Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

Contents

Preface *page* xvii

Installation xxvi

Notation xxix

1 Introduction 1
1.1 A Motivating Example 2
1.2 Key Components 4
1.3 Kinds of Machine Learning Problems 7
1.4 Roots 20
1.5 The Road to Deep Learning 22
1.6 Success Stories 25
1.7 The Essence of Deep Learning 27
1.8 Summary 29
1.9 Exercises 29

2 Preliminaries 30
2.1 Data Manipulation 30
2.1.1 Getting Started 30
2.1.2 Indexing and Slicing 33
2.1.3 Operations 34
2.1.4 Broadcasting 35
2.1.5 Saving Memory 36
2.1.6 Conversion to Other Python Objects 37
2.1.7 Summary 37
2.1.8 Exercises 38
2.2 Data Preprocessing 38
2.2.1 Reading the Dataset 38
2.2.2 Data Preparation 39
2.2.3 Conversion to the Tensor Format 40
2.2.4 Discussion 40
2.2.5 Exercises 40
2.3 Linear Algebra 41
2.3.1 Scalars 41

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

2.3.2 Vectors 42
2.3.3 Matrices 43
2.3.4 Tensors 44
2.3.5 Basic Properties of Tensor Arithmetic 45
2.3.6 Reduction 46
2.3.7 Non-Reduction Sum 47
2.3.8 Dot Products 48
2.3.9 Matrix–Vector Products 48
2.3.10 Matrix–Matrix Multiplication 49
2.3.11 Norms 50
2.3.12 Discussion 52
2.3.13 Exercises 53
2.4 Calculus 54
2.4.1 Derivatives and Differentiation 54
2.4.2 Visualization Utilities 56
2.4.3 Partial Derivatives and Gradients 58
2.4.4 Chain Rule 59
2.4.5 Discussion 59
2.4.6 Exercises 59
2.5 Automatic Differentiation 60
2.5.1 A Simple Function 60
2.5.2 Backward for Non-Scalar Variables 62
2.5.3 Detaching Computation 62
2.5.4 Gradients and Python Control Flow 63
2.5.5 Discussion 64
2.5.6 Exercises 64
2.6 Probability and Statistics 65
2.6.1 A Simple Example: Tossing Coins 66
2.6.2 A More Formal Treatment 69
2.6.3 Random Variables 69
2.6.4 Multiple Random Variables 70
2.6.5 An Example 73
2.6.6 Expectations 74
2.6.7 Discussion 76
2.6.8 Exercises 77
2.7 Documentation 78
2.7.1 Functions and Classes in a Module 79
2.7.2 Specific Functions and Classes 79

3 Linear Neural Networks for Regression 82
3.1 Linear Regression 82
3.1.1 Basics 83
3.1.2 Vectorization for Speed 88
3.1.3 The Normal Distribution and Squared Loss 88
3.1.4 Linear Regression as a Neural Network 90

© in this web service Cambridge University Press & Assessment www.cambridge.org

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

3.1.5 Summary 92
3.1.6 Exercises 92
3.2 Object-Oriented Design for Implementation 93
3.2.1 Utilities 94
3.2.2 Models 96
3.2.3 Data 97
3.2.4 Training 98
3.2.5 Summary 99
3.2.6 Exercises 99
3.3 Synthetic Regression Data 99
3.3.1 Generating the Dataset 99
3.3.2 Reading the Dataset 100
3.3.3 Concise Implementation of the Data Loader 101
3.3.4 Summary 102
3.3.5 Exercises 102
3.4 Linear Regression Implementation from Scratch 103
3.4.1 Defining the Model 103
3.4.2 Defining the Loss Function 104
3.4.3 Defining the Optimization Algorithm 104
3.4.4 Training 105
3.4.5 Summary 107
3.4.6 Exercises 107
3.5 Concise Implementation of Linear Regression 108
3.5.1 Defining the Model 109
3.5.2 Defining the Loss Function 110
3.5.3 Defining the Optimization Algorithm 110
3.5.4 Training 110
3.5.5 Summary 111
3.5.6 Exercises 112
3.6 Generalization 112
3.6.1 Training Error and Generalization Error 113
3.6.2 Underfitting or Overfitting? 115
3.6.3 Model Selection 117
3.6.4 Summary 118
3.6.5 Exercises 118
3.7 Weight Decay 119
3.7.1 Norms and Weight Decay 119
3.7.2 High-Dimensional Linear Regression 121
3.7.3 Implementation from Scratch 121
3.7.4 Concise Implementation 123
3.7.5 Summary 125
3.7.6 Exercises 125

