

Spis treści

O autorze	13
O recenzentach	14
Wstęp	15
Rozdział 1. Czym jest uczenie przez wzmacnianie	21
Uczenie nadzorowane	22
Uczenie nienadzorowane	22
Uczenie przez wzmacnianie	23
Trudności związane z uczeniem przez wzmacnianie	24
Formalne podstawy uczenia przez wzmacnianie	25
Nagroda	26
Agent	28
Środowisko	28
Akcje	28
Obserwacje	29
Teoretyczne podstawy uczenia przez wzmacnianie	32
Procesy decyzyjne Markowa	32
Polityka	42
Podsumowanie	42
Rozdział 2. Zestaw narzędzi OpenAI Gym	43
Anatomia agenta	43
Wymagania sprzętowe i programowe	46
Interfejs API biblioteki OpenAI Gym	47
Przestrzeń akcji	48
Przestrzeń obserwacji	48
Środowisko	50

Tworzenie środowiska	52
Sesja CartPole	54
Losowy agent dla środowiska CartPole	56
Dodatkowa funkcjonalność biblioteki Gym — opakowania i monitory	57
Opakowania	57
Monitory	59
Podsumowanie	62
Rozdział 3. Uczenie głębokie przy użyciu biblioteki PyTorch	63
Tensory	64
Tworzenie tensorów	64
Tensory skalarne	66
Operacje na tensorach	67
Tensory GPU	67
Gradienty	68
Tensory a gradienty	70
Bloki konstrukcyjne sieci neuronowych	73
Warstwy definiowane przez użytkownika	74
Funkcje straty i optymalizatory	77
Funkcje straty	77
Optymalizatory	78
Monitorowanie za pomocą narzędzia TensorBoard	79
Podstawy obsługi narzędzia TensorBoard	80
Narzędzia do tworzenia wykresów	82
Przykład — użycie sieci GAN z obrazami Atari	83
Biblioteka PyTorch Ignite	88
Zasady działania biblioteki Ignite	89
Podsumowanie	92
Rozdział 4. Metoda entropii krzyżowej	94
Taksonomia metod uczenia przez wzmacnianie	95
Praktyczne wykorzystanie entropii krzyżowej	96
Użycie entropii krzyżowej w środowisku CartPole	98
Użycie metody entropii krzyżowej w środowisku FrozenLake	106
Teoretyczne podstawy metody entropii krzyżowej	112
Podsumowanie	114
Rozdział 5. Uczenie tabelaryczne i równanie Bellmana	115
Wartość, stan i optymalność	116
Równanie optymalności Bellmana	118
Wartość akcji	120
Metoda iteracji wartości	123
Wykorzystanie iteracji wartości w praktyce	125
Q-uczenie w środowisku FrozenLake	131
Podsumowanie	132

Rozdział 6. Głębokie sieci Q	134
Rozwiązywanie realnego problemu z wykorzystaniem metody iteracji wartości	134
Q-uczenie tabelaryczne	136
Głębokie Q-uczenie	140
Interakcja ze środowiskiem	142
Optymalizacja za pomocą stochastycznego spadku wzdłuż gradientu (SGD)	142
Korelacja pomiędzy krokami	143
Własność Markowa	144
Ostateczna wersja procedury trenowania dla głębokich sieci Q	144
Użycie głębokiej sieci Q w grze Pong	145
Opakowania	146
Model głębokiej sieci Q	151
Trenowanie	152
Uruchomienie programu i sprawdzenie jego wydajności	160
Użycie modelu	163
Rzeczy do przetestowania	165
Podsumowanie	166
Rozdział 7. Biblioteki wyższego poziomu uczenia przez wzmacnianie	167
Dlaczego potrzebujemy bibliotek uczenia przez wzmacnianie?	168
Biblioteka PTAN	168
Selektory akcji	170
Agent	171
Źródło doświadczeń	175
Bufory doświadczeń	180
Klasa TargetNet	182
Klasy upraszczające współpracę z biblioteką Ignite	183
Rozwiązanie problemu środowiska CartPole za pomocą biblioteki PTAN	184
Inne biblioteki związane z uczeniem przez wzmacnianie	186
Podsumowanie	186
Rozdział 8. Rozszerzenia sieci DQN	187
Podstawowa, głęboka sieć Q	188
Wspólna biblioteka	188
Implementacja	192
Wyniki	194
Głęboka sieć Q o n krokach	195
Implementacja	198
Wyniki	198
Podwójna sieć DQN	199
Implementacja	200
Wyniki	201
Sieci zakłócone	203
Implementacja	203
Wyniki	206
Bufor priorytetowy	207
Implementacja	208
Wyniki	211

