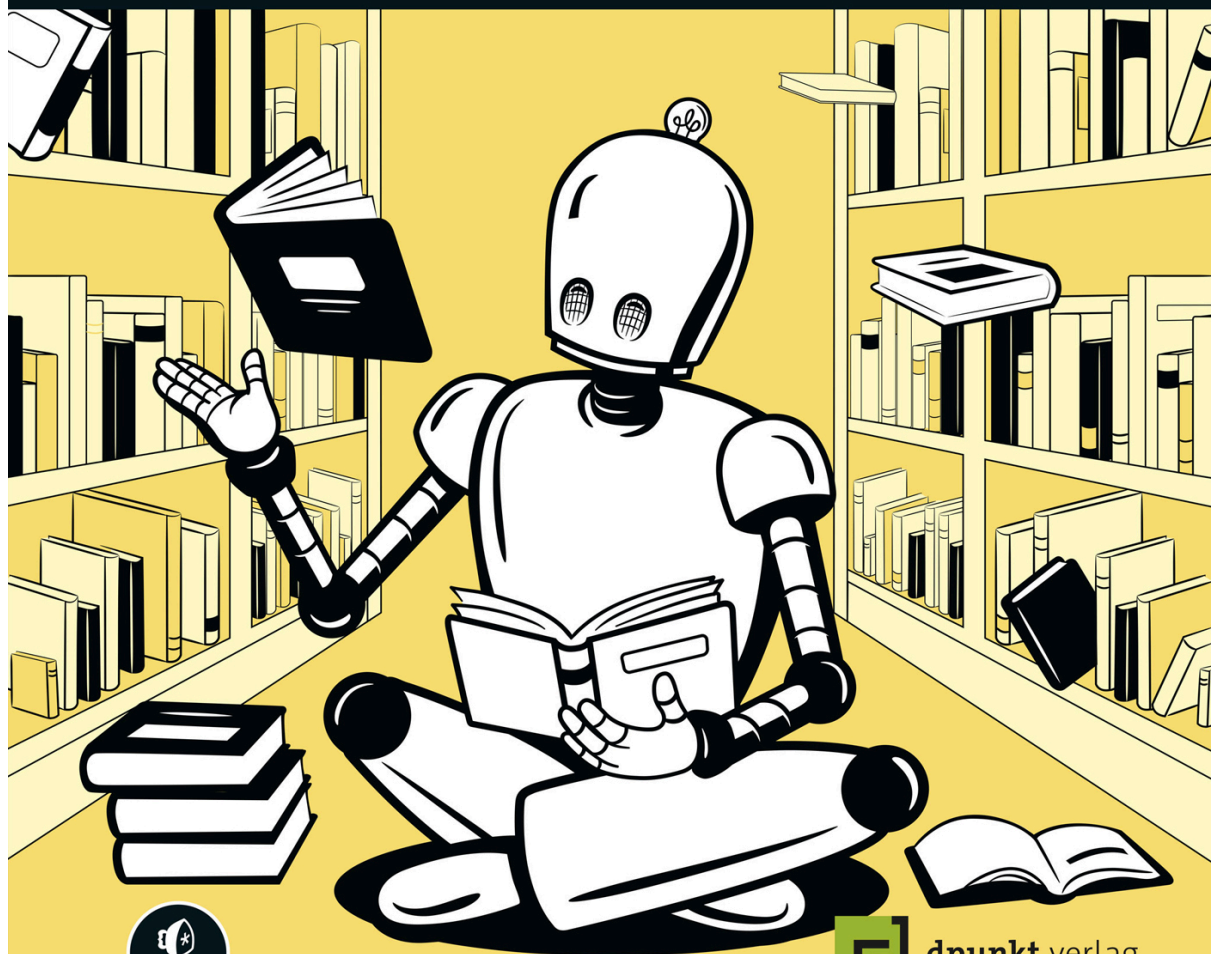


MACHINE LEARNING & KI

ZENTRALE KONZEPTE VERSTEHEN UND ANWENDEN

KOMPAKT

SEBASTIAN RASCHKA



dpunkt.verlag

Inhalt

Cover

Titel

Impressum

Inhalt

Vorwort

Danksagungen

Einleitung

Teil I Neuronale Netze und Deep Learning

1 Einbettungen, latenter Raum und Repräsentationen

1.1 Einbettungen

1.2 Latenter Raum

1.3 Repräsentation

1.4 Übungen

1.5 Referenzen

2 Selbstüberwachtes Lernen

2.1 Selbstüberwachtes Lernen vs. Transferlernen

2.2 Ungelabelte Daten nutzen

2.3 Selbstvorhersage und kontrastives selbstüberwachtes Lernen

2.4 Übungen

2.5 Referenzen

3 Few-Shot-Lernen

3.1 Datensätze und Terminologie

3.2 Übungen

4 Die Lotterie-Ticket-Hypothese

4.1 Das Lotterie-Ticket-Trainingsverfahren

4.2 Praktische Konsequenzen und Einschränkungen

4.3 Übungen

4.4 Referenzen

5 Überanpassung mit Daten verringern

5.1 Allgemeine Methoden

5.2 Übungen

5.3 Referenzen

6 Überanpassung durch Modellmodifikationen reduzieren

6.1 Allgemeine Methoden

6.2 Andere Methoden

6.3 Eine Regularisierungstechnik auswählen

6.4 Übungen

6.5 Referenzen

7 Multi-GPU-Trainingsparadigmen

7.1 Die Trainingsparadigmen

7.2 Empfehlungen

7.3 Übungen

7.4 Referenzen

8 Der Erfolg der Transformer

8.1 Der Aufmerksamkeitsmechanismus

8.2 Vortraining durch selbstüberwachtes Lernen

8.3 Große Anzahl von Parametern

8.4 Einfache Parallelisierung

8.5 Übungen

8.6 Referenzen

9 Generative KI-Modelle

9.1 Generative vs. diskriminative Modellierung

9.2 Arten von tiefen generativen Modellen

9.3 Empfehlungen

9.4 Übungen

9.5 Referenzen

10 Quellen der Zufälligkeit

10.1 Initialisierung der Modellgewichte

10.2 Sampling und Shuffling von Datensätzen

10.3 Nichtdeterministische Algorithmen

10.4 Verschiedene Laufzeitalgorithmen

10.5 Hardware und Treiber

10.6 Zufälligkeit und generative KI

10.7 Übungen

10.8 Referenzen

Teil II Computer Vision

11 Die Anzahl der Parameter berechnen

11.1 Wie man die Anzahl der Parameter ermittelt

11.2 Praktische Anwendungen

11.3 Übungen

12 Vollständig verbundene und konvolutionale Schichten

12.1 Szenario: Gleiche Größen von Kernel und Eingabe

12.2 Szenario: Kernel-Größe ist 1

12.3 Empfehlungen

12.4 Übungen

13 Große Trainingsmengen für Vision Transformer

13.1 Induktive Verzerrungen in CNNs

13.2 ViTs können CNNs übertreffen

13.3 Induktive Verzerrungen in ViTs

13.4 Empfehlungen

13.5 Übungen

