

Inhaltsverzeichnis

Vorwort	xiii
Danksagungen	xv
Einleitung	xvii
 Teil I Neuronale Netze und Deep Learning	
1 Einbettungen, latenter Raum und Repräsentationen	3
1.1 Einbettungen	3
1.2 Latenter Raum	5
1.3 Repräsentation	6
1.4 Übungen	6
1.5 Referenzen	7
2 Selbstüberwachtes Lernen	9
2.1 Selbstüberwachtes Lernen vs. Transferlernen	9
2.2 Ungelabelte Daten nutzen	11
2.3 Selbstvorhersage und kontrastives selbstüberwachtes Lernen . .	12
2.4 Übungen	14
2.5 Referenzen	14
3 Few-Shot-Lernen	17
3.1 Datensätze und Terminologie	17
3.2 Übungen	20
4 Die Lotterie-Ticket-Hypothese	21
4.1 Das Lotterie-Ticket-Trainingsverfahren	21
4.2 Praktische Konsequenzen und Einschränkungen	22

4.3	Übungen	23
4.4	Referenzen	23
5	Überanpassung mit Daten verringern	25
5.1	Allgemeine Methoden	25
5.2	Übungen	28
5.3	Referenzen	28
6	Überanpassung durch Modellmodifikationen reduzieren	31
6.1	Allgemeine Methoden	31
6.2	Andere Methoden	36
6.3	Eine Regularisierungstechnik auswählen	37
6.4	Übungen	37
6.5	Referenzen	37
7	Multi-GPU-Trainingsparadigmen	39
7.1	Die Trainingsparadigmen	39
7.2	Empfehlungen	43
7.3	Übungen	44
7.4	Referenzen	44
8	Der Erfolg der Transformer	45
8.1	Der Aufmerksamkeitsmechanismus	45
8.2	Vortraining durch selbstüberwachtes Lernen	47
8.3	Große Anzahl von Parametern	47
8.4	Einfache Parallelisierung	48
8.5	Übungen	48
8.6	Referenzen	49
9	Generative KI-Modelle	51
9.1	Generative vs. diskriminative Modellierung	51
9.2	Arten von tiefen generativen Modellen	52
9.3	Empfehlungen	59
9.4	Übungen	60
9.5	Referenzen	60
10	Quellen der Zufälligkeit	61
10.1	Initialisierung der Modellgewichte	61
10.2	Sampling und Shuffling von Datensätzen	62

10.3	Nichtdeterministische Algorithmen	63
10.4	Verschiedene Laufzeitalgorithmen	63
10.5	Hardware und Treiber	64
10.6	Zufälligkeit und generative KI	65
10.7	Übungen.	67
10.8	Referenzen	67

Teil II **Computer Vision**

11	Die Anzahl der Parameter berechnen	71
11.1	Wie man die Anzahl der Parameter ermittelt	71
11.2	Praktische Anwendungen	75
11.3	Übungen.	75
12	Vollständig verbundene und konvolutionale Schichten	77
12.1	Szenario: Gleiche Größen von Kernel und Eingabe.	78
12.2	Szenario: Kernel-Größe ist 1.	79
12.3	Empfehlungen	79
12.4	Übungen.	80
13	Große Trainingsmengen für Vision Transformer	81
13.1	Induktive Verzerrungen in CNNs.	81
13.2	ViTs können CNNs übertreffen	85
13.3	Induktive Verzerrungen in ViTs	85
13.4	Empfehlungen	87
13.5	Übungen.	87
13.6	Referenzen	87

Teil III **Natural Language Processing**

14	Die Verteilungshypothese	91
14.1	Word2vec, BERT und GPT	92
14.2	Trifft die Hypothese zu?	93
14.3	Übungen.	94
14.4	Referenzen	94

