The Science of Deep Learning

The Science of Deep Learning emerged from courses taught by the author that have provided thousands of students with training and experience for their academic studies, and prepared them for careers in deep learning, machine learning, and artificial intelligence in top companies in industry and academia.

The book begins by covering the foundations of deep learning, followed by key deep learning architectures. Subsequent parts on generative models and reinforcement learning may be used as part of a deep learning course or as part of a course on each topic. The book includes state-of-the-art topics such as Transformers, graph neural networks, variational autoencoders, and deep reinforcement learning, with a broad range of applications. The appendices provide equations for computing gradients in backpropagation and optimization, and best practices in scientific writing and reviewing.

The text presents an up-to-date guide to the field built upon clear visualizations using a unified notation and equations, lowering the barrier to entry for the reader. The accompanying website provides complementary code and hundreds of exercises with solutions.

Iddo Drori is a faculty member and associate professor at Boston University, a lecturer at MIT, and adjunct associate professor at Columbia University. He was a visiting associate professor at Cornell University in operations research and information engineering, and research scientist and adjunct professor at NYU Center for Data Science, Courant Institute, and NYU Tandon. He holds a PhD in computer science and was a postdoctoral research fellow at Stanford University in statistics. He also holds an MBA in organizational behavior and entrepreneurship and has a decade of industry research and leadership experience. His main research is in machine learning, AI, and computer vision, with 70 publications and over 5,100 citations, and he has taught over 35 courses in computer science. He has won multiple competitions in computer vision conferences and received multiple best paper awards in machine learning conferences.
The Science of Deep Learning

IDDO DRORI
Massachusetts Institute of Technology
Columbia University
# Contents

Preface .......................... xv
Acknowledgments .............. xvii
Abbreviations and Notation .... xix

## Part I Foundations

### 1 Introduction

1.1 Deep Learning .............. 3
1.2 Outline ..................... 4
1.2.1 Part I: Foundations: Backpropagation, Optimization, and Regularization .......................... 4
1.2.2 Part II: Architectures: CNNs, RNNs, GNNs, and Transformers ................... 5
1.2.3 Part III: Generative Models: GANs, VAEs, and Normalizing Flows .................. 6
1.2.4 Part IV: Reinforcement Learning .......................................................... 6
1.2.5 Part V: Applications .......................................................... 7
1.2.6 Appendices ..................... 7
1.3 Code ................................ 7
1.4 Exercises ...................... 8

### 2 Forward and Backpropagation

2.1 Introduction ................. 9
2.2 Fully Connected Neural Network .... 9
2.3 Forward Propagation ......... 11
2.3.1 Algorithm .................... 12
2.3.2 Example ..................... 12
2.3.3 Logistic Regression ............ 14
2.4 Non-linear Activation Functions .... 16
2.4.1 Sigmoid ..................... 16
2.4.2 Hyperbolic Tangent ............ 16
2.4.3 Rectified Linear Unit ........... 17
2.4.4 Swish ..................... 18
# Contents

2.4.5 Softmax .................................. 18
2.5 Loss Functions ............................... 19
2.6 Backpropagation .............................. 21
2.7 Differentiable Programming ................. 22
2.8 Computation Graph ........................... 22
  2.8.1 Example ................................ 22
  2.8.2 Logistic Regression ...................... 24
  2.8.3 Forward and Backpropagation .......... 25
2.9 Derivative of Non-linear Activation Functions .... 26
2.10 Backpropagation Algorithm ................. 28
  2.10.1 Example ................................ 29
2.11 Chain Rule for Differentiation ............... 30
  2.11.1 Two Functions in One Dimension ........ 30
  2.11.2 Three Functions in One Dimension ...... 31
  2.11.3 Two Functions in Higher Dimensions ... 31
2.12 Gradient of Loss Function .................. 32
2.13 Gradient Descent ............................ 32
2.14 Initialization and Normalization ............. 33
2.15 Software Libraries and Platforms ............ 33
2.16 Summary ................................. 33