4 Linear Neural Networks for Classification 126
4.1 Softmax Regression 126

© in this web service Cambridge University Press & Assessment www.cambridge.org

4.1.1 Classification 127
4.1.2 Loss Function 130
4.1.3 Information Theory Basics 131
4.1.4 Summary and Discussion 132
4.1.5 Exercises 133
4.2 The Image Classification Dataset 135
4.2.1 Loading the Dataset 135
4.2.2 Reading a Minibatch 136
4.2.3 Visualization 137
4.2.4 Summary 138
4.2.5 Exercises 138
4.3 The Base Classification Model 139
4.3.1 The `Classifier` Class 139
4.3.2 Accuracy 139
4.3.3 Summary 140
4.3.4 Exercises 140
4.4 Softmax Regression Implementation from Scratch 141
4.4.1 The Softmax 141
4.4.2 The Model 142
4.4.3 The Cross-Entropy Loss 142
4.4.4 Training 143
4.4.5 Prediction 144
4.4.6 Summary 145
4.4.7 Exercises 145
4.5 Concise Implementation of Softmax Regression 145
4.5.1 Defining the Model 146
4.5.2 Softmax Revisited 146
4.5.3 Training 147
4.5.4 Summary 148
4.5.5 Exercises 148
4.6 Generalization in Classification 148
4.6.1 The Test Set 149
4.6.2 Test Set Reuse 151
4.6.3 Statistical Learning Theory 152
4.6.4 Summary 154
4.6.5 Exercises 155
4.7 Environment and Distribution Shift 155
4.7.1 Types of Distribution Shift 156
4.7.2 Examples of Distribution Shift 158
4.7.3 Correction of Distribution Shift 160
4.7.4 A Taxonomy of Learning Problems 164
4.7.5 Fairness, Accountability, and Transparency in Machine Learning 166
4.7.6 Summary 166
4.7.7 Exercises 167

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

5 Multilayer Perceptrons 168
5.1 Multilayer Perceptrons 168
5.1.1 Hidden Layers 168
5.1.2 Activation Functions 172
5.1.3 Summary and Discussion 176
5.1.4 Exercises 176
5.2 Implementation of Multilayer Perceptrons 177
5.2.1 Implementation from Scratch 177
5.2.2 Concise Implementation 178
5.2.3 Summary 179
5.2.4 Exercises 180
5.3 Forward Propagation, Backward Propagation, and Computational Graphs 181
5.3.1 Forward Propagation 181
5.3.2 Computational Graph of Forward Propagation 182
5.3.3 Backpropagation 183
5.3.4 Training Neural Networks 184
5.3.5 Summary 184
5.3.6 Exercises 185
5.4 Numerical Stability and Initialization 185
5.4.1 Vanishing and Exploding Gradients 186
5.4.2 Parameter Initialization 188
5.4.3 Summary 190
5.4.4 Exercises 190
5.5 Generalization in Deep Learning 190
5.5.1 Revisiting Overfitting and Regularization 191
5.5.2 Inspiration from Nonparametrics 192
5.5.3 Early Stopping 193
5.5.4 Classical Regularization Methods for Deep Networks 194
5.5.5 Summary 194
5.5.6 Exercises 195
5.6 Dropout 195
5.6.1 Dropout in Practice 196
5.6.2 Implementation from Scratch 197
5.6.3 Concise Implementation 198
5.6.4 Summary 200
5.6.5 Exercises 200
5.7 Predicting House Prices on Kaggle 200
5.7.1 Downloading Data 201
5.7.2 Kaggle 201
5.7.3 Accessing and Reading the Dataset 202
5.7.4 Data Preprocessing 203
5.7.5 Error Measure 204
5.7.6 *K*-Fold Cross-Validation 205
5.7.7 Model Selection 206
5.7.8 Submitting Predictions on Kaggle 207