13.6 Referenzen

Teil III Natural Language Processing

14 Die Verteilungshypothese

14.1 Word2vec, BERT und GPT

14.2 Trifft die Hypothese zu?

14.3 Übungen

14.4 Referenzen

15 Datenvermehrung für Text

15.1 Ersetzen von Synonymen

15.2 Löschen von Wörtern

15.3 Vertauschen von Wortpositionen

15.4 Sätze mischen

15.5 Rauschinjektion

15.6 Rückübersetzung

15.7 Synthetische Daten

15.8 Empfehlungen

15.9 Übungen

15.10 Referenzen

16 Selbstaufmerksamkeit

16.1 Aufmerksamkeit in RNNs

16.2 Der Selbstaufmerksamkeitsmechanismus

16.3 Übungen

16.4 Referenzen

17 Encoder- und Decoder-Transformer

17.1 Der ursprüngliche Transformer

17.2 Encoder-Decoder-Hybride

17.3 Terminologie

17.4 Aktuelle Transformer-Modelle

17.5 Übungen

17.6 Referenzen

18 Transformer verwenden und feinabstimmen

18.1 Transformer für Klassifizierungsaufgaben verwenden

18.2 Kontextbezogenes Lernen, Indizierung und Prompt-Feinabstimmung

18.3 Parametereffiziente Feinabstimmung

18.4 Reinforcement Learning mit menschlicher Rückmeldung

18.5 Vortrainierte Sprachmodelle anpassen

18.6 Übungen

18.7 Referenzen

19 Generative LLMs evaluieren

19.1 Bewertungsmetriken für LLMs

19.2 Übungen

19.3 Referenzen

Teil IV Produktion und Deployment

20 Zustandsloses und zustandsbehaftetes Training

20.1 Zustandsloses (Re-)Training

20.2 Zustandsbehaftetes Training

20.3 Übungen

21 Datenzentrierte KI

21.1 Datenzentrierte vs. modellzentrierte KI

21.2 Empfehlungen

21.3 Übungen

21.4 Referenzen

22 Inferenz beschleunigen

22.1 Parallelisierung

22.2 Vektorisierung

22.3 Schleifenkachelung

22.4 Operatorfusion

22.5 Quantisierung

22.6 Übungen

22.7 Referenzen

23 Datenverteilungsverschiebungen

23.1 Kovariatenverschiebung

23.2 Labelverschiebung

23.3 Konzeptverschiebung

23.4 Domänenverschiebung

23.5 Arten von Datenverteilungsverschiebungen

23.6 Übungen

23.7 Referenzen

Teil V Vorhersageperformance und Modellevaluierung

24 Poisson- und ordinale Regression

24.1 Übungen

25 Konfidenzintervalle

25.1 Konfidenzintervalle definieren

25.2 Die Methoden

25.3 Empfehlungen

25.4 Übungen

25.5 Referenzen

26 Konfidenzintervalle vs. konforme Vorhersagen

26.1 Konfidenzintervalle und Vorhersageintervalle

26.2 Vorhersageintervalle und konforme Vorhersagen

26.3 Vorhersagebereiche, -intervalle und -mengen

26.4 Konforme Vorhersagen berechnen

26.5 Beispiel für eine konforme Vorhersage