3 Optimization ................................. 35
  3.1 Introduction ............................... 35
  3.2 Overview ................................ 35
    3.2.1 Optimization Problem Classes .......... 35
    3.2.2 Optimization Solution Methods .......... 37
    3.2.3 Derivatives and Gradients .............. 37
    3.2.4 Gradient Computation ................. 38
  3.3 First-Order Methods .................... 39
    3.3.1 Gradient Descent ..................... 39
    3.3.2 Step Size ................................ 41
    3.3.3 Mini-Batch Gradient Descent ............ 44
    3.3.4 Stochastic Gradient Descent ............ 44
    3.3.5 Adaptive Gradient Descent .............. 45
    3.3.6 Momentum .............................. 46
    3.3.7 Adagrad ................................ 47
    3.3.8 Adam: Adaptive Moment Estimation ...... 47
    3.3.9 Hypergradient Descent .................. 48
  3.4 Second-Order Methods ................... 49
    3.4.1 Newton’s Method ....................... 49
    3.4.2 Second-Order Taylor Approximation ..... 51
    3.4.3 Quasi-Newton Methods ................. 53
  3.5 Evolution Strategies ...................... 54
  3.6 Summary ................................ 55
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4</strong></td>
<td>Regularization</td>
<td>56</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>Generalization</td>
<td>56</td>
</tr>
<tr>
<td>4.3</td>
<td>Overfitting</td>
<td>58</td>
</tr>
<tr>
<td>4.4</td>
<td>Cross Validation</td>
<td>59</td>
</tr>
<tr>
<td>4.5</td>
<td>Bias and Variance</td>
<td>59</td>
</tr>
<tr>
<td>4.6</td>
<td>Vector Norms</td>
<td>59</td>
</tr>
<tr>
<td>4.7</td>
<td>Ridge Regression and Lasso</td>
<td>60</td>
</tr>
<tr>
<td>4.8</td>
<td>Regularized Loss Functions</td>
<td>61</td>
</tr>
<tr>
<td>4.9</td>
<td>Dropout Regularization</td>
<td>62</td>
</tr>
<tr>
<td>4.9.1</td>
<td>Random Least Squares with Dropout</td>
<td>63</td>
</tr>
<tr>
<td>4.9.2</td>
<td>Least Squares with Noise Input Distortion</td>
<td>64</td>
</tr>
<tr>
<td>4.10</td>
<td>Data Augmentation</td>
<td>64</td>
</tr>
<tr>
<td>4.11</td>
<td>Batch Normalization</td>
<td>65</td>
</tr>
<tr>
<td>4.12</td>
<td>Summary</td>
<td>65</td>
</tr>
</tbody>
</table>

