

Deep Reinforcement Learning

Das umfassende
Praxis-Handbuch

Moderne Algorithmen für Chatbots, Robotik,
diskrete Optimierung und Web-Automatisierung
inkl. Multiagenten-Methoden

Inhaltsverzeichnis

Über den Autor	17
Über die Korrektoren.....	17
Über den Fachkorrektor der deutschen Ausgabe	18
Einleitung.....	19
Teil I Grundlagen des Reinforcement Learnings.....	24
1 Was ist Reinforcement Learning?.....	25
1.1 Überwachtes Lernen	25
1.2 Unüberwachtes Lernen	26
1.3 Reinforcement Learning	26
1.4 Herausforderungen beim Reinforcement Learning	28
1.5 RL-Formalismen	28
1.5.1 Belohnung	29
1.5.2 Der Agent.....	31
1.5.3 Die Umgebung	31
1.5.4 Aktionen.....	31
1.5.5 Beobachtungen	32
1.6 Die theoretischen Grundlagen des Reinforcement Learnings.....	34
1.6.1 Markov-Entscheidungsprozesse.....	35
1.6.2 Markov-Prozess	35
1.6.3 Markov-Belohnungsprozess	39
1.6.4 Aktionen hinzufügen	42
1.6.5 Policy	44
1.7 Zusammenfassung	45
2 OpenAI Gym	47
2.1 Aufbau des Agenten	47
2.2 Anforderungen an Hard- und Software.....	50
2.3 OpenAI-Gym-API	51
2.3.1 Aktionsraum	52
2.3.2 Beobachtungsraum.....	52
2.3.3 Die Umgebung	54
2.3.4 Erzeugen der Umgebung	55
2.3.5 Die CartPole-Sitzung.....	57
2.4 Ein CartPole-Agent nach dem Zufallsprinzip	59

2.5	Zusätzliche Gym-Funktionalität: Wrapper und Monitor	60
2.5.1	Wrapper	61
2.5.2	Monitor	63
2.6	Zusammenfassung	66
3	Deep Learning mit PyTorch	67
3.1	Tensoren	67
3.1.1	Tensoren erzeugen	68
3.1.2	Skalare Tensoren	70
3.1.3	Tensor-Operationen	71
3.1.4	GPU-Tensoren	71
3.2	Gradienten	72
3.2.1	Tensoren und Gradienten	74
3.3	NN-Bausteine.	76
3.4	Benutzerdefinierte Schichten.	78
3.5	Verlustfunktionen und Optimierer	80
3.5.1	Verlustfunktionen.	81
3.5.2	Optimierer	81
3.6	Monitoring mit TensorBoard	83
3.6.1	Einführung in TensorBoard.	84
3.6.2	Plotten	85
3.7	Beispiel: GAN für Bilder von Atari-Spielen.	87
3.8	PyTorch Ignite	92
3.8.1	Konzepte	93
3.9	Zusammenfassung	97
4	Das Kreuzentropie-Verfahren	99
4.1	Klassifikation von RL-Verfahren	99
4.2	Kreuzentropie in der Praxis	100
4.3	Kreuzentropie beim CartPole	102
4.4	Kreuzentropie beim FrozenLake	111
4.5	Theoretische Grundlagen des Kreuzentropie-Verfahrens	118
4.6	Zusammenfassung	119
Teil II	Wertebasierte Verfahren	120
5	Tabular Learning und das Bellman'sche Optimalitätsprinzip	121
5.1	Wert, Zustand und Optimalität	121
5.2	Das Bellman'sche Optimalitätsprinzip	123
5.3	Aktionswert	126
5.4	Wertiteration	128
5.5	Wertiteration in der Praxis	130
5.6	Q-Learning in der FrozenLake-Umgebung.	136
5.7	Zusammenfassung	138