26.6 Die Vorteile der konformen Vorhersagen

26.7 Empfehlungen

26.8 Übungen

26.9 Referenzen

27 Geeignete Metriken

27.1 Die Kriterien

27.2 Der mittlere quadratische Fehler

27.3 Der Kreuzentropieverlust

27.4 Übungen

28 Das k in der k -fachen Kreuzvalidierung

28.1 Kompromisse bei der Auswahl von Werten für k

28.2 Geeignete Werte für k bestimmen

28.3 Übungen

28.4 Referenzen

29 Diskordanz zwischen Trainings- und Testdatensatz

29.1 Übungen

30 Begrenzte gelabelte Daten

30.1 Die Modellperformance mit begrenzten gelabelten Daten verbessern

30.2 Empfehlungen

30.3 Übungen

30.4 Referenzen

Nachwort

Lösungen zu den Übungen

Index

3 Few-Shot-Lernen

Was ist Few-Shot-Lernen? Wie unterscheidet es sich vom herkömmlichen Trainingsverfahren für überwachtes Lernen?

Few-Shot-Lernen ist eine Art des überwachten Lernens für kleine Trainingsdatensätze mit einem sehr kleinen Verhältnis von Beispielen zu Klassen. Beim regulären überwachten Lernen trainieren wir Modelle, indem wir über einem Trainingsdatensatz iterieren, wobei das Modell immer einen feststehenden Satz von Klassen sieht. Beim Few-Shot-Lernen arbeiten wir mit einem *Unterstützungsdatsatz*, aus dem wir mehrere Trainingsaufgaben erzeugen, um Trainingsepisoden zusammenzustellen, wobei jede Trainingsaufgabe aus verschiedenen Klassen besteht.

3.1 Datensätze und Terminologie

Beim überwachten Lernen passen wir ein Modell an einen Trainingsdatensatz an und bewerten es anhand eines Testdatensatzes. Der Trainingsdatensatz enthält normalerweise eine relativ große Anzahl von Beispielen pro Klasse. Zum Beispiel gilt im Zusammenhang mit überwachtem Lernen der Iris-Datensatz, der pro Klasse 50 Beispiele enthält, als winziger Datensatz. Andererseits wird bei Modellen für Deep Learning selbst ein Datensatz wie MNIST, der 5000 Trainingsbeispiele pro Klasse enthält, als sehr klein betrachtet.

Beim Few-Shot-Lernen ist die Anzahl der Beispiele wesentlich kleiner. Wenn wir die Few-Shot-Lernen-Aufgabe spezifizieren, sprechen wir normalerweise von *N-Way K-Shot*, wobei N für die Anzahl der Klassen und K für die Anzahl der Beispiele pro Klasse steht. Die gebräuchlichsten Werte sind $K = 1$ oder $K = 5$. Zum Beispiel gibt es bei einem 5-Way-1-Shot-Problem fünf Klassen mit jeweils nur einem Beispiel. Abbildung 3-1 zeigt eine 3-Way-1-Shot-Situation, um das Konzept anhand eines kleineren Beispiels zu verdeutlichen.