© in this web service Cambridge University Press & Assessment www.cambridge.org

5.7.9 Summary and Discussion 207
5.7.10 Exercises 208

6 Builders' Guide 209
6.1 Layers and Modules 209
6.1.1 A Custom Module 211
6.1.2 The Sequential Module 213
6.1.3 Executing Code in the Forward Propagation Method 213
6.1.4 Summary 215
6.1.5 Exercises 215
6.2 Parameter Management 215
6.2.1 Parameter Access 216
6.2.2 Tied Parameters 217
6.2.3 Summary 218
6.2.4 Exercises 218
6.3 Parameter Initialization 218
6.3.1 Built-in Initialization 219
6.3.2 Summary 221
6.3.3 Exercises 221
6.4 Lazy Initialization 221
6.4.1 Summary 222
6.4.2 Exercises 223
6.5 Custom Layers 223
6.5.1 Layers without Parameters 223
6.5.2 Layers with Parameters 224
6.5.3 Summary 225
6.5.4 Exercises 225
6.6 File I/O 225
6.6.1 Loading and Saving Tensors 226
6.6.2 Loading and Saving Model Parameters 227
6.6.3 Summary 228
6.6.4 Exercises 228
6.7 GPUs 228
6.7.1 Computing Devices 229
6.7.2 Tensors and GPUs 230
6.7.3 Neural Networks and GPUs 232
6.7.4 Summary 233
6.7.5 Exercises 234

7 Convolutional Neural Networks 235
7.1 From Fully Connected Layers to Convolutions 236
7.1.1 Invariance 236
7.1.2 Constraining the MLP 237
7.1.3 Convolutions 239
7.1.4 Channels 240

7.1.5 Summary and Discussion 241
7.1.6 Exercises 241
7.2 Convolutions for Images 242
7.2.1 The Cross-Correlation Operation 242
7.2.2 Convolutional Layers 244
7.2.3 Object Edge Detection in Images 244
7.2.4 Learning a Kernel 246
7.2.5 Cross-Correlation and Convolution 247
7.2.6 Feature Map and Receptive Field 247
7.2.7 Summary 248
7.2.8 Exercises 249
7.3 Padding and Stride 249
7.3.1 Padding 250
7.3.2 Stride 252
7.3.3 Summary and Discussion 253
7.3.4 Exercises 254
7.4 Multiple Input and Multiple Output Channels 254
7.4.1 Multiple Input Channels 254
7.4.2 Multiple Output Channels 256
7.4.3 1 × 1 Convolutional Layer 257
7.4.4 Discussion 258
7.4.5 Exercises 258
7.5 Pooling 259
7.5.1 Maximum Pooling and Average Pooling 260
7.5.2 Padding and Stride 262
7.5.3 Multiple Channels 263
7.5.4 Summary 264
7.5.5 Exercises 264
7.6 Convolutional Neural Networks (LeNet) 264
7.6.1 LeNet 265
7.6.2 Training 268
7.6.3 Summary 268
7.6.4 Exercises 269

8 Modern Convolutional Neural Networks 270
8.1 Deep Convolutional Neural Networks (AlexNet) 271
8.1.1 Representation Learning 272
8.1.2 AlexNet 275
8.1.3 Training 278
8.1.4 Discussion 279
8.1.5 Exercises 280
8.2 Networks Using Blocks (VGG) 280
8.2.1 VGG Blocks 281
8.2.2 VGG Network 282
8.2.3 Training 283

