

CONTENTS IN DETAIL

FOREWORD	xix
-----------------	------------

ACKNOWLEDGMENTS	xxi
------------------------	------------

INTRODUCTION	xxiii
---------------------	--------------

Who Is This Book For?	xxiii
What Will You Get Out of This Book?	xxiv
How to Read This Book	xxiv
Online Resources	xxviii

PART I NEURAL NETWORKS AND DEEP LEARNING

1 EMBEDDINGS, LATENT SPACE, AND REPRESENTATIONS	3
--	----------

Embeddings	3
Latent Space	5
Representation	6
Exercises	6
References	7

2 SELF-SUPERVISED LEARNING	9
---------------------------------------	----------

Self-Supervised Learning vs. Transfer Learning	9
Leveraging Unlabeled Data	11
Self-Prediction and Contrastive Self-Supervised Learning	11
Exercises	14
References	14

3	FEW-SHOT LEARNING	15
	Datasets and Terminology	15
	Exercises	18
4	THE LOTTERY TICKET HYPOTHESIS	19
	The Lottery Ticket Training Procedure	19
	Practical Implications and Limitations	20
	Exercises	21
	References	21
5	REDUCING OVERFITTING WITH DATA	23
	Common Methods	24
	Collecting More Data	24
	Data Augmentation	24
	Pretraining	25
	Other Methods	26
	Exercises	26
	References	26
6	REDUCING OVERFITTING WITH MODEL MODIFICATIONS	29
	Common Methods	30
	Regularization	30
	Smaller Models	31
	Caveats with Smaller Models	32
	Ensemble Methods	33
	Other Methods	34
	Choosing a Regularization Technique	35
	Exercises	35
	References	35
7	MULTI-GPU TRAINING PARADIGMS	37
	The Training Paradigms	37
	Model Parallelism	38
	Data Parallelism	38
	Tensor Parallelism	38

Pipeline Parallelism	40
Sequence Parallelism	40
Recommendations	41
Exercises	42
References	42

8 THE SUCCESS OF TRANSFORMERS 43

The Attention Mechanism	43
Pretraining via Self-Supervised Learning	45
Large Numbers of Parameters	45
Easy Parallelization	45
Exercises	46
References	47

9 GENERATIVE AI MODELS 49

Generative vs. Discriminative Modeling	49
Types of Deep Generative Models	50
Energy-Based Models	50
Variational Autoencoders	51
Generative Adversarial Networks	52
Flow-Based Models	53
Autoregressive Models	54
Diffusion Models	55
Consistency Models	56
Recommendations	57
Exercises	57
References	58

10 SOURCES OF RANDOMNESS 59

Model Weight Initialization	59
Dataset Sampling and Shuffling	60
Nondeterministic Algorithms	61
Different Runtime Algorithms	61
Hardware and Drivers	62
Randomness and Generative AI	62
Exercises	64
References	65

PART II COMPUTER VISION

11 CALCULATING THE NUMBER OF PARAMETERS 69

How to Find Parameter Counts	69
Convolutional Layers	70
Fully Connected Layers	72
Practical Applications	73
Exercises	73

12 FULLY CONNECTED AND CONVOLUTIONAL LAYERS 75

When the Kernel and Input Sizes Are Equal	76
When the Kernel Size Is 1	77
Recommendations	78
Exercises	78

13 LARGE TRAINING SETS FOR VISION TRANSFORMERS 79

Inductive Biases in CNNs	80
ViTs Can Outperform CNNs	82
Inductive Biases in ViTs	83
Recommendations	84
Exercises	85
References	85

PART III NATURAL LANGUAGE PROCESSING

14 THE DISTRIBUTIONAL HYPOTHESIS 89

Word2vec, BERT, and GPT	90
Does the Hypothesis Hold?	92
Exercises	92
References	92