Rywalizujące sieci DQN	212
Implementacja	214
Wyniki	215
Kategoryczne sieci DQN	216
Implementacja	218
Wyniki	223
Połączenie wszystkich metod	226
Wyniki	226
Podsumowanie	228
Bibliografia	228
Rozdział 9. Sposoby przyspieszania metod uczenia przez wzmacnianie	230
Dlaczego prędkość ma znaczenie?	230
Model podstawowy	233
Wykres obliczeniowy w bibliotece PyTorch	235
Różne środowiska	238
Granie i trenowanie w oddzielnych procesach	240
Dostrajanie opakowań	244
Podsumowanie testów	248
Rozwiązanie ekstremalne: CuLE	250
Podsumowanie	250
Bibliografia	250
Rozdział 10. Inwestowanie na giełdzie za pomocą metod uczenia przez wzmacnianie	251
Handel	251
Dane	252
Określenie problemu i podjęcie kluczowych decyzji	253
Środowisko symulujące giełdę	255
Modele	262
Kod treningowy	263
Wyniki	264
Model ze sprzężeniem wyprzedzającym	264
Model konwolucyjny	269
Rzeczy do przetestowania	270
Podsumowanie	271
Rozdział 11. Alternatywa — gradienty polityki	272
Wartości i polityka	272
Dlaczego polityka?	273
Reprezentacja polityki	274
Gradienty polityki	274
Metoda REINFORCE	275
Przykład środowiska CartPole	276
Wyniki	280
Porównanie metod opartych na polityce z metodami opartymi na wartościach	281
Ograniczenia metody REINFORCE	282
Wymagane jest ukończenie epizodu	282
Wariancja dużych gradientów	283

Eksploracja	283
Korelacja danych	284
Zastosowanie metody gradientu polityki w środowisku CartPole	284
Implementacja	284
Wyniki	287
Zastosowanie metody gradientu polityki w środowisku Pong	289
Implementacja	291
Wyniki	293
Podsumowanie	294
Rozdział 12. Metoda aktor-krytyk	295
Zmniejszenie poziomu wariancji	295
Wariancja w środowisku CartPole	297
Aktor-krytyk	300
Użycie metody A2C w środowisku Pong	302
Wyniki użycia metody A2C w środowisku Pong	307
Dostrajanie hiperparametrów	310
Podsumowanie	312
Rozdział 13. Asynchroniczna wersja metody aktor-krytyk	313
Korelacja i wydajność próbkowania	313
Zrównoleglenie metody A2C	315
Przetwarzanie wieloprocessorowe w języku Python	317
Algorytm A3C wykorzystujący zrównoleglenie na poziomie danych	318
Implementacja	318
Wyniki	324
Algorytm A3C wykorzystujący zrównoleglenie na poziomie gradientów	325
Implementacja	326
Wyniki	330
Podsumowanie	332
Rozdział 14. Trenowanie chatbotów z wykorzystaniem uczenia przez wzmacnianie	333
Czym są chatboty?	334
Trenowanie chatbotów	334
Podstawy głębokiego przetwarzania języka naturalnego	336
Rekurencyjne sieci neuronowe	336
Osadzanie słów	338
Architektura koder-dekoder	339
Trenowanie modelu koder-dekoder	340
Trenowanie z wykorzystaniem logarytmu prawdopodobieństwa	340
Algorytm „Bilingual Evaluation Understudy” (BLEU)	342
Zastosowanie uczenia przez wzmacnianie w modelu koder-dekoder	343
Krytyczna analiza trenowania sekwencji	344
Projekt chatbota	345
Przykładowa struktura	346
Moduły cornell.py i data.py	346
Wskaźnik BLEU i moduł utils.py	348
Model	348