**Part II  Architectures**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5</strong></td>
<td>Convolutional Neural Networks</td>
<td>69</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>69</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Representations Sharing Weights</td>
<td>69</td>
</tr>
<tr>
<td>5.2</td>
<td>Convolution</td>
<td>70</td>
</tr>
<tr>
<td>5.2.1</td>
<td>One-Dimensional Convolution</td>
<td>70</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Matrix Multiplication</td>
<td>71</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Two-Dimensional Convolution</td>
<td>73</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Separable Filters</td>
<td>76</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Properties</td>
<td>76</td>
</tr>
<tr>
<td>5.2.6</td>
<td>Composition</td>
<td>77</td>
</tr>
<tr>
<td>5.2.7</td>
<td>Three-Dimensional Convolution</td>
<td>77</td>
</tr>
<tr>
<td>5.3</td>
<td>Layers</td>
<td>78</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Convolution</td>
<td>78</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Pooling</td>
<td>78</td>
</tr>
<tr>
<td>5.4</td>
<td>Example</td>
<td>79</td>
</tr>
<tr>
<td>5.5</td>
<td>Architectures</td>
<td>80</td>
</tr>
<tr>
<td>5.6</td>
<td>Applications</td>
<td>86</td>
</tr>
<tr>
<td>5.7</td>
<td>Summary</td>
<td>89</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>6</strong></td>
<td>Sequence Models</td>
<td>91</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>91</td>
</tr>
<tr>
<td>6.2</td>
<td>Natural Language Models</td>
<td>91</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Bag of Words</td>
<td>91</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Feature Vector</td>
<td>92</td>
</tr>
<tr>
<td>6.2.3</td>
<td>N-grams</td>
<td>92</td>
</tr>
</tbody>
</table>
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.4</td>
<td>Markov Model</td>
<td>92</td>
</tr>
<tr>
<td>6.2.5</td>
<td>State Machine</td>
<td>92</td>
</tr>
<tr>
<td>6.2.6</td>
<td>Recurrent Neural Network</td>
<td>93</td>
</tr>
<tr>
<td>6.3</td>
<td>Recurrent Neural Network</td>
<td>93</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Architectures</td>
<td>94</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Loss Function</td>
<td>95</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Deep RNN</td>
<td>97</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Bidirectional RNN</td>
<td>98</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Backpropagation Through Time</td>
<td>99</td>
</tr>
<tr>
<td>6.4</td>
<td>Gated Recurrent Unit</td>
<td>102</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Update Gate</td>
<td>104</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Candidate Activation</td>
<td>105</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Reset Gate</td>
<td>105</td>
</tr>
<tr>
<td>6.4.4</td>
<td>Function</td>
<td>107</td>
</tr>
<tr>
<td>6.5</td>
<td>Long Short-Term Memory</td>
<td>108</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Forget Gate</td>
<td>109</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Input Gate</td>
<td>110</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Memory Cell</td>
<td>112</td>
</tr>
<tr>
<td>6.5.4</td>
<td>Candidate Memory</td>
<td>113</td>
</tr>
<tr>
<td>6.5.5</td>
<td>Output Gate</td>
<td>113</td>
</tr>
<tr>
<td>6.5.6</td>
<td>Peephole Connections</td>
<td>115</td>
</tr>
<tr>
<td>6.5.7</td>
<td>GRU vs. LSTM</td>
<td>115</td>
</tr>
<tr>
<td>6.6</td>
<td>Sequence to Sequence</td>
<td>117</td>
</tr>
<tr>
<td>6.7</td>
<td>Attention</td>
<td>118</td>
</tr>
<tr>
<td>6.8</td>
<td>Embeddings</td>
<td>121</td>
</tr>
<tr>
<td>6.9</td>
<td>Introduction to Transformers</td>
<td>122</td>
</tr>
<tr>
<td>6.10</td>
<td>Summary</td>
<td>123</td>
</tr>
<tr>
<td>7</td>
<td>Graph Neural Networks</td>
<td>124</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>124</td>
</tr>
<tr>
<td>7.2</td>
<td>Definitions</td>
<td>126</td>
</tr>
<tr>
<td>7.3</td>
<td>Embeddings</td>
<td>130</td>
</tr>
<tr>
<td>7.4</td>
<td>Node Similarity</td>
<td>132</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Adjacency-based Similarity</td>
<td>132</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Multi-hop Similarity</td>
<td>132</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Overlap Similarity</td>
<td>132</td>
</tr>
<tr>
<td>7.4.4</td>
<td>Random Walk Embedding</td>
<td>133</td>
</tr>
<tr>
<td>7.4.5</td>
<td>Graph Neural Network Properties</td>
<td>135</td>
</tr>
<tr>
<td>7.5</td>
<td>Neighborhood Aggregation in Graph Neural Networks</td>
<td>136</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Supervised Node Classification Using a GNN</td>
<td>138</td>
</tr>
<tr>
<td>7.6</td>
<td>Graph Neural Network Variants</td>
<td>138</td>
</tr>
<tr>
<td>7.6.1</td>
<td>Graph Convolution Network</td>
<td>138</td>
</tr>
<tr>
<td>7.6.2</td>
<td>GraphSAGE</td>
<td>139</td>
</tr>
<tr>
<td>7.6.3</td>
<td>Gated Graph Neural Networks</td>
<td>139</td>
</tr>
</tbody>
</table>
# Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.6.4</td>
<td>Graph Attention Networks</td>
<td>140</td>
</tr>
<tr>
<td>7.6.5</td>
<td>Message-Passing Networks</td>
<td>140</td>
</tr>
<tr>
<td>7.7</td>
<td>Applications</td>
<td>140</td>
</tr>
<tr>
<td>7.8</td>
<td>Software Libraries, Benchmarks, and Visualization</td>
<td>141</td>
</tr>
<tr>
<td>7.9</td>
<td>Summary</td>
<td>141</td>
</tr>
<tr>
<td>8</td>
<td>Transformers</td>
<td>142</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>142</td>
</tr>
<tr>
<td>8.2</td>
<td>General-Purpose Transformer-Based Architectures</td>
<td>143</td>
</tr>
<tr>
<td>8.2.1</td>
<td>BERT</td>
<td>143</td>
</tr>
<tr>
<td>8.3</td>
<td>Self-Attention</td>
<td>143</td>
</tr>
<tr>
<td>8.4</td>
<td>Multi-head Attention</td>
<td>144</td>
</tr>
<tr>
<td>8.5</td>
<td>Transformer</td>
<td>145</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Positional Encoding</td>
<td>145</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Encoder</td>
<td>145</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Decoder</td>
<td>146</td>
</tr>
<tr>
<td>8.5.4</td>
<td>Pre-training and Fine-tuning</td>
<td>146</td>
</tr>
<tr>
<td>8.6</td>
<td>Transformer Models</td>
<td>146</td>
</tr>
<tr>
<td>8.6.1</td>
<td>Autoencoding Transformers</td>
<td>146</td>
</tr>
<tr>
<td>8.6.2</td>
<td>Auto-regressive Transformers</td>
<td>147</td>
</tr>
<tr>
<td>8.6.3</td>
<td>Sequence-to-Sequence Transformers</td>
<td>147</td>
</tr>
<tr>
<td>8.6.4</td>
<td>GPT-3</td>
<td>148</td>
</tr>
<tr>
<td>8.7</td>
<td>Vision Transformers</td>
<td>148</td>
</tr>
<tr>
<td>8.8</td>
<td>Multi-modal Transformers</td>
<td>149</td>
</tr>
<tr>
<td>8.9</td>
<td>Text and Code Transformers</td>
<td>149</td>
</tr>
<tr>
<td>8.10</td>
<td>Summary</td>
<td>150</td>
</tr>
</tbody>
</table>