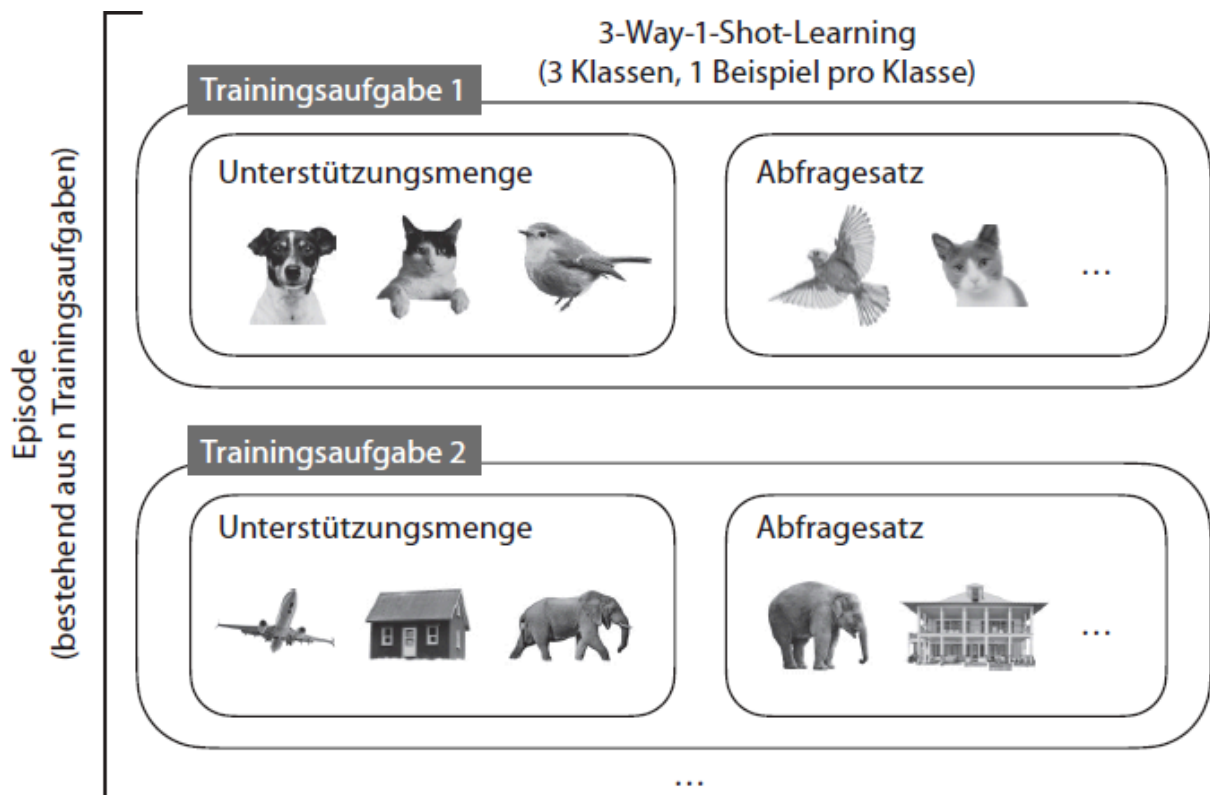


Abb. 3-1 Trainingsaufgaben im Few-Shot-Lernen

Anstatt das Modell an den Trainingsdatensatz anzupassen, können wir uns Few-Shot-Lernen als »Lernen zu lernen« vorstellen. Im Unterschied zum überwachten Lernen verwendet Few-Shot-Lernen keinen Trainingsdatensatz, sondern eine sogenannte Unterstützungsmenge, aus der wir Trainingsaufgaben auswählen, die das Use-Case-Szenario während der Vorhersage nachbilden. Mit jeder Trainingsaufgabe wird ein Bild abgefragt, das klassifiziert werden soll. Das Modell wird auf mehreren Trainingsaufgaben aus der Unterstützungsmenge trainiert, was man als *Episode* bezeichnet.

Beim Testen erhält das Modell dann eine neue Aufgabe mit anderen Klassen als beim Training. Die beim Training verwendeten Klassen nennt man auch Basisklassen. Die *Unterstützungsmenge* während des Trainings wird oftmals auch als *Basismenge* bezeichnet. Auch hier besteht die Aufgabe darin, abgefragte Bilder zu klassifizieren. Testaufgaben sind Trainingsaufgaben ähnlich, außer dass sich keine der Klassen beim Testen mit denen beim Training überschneiden. Abbildung 3-2 veranschaulicht dies.

Wie Abbildung 3-2 zeigt, enthalten die Unterstützungs- und Abfragemengen unterschiedliche Bilder derselben Klasse. Das Gleiche gilt für die Testphase. Beachten Sie aber, dass sich die Klassen in den Unterstützungs- und Abfragemengen von denjenigen Mengen unterscheiden, die beim Training erscheinen.