15	DATA AUGMENTATION FOR TEXT	93
	Synonym Replacement	93
	Word Deletion	94
	Word Position Swapping	94
	Sentence Shuffling	95
	Noise Injection	95
	Back Translation	96
	Synthetic Data	96
	Recommendations	97
	Exercises	97
	References	97
16	SELF-ATTENTION	99
	Attention in RNNs	99
	The Self-Attention Mechanism	101
	Exercises	103
	References	103
17	ENCODER- AND DECODER-STYLE TRANSFORMERS	105
	The Original Transformer	105
	Encoders	107
	Decoders	108
	Encoder-Decoder Hybrids	110
	Terminology	110
	Contemporary Transformer Models	111
	Exercises	112
	References	112
18	USING AND FINE-TUNING PRETRAINED TRANSFORMERS	113
	Using Transformers for Classification Tasks	113
	In-Context Learning, Indexing, and Prompt Tuning	116
	Parameter-Efficient Fine-Tuning	119
	Reinforcement Learning with Human Feedback	124
	Adapting Pretrained Language Models	124
	Exercises	125
	References	125

19	EVALUATING GENERATIVE LARGE LANGUAGE MODELS	127
Evaluation Metrics for LLMs		127
Perplexity		128
BLEU Score		129
ROUGE Score		131
BERTScore		132
Surrogate Metrics		133
Exercises		134
References		134

PART IV PRODUCTION AND DEPLOYMENT

20	STATELESS AND STATEFUL TRAINING	139
Stateless (Re)training		139
Stateful Training		140
Exercises		141

21	DATA-CENTRIC AI	143
Data-Centric vs. Model-Centric AI		143
Recommendations		145
Exercises		146
References		146

22	SPEEDING UP INFERENCE	147
Parallelization		147
Vectorization		148
Loop Tiling		149
Operator Fusion		150
Quantization		151
Exercises		152
References		152

23	DATA DISTRIBUTION SHIFTS	153
Covariate Shift		153
Label Shift		154
Concept Drift		155

Domain Shift	155
Types of Data Distribution Shifts	156
Exercises	157
References	157

PART V PREDICTIVE PERFORMANCE AND MODEL EVALUATION

24 POISSON AND ORDINAL REGRESSION 161

Exercises	162
-----------------	-----

25 CONFIDENCE INTERVALS 163

Defining Confidence Intervals	164
The Methods	166
Method 1: Normal Approximation Intervals	166
Method 2: Bootstrapping Training Sets	167
Method 3: Bootstrapping Test Set Predictions	169
Method 4: Retraining Models with Different Random Seeds	169
Recommendations	170
Exercises	171
References	171

26 CONFIDENCE INTERVALS VS. CONFORMAL PREDICTIONS 173

Confidence Intervals and Prediction Intervals	174
Prediction Intervals and Conformal Predictions	174
Prediction Regions, Intervals, and Sets	174
Computing Conformal Predictions	175
A Conformal Prediction Example	176
The Benefits of Conformal Predictions	177
Recommendations	178
Exercises	178
References	178

27 PROPER METRICS 179

The Criteria	179
The Mean Squared Error	180
The Cross-Entropy Loss	182
Exercises	183

28		
THE K IN K-FOLD CROSS-VALIDATION		185
Trade-offs in Selecting Values for k	186	
Determining Appropriate Values for k	187	
Exercises	188	
References	188	
29		
TRAINING AND TEST SET DISCORDANCE		189
Exercises	191	
30		
LIMITED LABELED DATA		193
Improving Model Performance with Limited Labeled Data	193	
Labeling More Data	193	
Bootstrapping the Data	194	
Transfer Learning	194	
Self-Supervised Learning	194	
Active Learning	195	
Few-Shot Learning	195	
Meta-Learning	196	
Weakly Supervised Learning	197	
Semi-Supervised Learning	198	
Self-Training	199	
Multi-Task Learning	199	
Multimodal Learning	200	
Inductive Biases	202	
Recommendations	202	
Exercises	204	
References	204	
AFTERWORD		205
APPENDIX: ANSWERS TO THE EXERCISES		207
INDEX		223