## Part III Generative Models

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Generative Adversarial Networks</td>
<td>153</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>153</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Progress</td>
<td>153</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Game Theory</td>
<td>154</td>
</tr>
<tr>
<td>9.1.3</td>
<td>Co-evolution</td>
<td>155</td>
</tr>
<tr>
<td>9.2</td>
<td>Minimax Optimization</td>
<td>155</td>
</tr>
<tr>
<td>9.3</td>
<td>Divergence between Distributions</td>
<td>157</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Least Squares GAN</td>
<td>158</td>
</tr>
<tr>
<td>9.3.2</td>
<td>$f$-GAN</td>
<td>158</td>
</tr>
<tr>
<td>9.4</td>
<td>Optimal Objective Value</td>
<td>158</td>
</tr>
<tr>
<td>9.5</td>
<td>Gradient Descent Ascent</td>
<td>158</td>
</tr>
<tr>
<td>9.6</td>
<td>Optimistic Gradient Descent Ascent</td>
<td>159</td>
</tr>
<tr>
<td>9.7</td>
<td>GAN Training</td>
<td>160</td>
</tr>
<tr>
<td>9.7.1</td>
<td>Discriminator Training</td>
<td>160</td>
</tr>
</tbody>
</table>
Contents

9.7.2 Generator Training ........................................ 160
9.7.3 Alternating Discriminator–Generator Training .......... 161
9.8 GAN Losses .................................................. 162
  9.8.1 Wasserstein GAN .................................. 162
  9.8.2 Unrolled GAN ...................................... 163
9.9 GAN Architectures ........................................... 164
  9.9.1 Progressive GAN ..................................... 164
  9.9.2 Deep Convolutional GAN ............................ 164
  9.9.3 Semi-Supervised GAN ............................... 164
  9.9.4 Conditional GAN ..................................... 164
  9.9.5 Image-to-Image Translation .......................... 165
  9.9.6 Cycle-Consistent GAN ............................... 165
  9.9.7 Registration GAN ..................................... 167
  9.9.8 Self-Attention GAN and BigGAN .................... 167
  9.9.9 Composition and Control with GANs .................. 167
  9.9.10 Instance Conditioned GAN .......................... 168
9.10 Evaluation .................................................. 168
  9.10.1 Inception Score ..................................... 168
  9.10.2 Frechet Inception Distance ......................... 169
9.11 Applications .................................................. 169
  9.11.1 Super Resolution and Restoration ..................... 169
  9.11.2 Style Synthesis ..................................... 169
  9.11.3 Image Completion ................................... 169
  9.11.4 De-raining ........................................... 170
  9.11.5 Map Synthesis ...................................... 170
  9.11.6 Pose Synthesis ..................................... 170
  9.11.7 Face Editing ......................................... 170
  9.11.8 Training Data Generation ............................ 170
  9.11.9 Text-to-Image Synthesis ............................ 170
  9.11.10 Medical Imaging .................................... 171
  9.11.11 Video Synthesis .................................... 171
  9.11.12 Motion Retargeting .................................. 171
  9.11.13 3D Synthesis ....................................... 171
  9.11.14 Graph Synthesis .................................... 172
  9.11.15 Autonomous Vehicles ............................... 172
  9.11.16 Text-to-Speech Synthesis .......................... 172
  9.11.17 Voice Conversion .................................... 172
  9.11.18 Music Synthesis ..................................... 172
  9.11.19 Protein Design ..................................... 172
  9.11.20 Natural Language Synthesis ......................... 173
  9.11.21 Cryptography ....................................... 173
9.12 Software Libraries, Benchmarks, and Visualization .. 173
9.13 Summary ..................................................... 173
## Contents