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

8.2.4 Summary 284
8.2.5 Exercises 284
8.3 Network in Network (NiN) 285
8.3.1 NiN Blocks 286
8.3.2 NiN Model 286
8.3.3 Training 288
8.3.4 Summary 288
8.3.5 Exercises 289
8.4 Multi-Branch Networks (GoogLeNet) 289
8.4.1 Inception Blocks 290
8.4.2 GoogLeNet Model 291
8.4.3 Training 293
8.4.4 Discussion 294
8.4.5 Exercises 294
8.5 Batch Normalization 295
8.5.1 Training Deep Networks 295
8.5.2 Batch Normalization Layers 298
8.5.3 Implementation from Scratch 299
8.5.4 LeNet with Batch Normalization 301
8.5.5 Concise Implementation 302
8.5.6 Discussion 303
8.5.7 Exercises 304
8.6 Residual Networks (ResNet) and ResNeXt 305
8.6.1 Function Classes 305
8.6.2 Residual Blocks 307
8.6.3 ResNet Model 309
8.6.4 Training 311
8.6.5 ResNeXt 311
8.6.6 Summary and Discussion 314
8.6.7 Exercises 315
8.7 Densely Connected Networks (DenseNet) 315
8.7.1 From ResNet to DenseNet 315
8.7.2 Dense Blocks 316
8.7.3 Transition Layers 317
8.7.4 DenseNet Model 318
8.7.5 Training 319
8.7.6 Summary and Discussion 319
8.7.7 Exercises 320
8.8 Designing Convolution Network Architectures 320
8.8.1 The AnyNet Design Space 321
8.8.2 Distributions and Parameters of Design Spaces 323
8.8.3 RegNet 325
8.8.4 Training 326
8.8.5 Discussion 326
8.8.6 Exercises 327

© in this web service Cambridge University Press & Assessment

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

9 Recurrent Neural Networks 328
9.1 Working with Sequences 330
9.1.1 Autoregressive Models 331
9.1.2 Sequence Models 333
9.1.3 Training 335
9.1.4 Prediction 336
9.1.5 Summary 338
9.1.6 Exercises 339
9.2 Converting Raw Text into Sequence Data 339
9.2.1 Reading the Dataset 340
9.2.2 Tokenization 340
9.2.3 Vocabulary 341
9.2.4 Putting It All Together 342
9.2.5 Exploratory Language Statistics 342
9.2.6 Summary 345
9.2.7 Exercises 345
9.3 Language Models 346
9.3.1 Learning Language Models 346
9.3.2 Perplexity 348
9.3.3 Partitioning Sequences 349
9.3.4 Summary and Discussion 351
9.3.5 Exercises 351
9.4 Recurrent Neural Networks 352
9.4.1 Neural Networks without Hidden States 352
9.4.2 Recurrent Neural Networks with Hidden States 353
9.4.3 RNN-Based Character-Level Language Models 355
9.4.4 Summary 356
9.4.5 Exercises 356
9.5 Recurrent Neural Network Implementation from Scratch 356
9.5.1 RNN Model 356
9.5.2 RNN-Based Language Model 358
9.5.3 Gradient Clipping 360
9.5.4 Training 361
9.5.5 Decoding 361
9.5.6 Summary 363
9.5.7 Exercises 363
9.6 Concise Implementation of Recurrent Neural Networks 364
9.6.1 Defining the Model 364
9.6.2 Training and Predicting 365
9.6.3 Summary 365
9.6.4 Exercises 366
9.7 Backpropagation Through Time 366
9.7.1 Analysis of Gradients in RNNs 366
9.7.2 Backpropagation Through Time in Detail 369
9.7.3 Summary 372

© in this web service Cambridge University Press & Assessment

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

9.7.4 Exercises 372

10 Modern Recurrent Neural Networks 373
10.1 Long Short-Term Memory (LSTM) 374
10.1.1 Gated Memory Cell 374
10.1.2 Implementation from Scratch 378
10.1.3 Concise Implementation 379
10.1.4 Summary 380
10.1.5 Exercises 380
10.2 Gated Recurrent Units (GRU) 381
10.2.1 Reset Gate and Update Gate 381
10.2.2 Candidate Hidden State 382
10.2.3 Hidden State 382
10.2.4 Implementation from Scratch 383
10.2.5 Concise Implementation 384
10.2.6 Summary 386
10.2.7 Exercises 386
10.3 Deep Recurrent Neural Networks 386
10.3.1 Implementation from Scratch 388
10.3.2 Concise Implementation 388
10.3.3 Summary 389
10.3.4 Exercises 390
10.4 Bidirectional Recurrent Neural Networks 390
10.4.1 Implementation from Scratch 391
10.4.2 Concise Implementation 392
10.4.3 Summary 392
10.4.4 Exercises 392
10.5 Machine Translation and the Dataset 393
10.5.1 Downloading and Preprocessing the Dataset 393
10.5.2 Tokenization 395
10.5.3 Loading Sequences of Fixed Length 396
10.5.4 Reading the Dataset 397
10.5.5 Summary 398
10.5.6 Exercises 398
10.6 The Encoder−Decoder Architecture 399
10.6.1 Encoder 399
10.6.2 Decoder 400
10.6.3 Putting the Encoder and Decoder Together 400
10.6.4 Summary 400
10.6.5 Exercises 401
10.7 Sequence-to-Sequence Learning for Machine Translation 401
10.7.1 Teacher Forcing 402
10.7.2 Encoder 402
10.7.3 Decoder 404
10.7.4 Encoder–Decoder for Sequence-to-Sequence Learning 406

