471 |
| Метод AlphaGo Zero | 472 |
| Обзор | 472 |

| | |
|----------------------------------|-----|
| Поиск по дереву Монте-Карло..... | 474 |
| Самостоятельная игра..... | 475 |
| Обучение и оценка | 476 |
| Бот для Connect4 | 477 |
| Модель игры | 478 |
| Реализация MCTS..... | 480 |
| Модель | 485 |
| Обучение | 487 |
| Тестирование и сравнение..... | 488 |
| Результаты для Connect4 | 488 |
| Резюме | 491 |
| Литература | 491 |
| Заключение | 492 |