Eksploracja zbioru danych	354
Trenowanie — entropia krzyżowa	356
Implementacja	356
Wyniki	360
Trenowanie — metoda SCST	362
Implementacja	363
Wyniki	369
Przetestowanie modeli przy użyciu danych	371
Bot dla komunikatora Telegram	374
Podsumowanie	376
 Rozdział 15. Środowisko TextWorld	 378
Fikcja interaktywna	378
Środowisko	382
Instalacja	382
Generowanie gry	382
Przestrzenie obserwacji i akcji	384
Dodatkowe informacje o grze	386
Podstawowa sieć DQN	388
Wstępne przetwarzanie obserwacji	390
Osadzenia i kodery	394
Model DQN i agent	396
Kod treningowy	398
Wyniki trenowania	399
Model generujący polecenia	403
Implementacja	405
Wyniki uzyskane po wstępnym trenowaniu	409
Kod treningowy sieci DQN	410
Wyniki uzyskane po trenowaniu sieci DQN	412
Podsumowanie	413
 Rozdział 16. Nawigacja w sieci	 414
Nawigacja w sieci	414
Automatyzacja działań w przeglądarce i uczenie przez wzmacnianie	415
Test porównawczy MiniWoB	416
OpenAI Universe	417
Instalacja	419
Akcje i obserwacje	420
Tworzenie środowiska	420
Stabilność systemu MiniWoB	422
Proste klikanie	423
Akcje związane z siatką	423
Przegląd rozwiązania	425
Model	425
Kod treningowy	426
Uruchamianie kontenerów	431
Proces trenowania	432
Testowanie wyuczonej polityki	435
Problemy występujące podczas prostego klikania	436

Obserwacje ludzkich działań	438
Zapisywanie działań	439
Format zapisywanych danych	441
Trenowanie z wykorzystaniem obserwacji działań	443
Wyniki	445
Gra w kółko i krzyżyk	447
Dodawanie opisów tekstowych	452
Implementacja	452
Wyniki	456
Rzeczy do przetestowania	458
Podsumowanie	460
Rozdział 17. Ciągła przestrzeń akcji	461
Dlaczego jest potrzebna ciągła przestrzeń akcji?	461
Przestrzeń akcji	462
Środowiska	463
Metoda A2C	465
Implementacja	466
Wyniki	469
Użycie modeli i zapisywanie plików wideo	470
Deterministyczne gradienty polityki	471
Eksploracja	473
Implementacja	473
Wyniki	478
Nagrywanie plików wideo	480
Dystrybucyjne gradienty polityki	480
Architektura	481
Implementacja	481
Wyniki	485
Nagrania wideo	487
Rzeczy do przetestowania	487
Podsumowanie	487
Rozdział 18. Metody uczenia przez wzmacnianie w robotyce	488
Roboty i robotyka	489
Złożoność robota	491
Przegląd sprzętu	492
Platforma	493
Sensory	495
Siłowniki	496
Szkielet	497
Pierwszy cel trenowania	501
Emulator i model	503
Plik z definicją modelu	504
Klasa robota	508
Trenowanie zgodnie z algorytmem DDPG i uzyskane wyniki	513
Sterowanie sprzętem	516
MicroPython	517
Obsługa czujników	520

Sterowanie serwomechanizmami	531
Przenoszenie modelu do sprzętu	536
Połączenie wszystkiego w całość	542
Eksperymentowanie z polityką	545
Podsumowanie	545
Rozdział 19. Regiony zaufania — PPO, TRPO, ACKTR i SAC	547
Biblioteka Roboschool	548
Model bazowy A2C	549
Implementacja	549
Wyniki	551
Nagrywanie plików wideo	555
Algorytm PPO	555
Implementacja	556
Wyniki	559
Algorytm TRPO	561
Implementacja	562
Wyniki	563
Algorytm ACKTR	565
Implementacja	565
Wyniki	565
Algorytm SAC	566
Implementacja	567
Wyniki	569
Podsumowanie	571
Rozdział 20. Optymalizacja typu „czarna skrzynka” w przypadku uczenia przez wzmacnianie	572
Metody typu „czarna skrzynka”	572
Strategie ewolucyjne	573
Testowanie strategii ewolucyjnej w środowisku CartPole	574
Testowanie strategii ewolucyjnej w środowisku HalfCheetah	580
Algorytmy genetyczne	586
Testowanie algorytmu genetycznego w środowisku CartPole	587
Dostrajanie algorytmu genetycznego	589
Testowanie algorytmu genetycznego w środowisku HalfCheetah	590
Podsumowanie	593
Bibliografia	594
Rozdział 21. Zaawansowana eksploracja	595
Dlaczego eksploracja jest ważna?	595
Co złego jest w metodzie epsilonu zachłannego?	596
Alternatywne sposoby eksploracji	599
Sieci zakłócone	600
Metody oparte na liczebności	600
Metody oparte na prognozowaniu	601
Eksperymentowanie w środowisku MountainCar	602
Metoda DQN z wykorzystaniem strategii epsilonu zachłannego	603
Metoda DQN z wykorzystaniem sieci zakłóconych	605