10 Variational Autoencoders 174

10.1 Introduction .................................... 174
10.2 Variational Inference ............................. 174
10.2.1 Reverse KL ................................ 176
10.2.2 Score Gradient ............................ 178
10.2.3 Reparameterization Gradient ................. 179
10.2.4 Forward KL ................................ 180
10.3 Variational Autoencoder ......................... 181
10.3.1 Autoencoder ............................... 181
10.3.2 Variational Autoencoder ..................... 182
10.4 Generative Flows ................................ 184
10.5 Denoising Diffusion Probabilistic Model ............ 186
10.5.1 Forward Noising Process ..................... 186
10.5.2 Reverse Generation by Sampling ............... 186
10.6 Geometric Variational Inference .................. 187
10.6.1 Moser Flow ................................ 188
10.6.2 Riemannian Score-Based Generative Models ...... 188
10.7 Software Libraries ................................ 189
10.8 Summary ..................................... 189

Part IV Reinforcement Learning

11 Reinforcement Learning 193

11.1 Introduction .................................. 193
11.2 Multi-Armed Bandit ............................. 193
11.2.1 Greedy Approach ............................ 194
11.2.2 $\epsilon$-greedy Approach ..................... 194
11.2.3 Upper Confidence Bound ..................... 196
11.3 State Machines ................................ 196
11.4 Markov Processes ............................... 196
11.5 Markov Decision Processes ....................... 199
11.5.1 State of Environment and Agent ............. 200
11.6 Definitions .................................. 202
11.6.1 Policy ................................ 202
11.6.2 State Action Diagram ....................... 203
11.6.3 State Value Function ......................... 204
11.6.4 Action Value Function ....................... 207
11.6.5 Reward ................................ 209
11.6.6 Model ................................ 210
11.6.7 Agent Types ................................ 210
11.6.8 Problem Types ............................. 211
11.6.9 Agent Representation of State ................. 211
11.6.10 Bellman Expectation Equation for State Value Function 212
Contents

11.6.11 Bellman Expectation Equation for Action Value Function 214
11.7 Optimal Policy ........................................ 214
   11.7.1 Optimal Value Function ......................... 215
   11.7.2 Bellman Optimality Equation for $V_*$ ............ 215
   11.7.3 Bellman Optimality Equation for $Q_*$ .............. 216
11.8 Planning by Dynamic Programming with a Known MDP .... 218
   11.8.1 Iterative Policy Evaluation ....................... 218
   11.8.2 Policy Iteration .................................. 218
   11.8.3 Infinite Horizon Value Iteration .................. 219
11.9 Reinforcement Learning ................................. 219
   11.9.1 Model-Based Reinforcement Learning .............. 221
   11.9.2 Policy Search .................................... 222
   11.9.3 Monte Carlo Sampling ............................ 222
   11.9.4 Temporal Difference Sampling .................... 222
   11.9.5 $Q$-Learning .................................... 224
   11.9.6 Sarsa ........................................... 226
   11.9.7 On-Policy vs. Off-Policy Methods ................. 227
   11.9.8 Sarsa($\lambda$) ................................... 227
11.10 Maximum Entropy Reinforcement Learning ............. 227
11.11 Summary ............................................. 228

12 Deep Reinforcement Learning 229
   12.1 Introduction ........................................ 229
   12.2 Function Approximation ............................. 230
       12.2.1 State Value Function Approximation ............ 230
       12.2.2 Action Value Function Approximation ........... 231
   12.3 Value-Based Methods ................................. 232
       12.3.1 Experience Replay .............................. 232
       12.3.2 Neural Fitted $Q$-Iteration ..................... 233
       12.3.3 Deep $Q$-Network .............................. 233
       12.3.4 Target Network ................................ 234
       12.3.5 Algorithm ..................................... 234
       12.3.6 Prioritized Replay .............................. 234
       12.3.7 Double DQN .................................... 235
       12.3.8 Dueling Networks ............................... 235
   12.4 Policy-Based Methods ................................. 236
       12.4.1 Policy Gradient ................................ 237
       12.4.2 REINFORCE ..................................... 238
       12.4.3 Subtracting a Baseline ......................... 238
   12.5 Actor–Critic Methods ................................ 239
       12.5.1 Advantage Actor–Critic ......................... 240
       12.5.2 Asynchronous Advantage Actor–Critic .......... 240
       12.5.3 Importance Sampling ......................... 241
       12.5.4 Surrogate Loss ................................ 241
# Contents

12.5.5 Natural Policy Gradient ........................................ 242
12.5.6 Trust Region Policy Optimization .......................... 243
12.5.7 Proximal Policy Optimization ............................... 244
12.5.8 Deep Deterministic Policy Gradient ....................... 244