6	Deep Q-Networks	139
6.1	Wertiteration in der Praxis	139
6.2	Tabular Q-Learning	140
6.3	Deep Q-Learning	145
	6.3.1 Interaktion mit der Umgebung	147
	6.3.2 SGD-Optimierung	147
	6.3.3 Korrelation der Schritte	148
	6.3.4 Die Markov-Eigenschaft	148
	6.3.5 Die endgültige Form des DQN-Trainings	149
6.4	DQN mit Pong	150
	6.4.1 Wrapper	151
	6.4.2 DQN-Modell	156
	6.4.3 Training	158
	6.4.4 Ausführung und Leistung	167
	6.4.5 Das Modell in Aktion	170
6.5	Weitere Möglichkeiten	172
6.6	Zusammenfassung	173
7	Allgemeine RL-Bibliotheken	175
7.1	Warum RL-Bibliotheken?	175
7.2	Die PTAN-Bibliothek	176
	7.2.1 Aktionsselektoren	177
	7.2.2 Der Agent	179
	7.2.3 Quelle der Erfahrungswerte	183
	7.2.4 Replay Buffer für Erfahrungswerte	189
	7.2.5 Die TargetNet-Klasse	191
	7.2.6 Hilfsfunktionen für Ignite	193
7.3	Lösung der CartPole-Umgebung mit PTAN	194
7.4	Weitere RL-Bibliotheken	196
7.5	Zusammenfassung	197
8	DQN-Erweiterungen	199
8.1	Einfaches DQN	199
	8.1.1 Die Bibliothek common	200
	8.1.2 Implementierung	205
	8.1.3 Ergebnisse	207
8.2	N-Schritt-DQN	208
	8.2.1 Implementierung	211
	8.2.2 Ergebnisse	211
8.3	Double DQN	212
	8.3.1 Implementierung	213
	8.3.2 Ergebnisse	215
8.4	Verrauschte Netze	216
	8.4.1 Implementierung	217
	8.4.2 Ergebnisse	219

8.5	Priorisierter Replay Buffer	220
8.5.1	Implementierung	221
8.5.2	Ergebnisse	225
8.6	Rivalisierendes DQN	227
8.6.1	Implementierung	228
8.6.2	Ergebnisse	229
8.7	Kategoriales DQN	230
8.7.1	Implementierung	232
8.7.2	Ergebnisse	239
8.8	Alles miteinander kombinieren	241
8.8.1	Ergebnisse	242
8.9	Zusammenfassung	243
8.10	Quellenangaben	244
9	Beschleunigung von RL-Verfahren	245
9.1	Die Bedeutung der Geschwindigkeit	245
9.2	Der Ausgangspunkt	248
9.3	Der Berechnungsgraph in PyTorch	250
9.4	Mehrere Umgebungen	252
9.5	Spielen und Trainieren in separaten Prozessen	255
9.6	Optimierung der Wrapper	259
9.7	Zusammenfassung der Benchmarks	265
9.8	Atari-Emulation: CuLE	265
9.9	Zusammenfassung	266
9.10	Quellenangaben	266
10	Aktienhandel per Reinforcement Learning	267
10.1	Börsenhandel	267
10.2	Daten	268
10.3	Aufgabenstellungen und Grundsatzentscheidungen	269
10.4	Die Handelsumgebung	270
10.5	Modelle	279
10.6	Trainingscode	281
10.7	Ergebnisse	281
10.7.1	Das Feedforward-Modell	281
10.7.2	Das Faltungsmodell	287
10.8	Weitere Möglichkeiten	288
10.9	Zusammenfassung	289
Teil III	Policybasierte Verfahren	290
11	Eine Alternative: Policy Gradients	291
11.1	Werte und Policy	291
11.1.1	Warum Policy?	292

11.1.2	Repräsentation der Policy	292
11.1.3	Policy Gradients	293
11.2	Das REINFORCE-Verfahren	294
11.2.1	Das CartPole-Beispiel	295
11.2.2	Ergebnisse	299
11.2.3	Policybasierte und wertebasierte Verfahren	300
11.3	Probleme mit REINFORCE	301
11.3.1	Notwendigkeit vollständiger Episoden	301
11.3.2	Große Varianz der Gradienten	302
11.3.3	Exploration	302
11.3.4	Korrelation zwischen Beispielen	303
11.4	PG mit CartPole	303
11.4.1	Implementierung	303
11.4.2	Ergebnisse	306
11.5	PG mit Pong	310
11.5.1	Implementierung	311
11.5.2	Ergebnisse	312
11.6	Zusammenfassung	313
12	Das Actor-Critic-Verfahren	315
12.1	Verringern der Varianz	315
12.2	Varianz der CartPole-Umgebung	317
12.3	Actor-Critic	320
12.4	A2C mit Pong	322
12.5	A2C mit Pong: Ergebnisse	328
12.6	Optimierung der Hyperparameter	331
12.6.1	Lernrate	332
12.6.2	Beta	333
12.6.3	Anzahl der Umgebungen	333
12.6.4	Batchgröße	333
12.7	Zusammenfassung	333
13	Asynchronous Advantage Actor Critic	335
13.1	Korrelation und Stichprobeneffizienz	335
13.2	Ein weiteres A zu A2C hinzufügen	336
13.3	Multiprocessing in Python	339
13.4	A3C mit Datenparallelität	339
13.4.1	Implementierung	339
13.4.2	Ergebnisse	346
13.5	A3C mit Gradientenparallelität	347
13.5.1	Implementierung	348
13.5.2	Ergebnisse	353
13.6	Zusammenfassung	354
14	Chatbot-Training per Reinforcement Learning	355
14.1	Chatbots – ein Überblick	355