15	Datenvermehrung für Text	95
15.1	Ersetzen von Synonymen	95
15.2	Löschen von Wörtern	96
15.3	Vertauschen von Wortpositionen	96
15.4	Sätze mischen	97
15.5	Rauschinjektion	97
15.6	Rückübersetzung	98
15.7	Synthetische Daten	98
15.8	Empfehlungen	99
15.9	Übungen	99
15.10	Referenzen	99
16	Selbstaufmerksamkeit	101
16.1	Aufmerksamkeit in RNNs	101
16.2	Der Selbstaufmerksamkeitsmechanismus	103
16.3	Übungen	104
16.4	Referenzen	105
17	Encoder- und Decoder-Transformer	107
17.1	Der ursprüngliche Transformer	107
17.2	Encoder-Decoder-Hybride	112
17.3	Terminologie	112
17.4	Aktuelle Transformer-Modelle	113
17.5	Übungen	114
17.6	Referenzen	114
18	Transformer verwenden und feinabstimmen	117
18.1	Transformer für Klassifizierungsaufgaben verwenden	117
18.2	Kontextbezogenes Lernen, Indizierung und Prompt- Feinabstimmung	120
18.3	Parametereffiziente Feinabstimmung	122
18.4	Reinforcement Learning mit menschlicher Rückmeldung . . .	127
18.5	Vortrainierte Sprachmodelle anpassen	128
18.6	Übungen	128
18.7	Referenzen	129

19	Generative LLMs evaluieren	131
19.1	Bewertungsmetriken für LLMs	131
19.2	Übungen	138
19.3	Referenzen	138

Teil IV Produktion und Deployment

20	Zustandsloses und zustandsbehaftetes Training	143
20.1	Zustandsloses (Re-)Training	143
20.2	Zustandsbehaftetes Training	144
20.3	Übungen	145
21	Datenzentrierte KI	147
21.1	Datenzentrierte vs. modellzentrierte KI	147
21.2	Empfehlungen	149
21.3	Übungen	150
21.4	Referenzen	150
22	Inferenz beschleunigen	151
22.1	Parallelisierung	151
22.2	Vektorisierung	152
22.3	Schleifenkachelung	153
22.4	Operatorfusion	154
22.5	Quantisierung	155
22.6	Übungen	156
22.7	Referenzen	156
23	Datenverteilungsverschiebungen	157
23.1	Kovariatenverschiebung	157
23.2	Labelverschiebung	158
23.3	Konzeptverschiebung	159
23.4	Domänenverschiebung	159
23.5	Arten von Datenverteilungsverschiebungen	160
23.6	Übungen	161
23.7	Referenzen	161

Teil V Vorhersageperformance und Modellevaluierung

24	Poisson- und ordinale Regression	165
24.1	Übungen	166
25	Konfidenzintervalle	167
25.1	Konfidenzintervalle definieren	167
25.2	Die Methoden	170
25.3	Empfehlungen	175
25.4	Übungen	175
25.5	Referenzen	175
26	Konfidenzintervalle vs. konforme Vorhersagen	177
26.1	Konfidenzintervalle und Vorhersageintervalle	177
26.2	Vorhersageintervalle und konforme Vorhersagen	178
26.3	Vorhersagebereiche, -intervalle und -mengen.	178
26.4	Konforme Vorhersagen berechnen	179
26.5	Beispiel für eine konforme Vorhersage	180
26.6	Die Vorteile der konformen Vorhersagen	181
26.7	Empfehlungen	182
26.8	Übungen	182
26.9	Referenzen	183
27	Geeignete Metriken	185
27.1	Die Kriterien	185
27.2	Der mittlere quadratische Fehler	186
27.3	Der Kreuzentropieverlust	188
27.4	Übungen	189
28	Das k in der k-fachen Kreuzvalidierung	191
28.1	Kompromisse bei der Auswahl von Werten für k	192
28.2	Geeignete Werte für k bestimmen	194
28.3	Übungen	194
28.4	Referenzen	195
29	Diskordanz zwischen Trainings- und Testdatensatz	197
29.1	Übungen	199

30	Begrenzte gelabelte Daten	201
30.1	Die Modellperformance mit begrenzten gelabelten Daten verbessern	201
30.2	Empfehlungen	210
30.3	Übungen.	211
30.4	Referenzen	212
	Nachwort	213
	Lösungen zu den Übungen	215
	Index	235