12.6 Model-Based Reinforcement Learning ....................... 245
12.6.1 Monte Carlo Tree Search ................................. 245
12.6.2 Expert Iteration and AlphaZero .......................... 246
12.6.3 World Models ............................................ 247

12.7 Imitation Learning ............................................ 248

12.8 Exploration .................................................. 249
12.8.1 Sparse Rewards .......................................... 249

12.9 Summary ....................................................... 249

---

## Part V Applications

13 Applications 253

13.1 Introduction .................................................. 253

13.2 Autonomous Vehicles ......................................... 253

13.3 Climate Change and Climate Monitoring .................... 255
13.3.1 Predicting Ocean Biogeochemistry ....................... 255
13.3.2 Predicting Atlantic Multidecadal Variability .......... 258
13.3.3 Predicting Wildfire Growth ............................... 261

13.4 Computer Vision ............................................... 262
13.4.1 Kinship Verification ...................................... 262
13.4.2 Image-to-3D ................................................ 263
13.4.3 Image2LEGO® .............................................. 264
13.4.4 Imaging through Scattering Media ....................... 265
13.4.5 Contrastive Language-Image Pre-training ............... 269

13.5 Speech and Audio Processing .................................. 269
13.5.1 Audio Reverb Impulse Response Synthesis .............. 269
13.5.2 Voice Swapping ............................................ 270
13.5.3 Explainable Musical Phrase Completion .................. 271

13.6 Natural Language Processing .................................. 273
13.6.1 Quantifying and Alleviating Distribution Shifts in Founda-
      tion Models on Review Classification ....................... 273

13.7 Automated Machine Learning .................................. 275

13.8 Education ..................................................... 278
13.8.1 Learning-to-Learn STEM Courses ........................ 278

13.9 Proteomics ...................................................... 280
13.9.1 Protein Structure Prediction ............................ 280
13.9.2 Protein Docking .......................................... 284

13.10 Combinatorial Optimization .................................. 285
13.10.1 Problems over Graphs .................................. 288
13.10.2 Learning Graph Algorithms as Single-Player Games . 289
13.11 Physics ........................................... 289
13.11.1 Pedestrian Wind Estimation in Urban Environments . 289
13.11.2 Fusion Plasma .................................. 290
13.12 Summary ........................................... 291

Appendix A Matrix Calculus 293
A.1 Gradient Computations for Backpropagation .......... 293
A.1.1 Scalar by Vector ................................ 293
A.1.2 Scalar by Matrix ................................ 293
A.1.3 Vector by Vector ................................ 294
A.1.4 Matrix by Scalar ................................ 294
A.2 Gradient Computations for Optimization .......... 294
A.2.1 Dot Product by Vector .......................... 294
A.2.2 Quadratic Form by Vector ....................... 295

Appendix B Scientific Writing and Reviewing Best Practices 296
B.1 Writing Best Practices ............................... 296
B.1.1 Introduction ..................................... 296
B.1.2 Methods ......................................... 296
B.1.3 Figures and Tables ............................... 297
B.1.4 Results .......................................... 297
B.1.5 Abbreviations and Notation ....................... 297
B.2 Reviewing Best Practices ............................ 297
B.2.1 Ranking .......................................... 298
B.2.2 Rebuttal .......................................... 298

References 299
Index 335
Preface

This book provides comprehensive and clear coverage of deep learning, which has transformed the field of artificial intelligence. The book is distinctive in that it uses a unified notation, high-quality illustrated figures and the most up-to-date material in the field, and is accompanied by hundreds of code samples, exercises, and solutions on each topic, automatically generated by program synthesis. *The Science of Deep Learning* emerged from courses taught by the author in the past five years that have provided thousands of students with training and experience for their academic studies, and prepared them for careers in deep learning, machine learning, and artificial intelligence in leading companies in industry and academia. The motivation for the book is to provide a guide to the field built upon clear visualizations using a unified notation and equations. The content is self-contained, using a unified language so that students, teachers, and researchers in academia and industry can use the book without having to overcome the barriers to entry of the specific language and notation of each topic. Introductory topics are represented using both basic linear algebra and graphs simultaneously, along with the corresponding algorithms.

Coverage

The material is presented in five main parts:

1. Part I, on the foundations of deep learning, includes Chapters 1–4, which covers core deep learning material on forward and backpropagation, optimization, and regularization.
2. Part II, on deep learning architectures, includes Chapters 5–8. This covers key architectures including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTMs), gated recurrent units (GRUs), graph neural networks (GNNs), and Transformers.
3. Part III, on generative models, comprises Chapters 9 and 10, which cover generative adversarial networks (GANs) and variational autoencoders (VAEs).
4. Part IV addresses reinforcement learning and deep reinforcement learning (Chapters 11 and 12).
5. Part V, on applications (Chapter 13), covers a broad range of deep learning...
applications, which are also distributed among the chapters by topic and relevance.