Metoda DQN z licznikami stanów	607
Optymalizacja bliskiej polityki	609
Metoda PPO z wykorzystaniem sieci zakłóconych	611
Metoda PPO wykorzystująca eksplorację opartą na liczebności	614
Metoda PPO wykorzystująca destylację sieci	616
Eksperymentowanie ze środowiskami Atari	618
Metoda DQN z wykorzystaniem strategii epsilon zachłannego	619
Klasyczna metoda PPO	620
Metoda PPO z wykorzystaniem destylacji sieci	621
Metoda PPO z wykorzystaniem sieci zakłóconych	623
Podsumowanie	624
Bibliografia	624
Rozdział 22. Alternatywa dla metody bezmodelowej — agent wspomagany wyobraźnią	625
Metody oparte na modelu	626
Porównanie metody opartej na modelu z metodą bezmodelową	626
Niedoskonałości modelu	627
Agent wspomagany wyobraźnią	628
Model środowiskowy	630
Polityka wdrożenia	631
Koder wdrożeń	631
Wyniki zaprezentowane w artykule	631
Użycie modelu I2A w grze Breakout	631
Podstawowy agent A2C	632
Trenowanie modelu środowiskowego	633
Agent wspomagany wyobraźnią	636
Wyniki eksperymentów	641
Agent podstawowy	641
Trenowanie wag modelu środowiskowego	643
Trenowanie przy użyciu modelu I2A	645
Podsumowanie	647
Bibliografia	647
Rozdział 23. AlphaGo Zero	648
Gry planszowe	648
Metoda AlphaGo Zero	649
Wprowadzenie	650
Przeszukiwanie drzewa metodą Monte Carlo (MCTS)	651
Granie modelu z samym sobą	652
Trenowanie i ocenianie	653
Bot dla gry Czwórki	653
Model gry	654
Implementacja algorytmu przeszukiwania drzewa metodą Monte Carlo (MCTS)	656
Model	661
Trenowanie	663
Testowanie i porównywanie	663

Wyniki uzyskane w grze Czwórki	664
Podsumowanie	666
Bibliografia	667
Rozdział 24. Użycie metod uczenia przez wzmacnianie w optymalizacji dyskretnej	668
Rola uczenia przez wzmacnianie	668
Kostka Rubika i optymalizacja kombinatoryczna	669
Optymalność i liczba boska	670
Sposoby układania kostki	671
Reprezentacja danych	672
Akcje	672
Stany	673
Proces trenowania	676
Architektura sieci neuronowej	676
Trenowanie	678
Aplikacja modelowa	679
Wyniki	681
Analiza kodu	682
Środowiska kostki	683
Trenowanie	687
Proces wyszukiwania	688
Wyniki eksperymentu	689
Kostka 2×2	691
Kostka 3×3	693
Dalsze usprawnienia i eksperymenty	694
Podsumowanie	695
Rozdział 25. Metoda wieloagentowa	696
Na czym polega działanie metody wieloagentowej?	696
Formy komunikacji	697
Użycie uczenia przez wzmacnianie	698
Środowisko MAgent	698
Instalacja	698
Przegląd rozwiązania	699
Środowisko losowe	699
Głęboka sieć Q obsługująca tygrysy	704
Trenowanie i wyniki	707
Współpraca między tygrysami	709
Trenowanie tygrysów i jeleni	712
Walka pomiędzy równorzędnymi aktorami	714
Podsumowanie	714