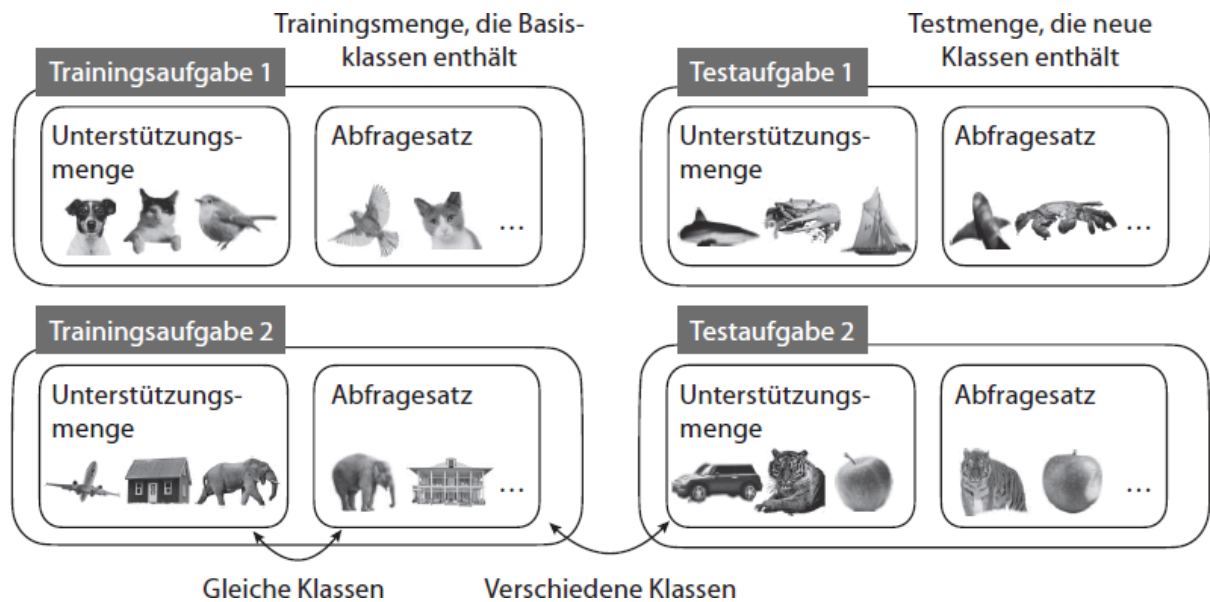


Abb. 3-2 Klassen beim Training und beim Testen

Few-Shot-Lernen gibt es in vielen verschiedenen Arten. Bei der gängigsten Variante, dem Meta-Lernen, geht es beim Training im Wesentlichen darum, die Parameter des Modells so zu aktualisieren, dass es sich gut an eine neue Aufgabe *anpassen* kann. Eine Strategie beim Few-Shot-Lernen besteht prinzipiell darin, ein Modell zu lernen, das Einbettungen erzeugt, bei denen wir die Zielklasse über eine Suche nach den nächsten Nachbarn unter den Bildern in der Unterstützungsmenge finden können. Abbildung 3-3 veranschaulicht diesen Ansatz.

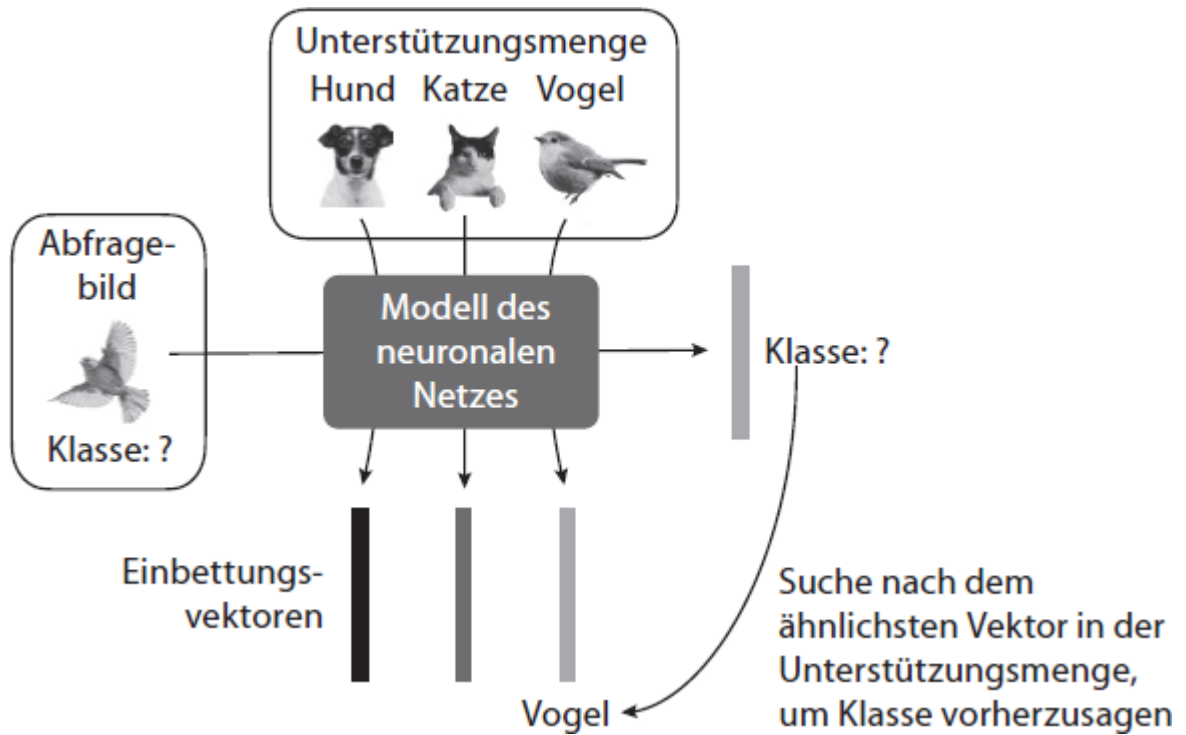


Abb. 3-3 Lernen von Einbettungen, die für die Klassifizierung geeignet sind

Das Modell lernt, wie man gute Einbettungen aus der Unterstützungsmenge erzeugt, um das Abfragebild zu klassifizieren, indem es den ähnlichsten Einbettungsvektor sucht.

3.2 Übungen

1. MNIST (<https://de.wikipedia.org/wiki/MNIST>) ist ein klassischer und beliebter Datensatz für maschinelles Lernen, der aus 50.000 handgeschriebenen Ziffern aus 10 Klassen besteht, die den Ziffern 0 bis 9 entsprechen. Wie kann man den MNIST-Datensatz für einen One-Shot-Klassifizierungskontext unterteilen?
2. Nennen Sie einige reale Anwendungen oder Use Cases für Few-Shot-Lernen.