6. Appendices provide equations for computing gradients in backpropagation and optimization, and best practices in scientific writing and reviewing.

Contribution

The book contributes to the literature in the field in that it uses rigorous math with a unified notation. In addition, over the past five years, during the course of instruction, advanced topics have been simplified to become part of the core, while bringing in new topics in the field as advanced topics. The key advantages of this book are that it is up-to-date, with the latest advances in the field including unique content; the math is rigorous, using a unified notation; and the book presents comprehensive algorithms and uses high-quality figures.

Audience and Prerequisite Knowledge

The book is intended for students and researchers in academia and industry, as well as lecturers in academia. The book is primarily intended for computer science undergraduate and graduate students, as well as advanced PhD students. This book has been used for teaching students mainly in computer science, electrical engineering, data science, statistics, and operations research. The required background is linear algebra and calculus. Optional background is machine learning and programming experience. The book is also applicable for a wide audience of students pursuing degrees in STEM fields with the required background. The book is useful for researchers in academia and industry, as well as data scientists and algorithm developers of artificial intelligence. Finally, the book may be used by lecturers in academia for teaching a course on deep learning, and chapters may be used in teaching topics in courses on machine learning, data science, optimization, and reinforcement learning at the undergraduate and graduate levels.

Usage

The first four parts of the book have been used as a textbook in courses on deep learning. The third part, on generative models, may be used as part of a course on unsupervised learning. The fourth part may be used as part of a course on reinforcement learning or deep reinforcement learning. Appendix A is useful for computing the gradients in backpropagation and optimization. Appendix B may be used in project-based courses for providing best practices in scientific writing and reviewing.
Acknowledgments

I developed the material for this book over the past five years while teaching a dozen deep learning classes. The book is now used as the new textbook in deep learning at Columbia University. I would like to thank Columbia University students that read each chapter before class and reported errata that I fixed and improved the book, MIT students and colleagues for reading the book and providing feedback, Leslie Goodman and Gary Daniel Smith for copy editing, Michal Solel Elnekave and Nikhil Singh, who were helpful in producing high-quality figures in Illustrator from sketches, and Maggie Jeffers and Lauren Cowles, my editors who encouraged me to see the book to completion. I wish to thank numerous colleagues, particularly Kyunghyun Cho and Claudio Silva of NYU, Nakul Verma and Itsik Pe’er of Columbia University, Tonio Buonassisi and Gilbert Strang of MIT, Dov Te’eni of Tel Aviv University, and Madeleine Udell and David Williamson of Cornell University. Finally, I’d like to thank my family, Adi, Danielle, Yael, Sharon, and Gilly, my sister Dr. Tali Drori Snir, mom Nili and special thanks to my dad, Prof. Israel Drori, who has published over a dozen books and gets to read my first one.
# Abbreviations

- **A2C**: advantage actor–critic
- **A3C**: asynchronous advantage actor–critic
- **AMV**: Atlantic Multidecadal Variability
- **AVI**: amortized variational inference
- **BBVI**: black-box variational inference
- **BCE**: binary cross entropy
- **BERT**: bidirectional encoder representations from Transformers
- **BFGS**: Broyden–Fletcher–Goldfarb–Shanno (correction)
- **CFD**: computational fluid dynamics
- **CGAN**: conditional GAN
- **CLIP**: contrastive language-image pre-training
- **CNN**: convolutional neural network
- **DAG**: directed acyclic graph
- **DCGAN**: deep convolutional generative adversarial network
- **DDPG**: deep deterministic policy gradient
- **DDPM**: denoising diffusion probabilistic model
- **DFP**: Davidon–Fletcher–Powell (correction)
- **DMD**: digital micromirror device
- **DQN**: deep Q-network
- **ELBO**: evidence lower bound
- **EMD**: Earth mover’s distance
- **ENSO**: El Niño-Southern Oscillation
- **ESM**: Earth system model
- **FID**: Frechet inception distance
- **FKL**: forward KL
- **GAE**: generalized advantage estimation
- **GAN**: generative adversarial network
- **GAT**: graph attention network
- **GCN**: graph convolutional network
- **GDA**: Gradient descent ascent
- **GNN**: graph neural network
- **GPT-3**: generative pre-trained Transformer 3
- **GRU**: gated recurrent unit
- **IID**: independent and identically distributed
Abbreviations and Notation