© in this web service Cambridge University Press & Assessment www.cambridge.org

10.7.5 Loss Function with Masking 406
10.7.6 Training 406
10.7.7 Prediction 407
10.7.8 Evaluation of Predicted Sequences 408
10.7.9 Summary 409
10.7.10 Exercises 409
10.8 Beam Search 410
10.8.1 Greedy Search 410
10.8.2 Exhaustive Search 411
10.8.3 Beam Search 412
10.8.4 Summary 413
10.8.5 Exercises 413

11 Attention Mechanisms and Transformers 414
11.1 Queries, Keys, and Values 416
11.1.1 Visualization 417
11.1.2 Summary 419
11.1.3 Exercises 419
11.2 Attention Pooling by Similarity 420
11.2.1 Kernels and Data 420
11.2.2 Attention Pooling via Nadaraya–Watson Regression 422
11.2.3 Adapting Attention Pooling 423
11.2.4 Summary 424
11.2.5 Exercises 424
11.3 Attention Scoring Functions 425
11.3.1 Dot Product Attention 426
11.3.2 Convenience Functions 426
11.3.3 Scaled Dot Product Attention 428
11.3.4 Additive Attention 429
11.3.5 Summary 431
11.3.6 Exercises 431
11.4 The Bahdanau Attention Mechanism 432
11.4.1 Model 433
11.4.2 Defining the Decoder with Attention 433
11.4.3 Training 435
11.4.4 Summary 436
11.4.5 Exercises 437
11.5 Multi-Head Attention 437
11.5.1 Model 438
11.5.2 Implementation 438
11.5.3 Summary 440
11.5.4 Exercises 440
11.6 Self-Attention and Positional Encoding 440
11.6.1 Self-Attention 441
11.6.2 Comparing CNNs, RNNs, and Self-Attention 441

© in this web service Cambridge University Press & Assessment

Cambridge University Press & Assessment
978-1-009-38943-3 — Dive into Deep Learning
Aston Zhang , Zachary C. Lipton , Mu Li , Alexander J. Smola
Table of Contents
More Information

11.6.3 Positional Encoding 442
11.6.4 Summary 445
11.6.5 Exercises 445
11.7 The Transformer Architecture 446
11.7.1 Model 446
11.7.2 Positionwise Feed-Forward Networks 447
11.7.3 Residual Connection and Layer Normalization 448
11.7.4 Encoder 449
11.7.5 Decoder 451
11.7.6 Training 453
11.7.7 Summary 456
11.7.8 Exercises 456
11.8 Transformers for Vision 457
11.8.1 Model 457
11.8.2 Patch Embedding 459
11.8.3 Vision Transformer Encoder 459
11.8.4 Putting It All Together 460
11.8.5 Training 461
11.8.6 Summary and Discussion 462
11.8.7 Exercises 462
11.9 Large-Scale Pretraining with Transformers 462
11.9.1 Encoder-Only 463
11.9.2 Encoder–Decoder 465
11.9.3 Decoder-Only 467
11.9.4 Scalability 470
11.9.5 Large Language Models 471
11.9.6 Summary and Discussion 472
11.9.7 Exercises 473

Appendix: Tools for Deep Learning 474
A.1 Using Jupyter Notebooks 474
A.2 Using Amazon SageMaker 479
A.3 Using AWS EC2 Instances 482
A.4 Using Google Colab 490
A.5 Selecting Servers and GPUs 491
A.6 Contributing to This Book 495
A.7 The d2l API Document 500

References 516

Index 539

© in this web service Cambridge University Press & Assessment