Index

.632[+]-Bootstrapping-Regel 176

[CLS] 110

[SEP] 110

1-aus-n-Codierung 3, 215

A

- Abstand 5
 - euklidischer 187
 - L2 187
 - Wahrscheinlichkeitsverteilungen 188
- Adam-Optimierer 44, 219
- Adaptermethoden 126
- Add-&-Norm-Schritt 109
- Adversariale Validierung 198
- Ähnlichkeit 5
 - Kosinus- 226
- Aktives Lernen 203
- Algorithmen
 - Laufzeit- 63
 - nichtdeterministische 63
- Äquivalenz 77, 222
- Architekturen
 - EfficientNetV2 87
 - Transformer, ursprüngliche 107
- Aufgaben
 - Downstream 11, 47
 - Hilfs- 207
 - Multi-Task-Lernen 207–208
- Aufmerksamkeit
 - Bahdanau- 101
 - räumliche 224
 - Selbst- 101
- Autoencoder 5
 - maskierte 12, 223
 - variationale 53
- Autoregressive Decodierung 112
- Autoregressive Modelle 56

B

Bag-of-Words 215

Bahdanau-Aufmerksamkeit 101

BART, Encoder-Decoder-Modelle 112

Batched Inferenz 151

Bedeutungsgewichtung 158

Berry-Esseen 171

BERT

- Bidirectional Encoder Representations from Transformers 109

- F1-Maß 226

BERTScore 136

Bibliotheken

- Cleanlab 150

- CUDA Deep Neural Network (cuDNN) 64

- MAPIE 183

Bildhistogramme 215

Bilingual Evaluation Understudy (BLEU) 131

Binomialverteilung 170

BLAS (Basic Linear Algebra Subprograms) 152

BLEU 133

Bootstrapping 202

- Out-of-Bag- 171

- Trainingsdatensätze 171

- Vorhersagen 173

C

Carbin, Michael 21

ChatGPT 127

Checkpoints 37, 218

Cleanlab 150

CNNs, induktive Verzerrungen 81

Concept Drift 159

Continuous Bag-Of-Words (CBOW) 92

Convolutional Neural Network (CNN) 4

CUDA Deep Neural Network (cuDNN) 64

D

Data Augmentation 26

Daten

beschriftet 201

Bootstrapping 202

kategoriale 3

Lokalität 155

ordinale 165

synthetische 98

Datenparallelität 40

Datensätze 17

ImageNet 9, 27

Iris 17

Stimmungen 127

Unterstützungs- 17

Datenvermehrung 26

Ersetzen von Synonymen 95

Generieren synthetischer Daten 98

Löschen von Wörtern 96

Mischen von Sätzen 97

Permutation 96

Rauschinjektion 97

Rückübersetzung 98

Text 95

Vertauschen von Wortpositionen 96

Datenverteilungsverschiebungen 160

Datenzentrierte KI 148

Decoder 110

Decodierung, autoregressive 112

Deep-Boltzmann-Maschinen (DBMs) 52

DeepSpeed 219

Differenzialgleichung 59

Diffusionsmodelle 58

Domain Shift 159

Domänenverschiebung 159

Double Descent 34

Downstream 11, 47

Modell 117

Dreiecksungleichung 186

Dropout 32, 63

E

EfficientNetV2 87

Eigenschaften, emergente 112

Einbettungen 3

- Abstand 5

- Ähnlichkeit 5

- BERTScore 136

- latenter Raum 5

Einbettungsvektoren 3

Emergente Eigenschaften 112

Encoder 109

Encoder-Decoder-Modelle 112

- BART 112

- T5 112

Ensemble-Methoden 35, 219

Episoden 18

Ersetzen von Synonymen 95

Euklidischer Abstand 13, 185

Extrinsische Metriken 132

F

F1-Maß 135, 226

Faltung, Konvolution 64

Faltungen 191

Faltungsschichten 72

Fehler

 mittlerer absoluter 189

 Standardfehler der Regression 189

Feinabstimmung

 I, II 118

 parametereffiziente 122

 Präfix- 123

 Prompts 121

Felder 85

Few-Shot-Lernen 17, 204

Flussbasierte Modelle 55

Folds 191

Frankle, Jonathan 21

Fréchet-Abstand 220

Frühes Stoppen 32

Funktionen, stats.zscore 170

Fusion

 Operatoren 154

 Schleifen 154

G

Gaußsches Rauschen 58

Genauigkeit 135

Generalisierung, gestapelte 35

Generative Adversarial Networks (GANs) 54

Generieren synthetischer Daten 98

Gestapelte Generalisierung 35

Gewichte, initialisieren 61

Gewichtszurücknahme 32

Gewöhnliche Differenzialgleichung 59

Gibbs-Sampling 53

GPT

- emergente Eigenschaften 112

- Generative Pretrained Transformer 111

Gradientenabstieg, stochastischer 221

Grokking 34

Grundgesamtheit, statistische 167

H

- Halbüberwachtes Lernen 206
- Hard Parameter Sharing 208
- Hauptkomponentenanalyse 126
- Hilfsaufgaben 207
- Hinton, Geoffrey 6
- Holdout-Validierung 62
- Homofone 94, 223