IR      impulse response
IS      inception score
JS      Jensen–Shannon (divergence)
KL      Kullback–Leibler
LSTM    long short-term memory
MAE     mean absolute error
MC      Monte Carlo
MCMC    Markov chain Monte Carlo
MCTS    Monte Carlo tree search
MDP     Markov decision process
MFVI    mean-field variational inference
MST     minimum spanning tree
NAS     neural architecture search
NPCC    negative Pearson correlation coefficient
NPG     natural policy gradient
OGDA    optimistic gradient descent ascent
PCA     principle component analysis
PPO     proximal policy optimization
ReLU    rectified linear unit function
RFIW    Recognizing Families in the Wild
RKL     reverse KL
RNN     recurrent neural network
seq2seq sequence-to-sequence (models)
SGAN    semi-supervised GAN
SGD     stochastic gradient descent
SLM     spatial light modulator
SSP     single-source shortest paths
SSS     sea surface salinity
SST     sea-surface temperatures
TD      temporal difference
TRPO    trust region policy optimization
TSP     traveling salesman problem
UCB     upper confidence bound
VAE     variational autoencoder
VI      variational inference
VQ-VAE  vector quantized variational autoencoder
VRN     Volumetric Regression Network
VRP     vehicle routing problem
WGAN    Wasserstein GAN
Abbreviations and Notation

Notation

General

\( \mathbb{E} \)  
expectation

\( \mathbb{R} \)  
real numbers

\( I \)  
identity matrix

\( X^T \)  
matrix transpose

\( \|x\|_p \)  
\( \ell_p \) norm of vector \( x \)

\( \cap \)  
intersection

\( \cup \)  
union

\( \| \)  
concatenation of vectors

\( \mathcal{N}(\mu, \sigma) \)  
Gaussian distribution with mean \( \mu \) and standard deviation \( \sigma \)

\( \nabla_x y \)  
gradient of \( y \) with respect to \( x \)

Neural Networks

\( \alpha \)  
learning rate

\( \theta \)  
learning parameter

\( \ell \)  
network layer index

\( W^\ell \)  
weight matrix of layer

\( z^\ell \)  
pre-activation vector of layer

\( Z^\ell \)  
pre-activation matrix of layer

\( a^\ell \)  
activation vector of layer

\( A^\ell \)  
activation matrix of layer

\( f^\ell \)  
non-linear activation function of layer

\( \sigma \)  
sigmoid function

\( \mathcal{L} \)  
loss function

\( \mathcal{R} \)  
regularization function

Convolutional Neural Networks

\( f * g \)  
convolution of functions \( f \) and \( g \)

Sequence Models

\( x_t \)  
input vector at time \( t \)

\( h_t \)  
hidden vector at time \( t \)

\( y_t \)  
output vector at time \( t \)

\( U \)  
weight matrix applied to input vector shared across time

\( V \)  
weight matrix applied to hidden vector shared across time

\( W \)  
weight matrix applied to previous hidden vector shared across time
### Abbreviations and Notation

#### Graph Neural Networks
- $G$: graph
- $V$: graph nodes
- $E$: graph edges
- $N(i)$: neighbors of node $i$
- $A$: graph adjacency matrix
- $D$: graph diagonal degree matrix
- $L$: graph Laplacian matrix
- $L_{\text{sym}}$: symmetric normalized Laplacian matrix
- $L_{\text{rw}}$: random walk normalized Laplacian matrix
- $h^\ell$: embedding vector of layer $\ell$

#### Generative Models
- $D$: GAN discriminator
- $G$: GAN generator
- $D_{\text{KL}}$: Kullback–Leibler divergence
- $D_{\text{JS}}$: Jenson–Shannon divergence

#### Reinforcement Learning
- $s$: state
- $S$: set of states
- $a$: action
- $A$: set of actions
- $T(s,a,s')$: transition function from state $s$ and action $a$ to next state $s'$
- $r$: reward
- $R(s,a)$: reward for state $s$ and action $a$
- $\gamma$: reward discount factor
- $g_t$: return at time step $t$
- $\pi$: policy
- $\pi(a|s)$: probability of taking action $a$ in state $s$ under policy $\pi$
- $h$: horizon
- $V^h_\pi(s)$: state value function with respect to policy $\pi$ with horizon $h$ of state $s$
- $Q^h_\pi(s,a)$: action value function with respect to policy $\pi$ with horizon $h$ of state $s$ and action $a$
- $\pi^*$: optimal policy
- $V_*(s)$: state value function with respect to optimal policy $\pi^*$ for state $s$
- $Q_*(s,a)$: action value function with respect to optimal policy $\pi^*$ for state $s$ and action $a$