I

- ImageNet 9, 27
- Inception Score 136
- In-Context Lernen 120
- Induktive Verzerrungen 81, 210
- Inferenz
 - Batched 151
 - Modell- 151
 - stapelorientierte 151
 - variationale 53
- Inpainting 202, 216
- Instruct-GPT 127
- Inter-Operatorparallelität 39
- Intervalle
 - Normal-Approximation 170
 - Vorhersage- 177
- Intra-Operatorparallelität 40
- Intrinsische Metriken 132
- Invarianz
 - räumliche 82
 - Translations- 82
- Iris 17
- Iteratives Magnitude-Pruning 22
- Iteratives Pruning 33

K

- k-fache Kreuzvalidierung 35
- KI, datenzentrierte 148
- Klassifizierungskopf 225
- Konfidenzintervalle 167
 - Binomialverteilung 170
 - Bootstrapping 171
 - konstruieren 167
 - symmetrische 169
 - Vorhersageintervalle 177
- Konfidenzniveaus 170
- Konforme Vorhersagen 178
- Konsistenzmodelle 59
- Kontrastives selbstüberwachtes Lernen 12
- Konvolution, Faltung 64
- Konvolutionale Schichten 72
- Kosinusähnlichkeit 226
- Kovariatenverschiebung 157
- Kreuzentropie 188
 - Verlustfunktionen 132
- Kreuzvalidierung
 - k-fache 35
 - k-Werte auswählen 192
 - Leave-One-Out- (LOOCV) 194
- Krizhevsky, Alex 6
- Kullback-Leibler-Divergenz 33

L

L2-Norm 13

Label-Shift 158

Latenter Raum 3, 5

Leave-One-Out-Kreuzvalidierung (LOOCV) 194

Lernen

aktives 203

Few-Shot- 17, 204

halbüberwachtes 206

In-Context 120

kontrastives 13

Little-Shot- 120

Meta- 204

multimodales 208

Multi-Task- 207

RLHF 127

schwach überwachtes 205

selbstüberwachtes 9, 12, 202

Transfer- 202

verstärkendes 127

Zero-Shot- 120

Little-Shot-Lernen 120

LlamaIndex 129

Lokalität 155

Loop tiling 153

Löschen von Wörtern 96

Lotterie-Ticket-Hypothese 21, 217

Low-Rank Adaptation (LoRA) 126

Low-Rank-Transformationen 126

M

MAE (Mean Absolute Error) 189

MAPIE 183

Maschinelles Lernen xvii

Maskierte Autoencoder 223

Maskierte Sprachmodellierung 92, 109

Maskierter Autoencoder 12

Mean Absolute Error (MAE) 189

Meta-Lernen 204

Metriken

- BERTScore 136

- Bilingual Evaluation Understudy (BLEU) 131

- BLEU 133

- euklidischer Abstand 185

- extrinsische 132

- intrinsische 132

- Perplexity 131–132

- präzisionsbasierte 133

- Recall-Oriented Understudy for Gisting Evaluation (ROUGE) 132

- referenzbasierte 133

- Surrogate 138

Mischen von Sätzen 97

Mittlerer absoluter Fehler 189

Mittlerer quadratischer Fehler 186

Modelle

- autoregressive 56

- Bag-of-Words 215

- BART 112

- Deep-Boltzmann-Maschinen (DBMs) 52

- Diffusions- 58

- diskriminative 51

- Encoder-Decoder- 112

- Ensemble-Methoden 35

- flussbasierte 55

- generative 51

- Gewichte initialisieren 61

- Konsistenz- 59

- maskierte Sprach- 92

- PixelCNN 57

- T5 112

- VideoBERT 209
- Modellinferenz 151
- Modellparallelität 39
- Multi-Task-Lernen 208
- N**
- Naive Bayes 51
- Neuronale Netze
 - Faltungsschichten 72
 - kontrastives selbstüberwachtes Lernen 12
 - konvolutionale Schichten 72
 - Schiebefenster 78
- Neuronen, tote 217
- n-Gramme 131
- Nichtkonformitätsmaß 180
- Non-Linear Independent Components Estimation (NICE) 55
- Normal-Approximation 170
- Normalisierende Flüsse 55
- Normalverteilung 165
 - Verschiebung 157
- Nucleus Sampling 65–66
- N-Way K-Shot 17

O

One-hot encoding 3

Operatorfusion 154

Optimierer, Adam- 44

Orakel 203

Ordinale Daten 165

Ordinary Differential Equation (ODE) 59

Out-of-Bag-Bootstrapping 171

Overfitting 25

P

Parallelisierung 48

Parallelität

- Daten- 40

- Inter-Operator- 39

- Intra-Operator- 40

- Modell- 39

- Pipeline- 42

- Sequenz- 42

- Tensor- 40

Parameter, Anzahl 47, 71

Parametereffiziente Feinabstimmungen 122

Patches 85

Permutation 96

Perplexity 131–132

Perzentile 173

Pipeline-Parallelität 42

PixelCNN 57

Poisson-Verteilung 165

Populationsparameter 167

Positionseinbettungen, relative 84

Positive-Unlabeled Lernen 206

PPO (Proximal Policy Optimization) 127

Präfix-Feinabstimmung 123, 226

Prior-Wahrscheinlichkeitsverschiebung 158

Prompting 117

Prompts, Feinabstimmung 121

Proximal Policy Optimization (PPO) 127

Pruning 156, 217

- iteratives 33

- iteratives Magnitude- 22

- strukturiertes 22

- unstrukturiertes 22

Pseudo-Labeler 207

PU-Lernen 206

Q

Quantisierung 155

R

Raum, latenter 5, 53

Rauschen, gaußsches 58

Rauschinjektion 97

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) 132

Regularisierung 32

- Dropout 32

- frühes Stoppen 32

- Gewichtszurücknahme 32

- Grokking 34

- Klassifizierer, lineare 117

- Multi-Task-Lernen 208

- Stärke 32

- Weight Decay 32

Reinforcement Learning with Human Feedback (RLHF) 127

Relative Positionseinbettungen 84

Reparametrisierung 155

Repräsentationen 6

- latente 3

Retraining 143

RLHF (Reinforcement Learning with Human Feedback) 127

RMSE (Root Mean Square Error) 189

Root Mean Square Error (RMSE) 189

ROUGE 135

- F1-Maß 135

ROUGE-1 135

ROUGE-L 135

ROUGE-N 135

ROUGE-S 135

Rückübersetzung 98

S

Satz von Berry-Esseen 171
Sätze mischen 97
Scaled Dot-Product Attention 42
Schichtnormalisierung 108
Schiebefenster 78
Schleifenfusion 154
Schleifenkachelung 153
Schwach überwachtetes Lernen 205
Score-Methode 180
Selbstaufmerksamkeit 101, 103
Selbstaufmerksamkeitsmechanismus 45
Selbstüberwachtes Lernen 9
 kontrastives 12
 Vorausgaben 10
 Vortraining 47
Selbstvorhersage 12
Self-Attention 101
Sentence Shuffling 97
seq2seq 112
Sequenzparallelität 42
SGD (Stochastic Gradient Descent) 75, 221
Siamesischer Netzaufbau 13
SIMD-Prinzip (Single Instruction, Multiple Data) 152
Singulärvektorzerlegung 126
Skip-Bigramme 135
Skip-Gram 92
Soft Parameter Sharing 208
Sprachgewandtheit 135
Sprachmodellierung, maskierte 92, 109
Stacking 36
Standardfehler 174
Standardfehler der Regression 189
Stapelorientierte Inferenz 151
Statistische Grundgesamtheit 167
stats.zscore 170
Stichproben, Gibbs-Sampling 53
Stimmungen 127
Stochastic Gradient Descent (SGD) 75, 221
Stochastischer Gradientenabstieg 75, 221

Surrogate 138
Sutskever, Ilya 6
Synonyme
 ersetzen 95
 Thesaurus 95

T

T5 112

Temperaturskalierung 65

Tensoren 153

Tensor-Parallelität 40

Thesaurus 95

Tokens

[CLS] 110

[SEP] 110

Top-k-Sampling 65, 221

Top-p-Sampling 66

Tote Neuronen 217

Training

Quantisierung 155

zustandsbehaftetes 144

Trajektorie 59

Transferlernen 9, 202

Transformer 45

Decoder 110

Encoder 109

Felder 85

Patches 85

Selbstaufmerksamkeitsmechanismus 45

ursprünglicher 107

Vision 81

Vortraining 47

Translationsäquivarianz 82

Translationsinvarianz 82

Trefferquote 135

t-Verteilung 174

U

Überanpassung 25

Unterstützungsdatensätze 17

Unüberwachtes Vortraining 202

V

- Validierung, adversariale 198
- Variationale Autoencoder (VAEs) 53, 220
- Variationale Inferenz 53
- Vektorisierung 152
 - SIMD-Prinzip (Single Instruction, Multiple Data) 152
- Verlustfunktionen
 - Autoencoder 54
 - Kreuzentropie 132
 - LLMs 131
 - Regularisierung 32
- Verschiebungen
 - Datenverteilungs- 160
 - Domänen- 159
 - Konzept- 159
 - Kovariaten- 157
 - Label- 158
 - Verteilungen 157
- Vertauschen von Wortpositionen 96
- Verteilungen
 - Binomial- 170
 - multivariate 51
 - Normal- 165
 - Normal-Approximation 170
 - Poisson- 165
 - t- 174
- Verteilungshypothese 91, 223
- Verzerrungen, induktive 81, 210
- VideoBERT 209
- Vision Transformer (ViT) 81
- ViT
 - Positionsinformationen 84
 - Vision Transformer 81
- Voraufgaben 10
- Vorhersagebereich 178
- Vorhersageintervall 178
- Vorhersageintervalle 177
- Vorhersagemenge 178
- Vorhersagen, konforme 178
- Vortraining 27

unüberwachtes 202

W

Weak Supervision 205

Weight Decay 32

Wissensdestillation 33, 156

Word2vec 92, 226

- Continuous Bag-Of-Words (CBOW) 92

- Skip-Gram 92

WordNet 95

Worteinbettungen 92

Wörter, löschen 96

Wortpositionen, vertauschen 96

Wortverschiebung 96

Z

Zero-Shot-Lernen 120

Zufälligkeit, Design 65

z-Werte 170