

---

# Table of Contents

<b>Preface.....</b>	<b>xv</b>
<b>1. Exploring the Landscape of Artificial Intelligence.....</b>	<b>1</b>
An Apology	2
The Real Introduction	3
What Is AI?	3
Motivating Examples	5
A Brief History of AI	6
Exciting Beginnings	6
The Cold and Dark Days	8
A Glimmer of Hope	8
How Deep Learning Became a Thing	12
Recipe for the Perfect Deep Learning Solution	15
Datasets	16
Model Architecture	18
Frameworks	20
Hardware	23
Responsible AI	26
Bias	27
Accountability and Explainability	30
Reproducibility	30
Robustness	31
Privacy	31
Summary	32
Frequently Asked Questions	32

<b>2. What's in the Picture: Image Classification with Keras. ....</b>	<b>35</b>
Introducing Keras	36
Predicting an Image's Category	37
Investigating the Model	42
ImageNet Dataset	42
Model Zoos	44
Class Activation Maps	46
Summary	48
<b>3. Cats Versus Dogs: Transfer Learning in 30 Lines with Keras. ....</b>	<b>49</b>
Adapting Pretrained Models to New Tasks	50
A Shallow Dive into Convolutional Neural Networks	51
Transfer Learning	53
Fine Tuning	54
How Much to Fine Tune	55
Building a Custom Classifier in Keras with Transfer Learning	56
Organize the Data	57
Build the Data Pipeline	59
Number of Classes	60
Batch Size	61
Data Augmentation	61
Model Definition	65
Train the Model	65
Set Training Parameters	65
Start Training	66
Test the Model	68
Analyzing the Results	68
Further Reading	76
Summary	78
<b>4. Building a Reverse Image Search Engine: Understanding Embeddings. ....</b>	<b>79</b>
Image Similarity	80
Feature Extraction	83
Similarity Search	86
Visualizing Image Clusters with t-SNE	90
Improving the Speed of Similarity Search	94
Length of Feature Vectors	94
Reducing Feature-Length with PCA	96
Scaling Similarity Search with Approximate Nearest Neighbors	100
Approximate Nearest-Neighbor Benchmark	101

Which Library Should I Use?	102
Creating a Synthetic Dataset	102
Brute Force	102
Annoy	103
NGT	104
Faiss	104
Improving Accuracy with Fine Tuning	104
Fine Tuning Without Fully Connected Layers	108
Siamese Networks for One-Shot Face Verification	109
Case Studies	110
Flickr	111
Pinterest	111
Celebrity Doppelgangers	112
Spotify	113
Image Captioning	114
Summary	116
 <b>5. From Novice to Master Predictor: Maximizing Convolutional</b>	
<b>Neural Network Accuracy.....</b>	<b>117</b>
Tools of the Trade	118
TensorFlow Datasets	119
TensorBoard	120
What-If Tool	123
tf-explain	128
Common Techniques for Machine Learning Experimentation	130
Data Inspection	130
Breaking the Data: Train, Validation, Test	131
Early Stopping	132
Reproducible Experiments	132
End-to-End Deep Learning Example Pipeline	133
Basic Transfer Learning Pipeline	133
Basic Custom Network Pipeline	135
How Hyperparameters Affect Accuracy	136
Transfer Learning Versus Training from Scratch	137
Effect of Number of Layers Fine-Tuned in Transfer Learning	138
Effect of Data Size on Transfer Learning	139
Effect of Learning Rate	140
Effect of Optimizers	141
Effect of Batch Size	141
Effect of Resizing	142

Effect of Change in Aspect Ratio on Transfer Learning	143
Tools to Automate Tuning for Maximum Accuracy	144
Keras Tuner	144
AutoAugment	146
AutoKeras	147
Summary	148
<b>6. Maximizing Speed and Performance of TensorFlow: A Handy Checklist.....</b>	<b>149</b>
GPU Starvation	149
nvidia-smi	150
TensorFlow Profiler + TensorBoard	152
How to Use This Checklist	153
Performance Checklist	154
Data Preparation	154
Data Reading	154
Data Augmentation	154
Training	154
Inference	155
Data Preparation	155
Store as TFRecords	155
Reduce Size of Input Data	157
Use TensorFlow Datasets	157
Data Reading	158
Use tf.data	158
Prefetch Data	158
Parallelize CPU Processing	159
Parallelize I/O and Processing	159
Enable Nondeterministic Ordering	160
Cache Data	160
Turn on Experimental Optimizations	161
Autotune Parameter Values	163
Data Augmentation	164
Use GPU for Augmentation	164
Training	165
Use Automatic Mixed Precision	165
Use Larger Batch Size	166
Use Multiples of Eight	168
Find the Optimal Learning Rate	168
Use tf.function	170
Overtrain, and Then Generalize	172

Install an Optimized Stack for the Hardware	173
Optimize the Number of Parallel CPU Threads	175
Use Better Hardware	176
Distribute Training	177
Examine Industry Benchmarks	178
Inference	180
Use an Efficient Model	180
Quantize the Model	183
Prune the Model	185
Use Fused Operations	186
Enable GPU Persistence	186
Summary	187
<b>7. Practical Tools, Tips, and Tricks. ....</b>	<b>189</b>
Installation	189
Training	191
Model	192
Data	193
Privacy	196
Education and Exploration	197
One Last Question	198
<b>8. Cloud APIs for Computer Vision: Up and Running in 15 Minutes. ....</b>	<b>201</b>
The Landscape of Visual Recognition APIs	203
Clarifai	203
Microsoft Cognitive Services	204
Google Cloud Vision	204
Amazon Rekognition	205
IBM Watson Visual Recognition	206
Algorithmia	208
Comparing Visual Recognition APIs	209
Service Offerings	209
Cost	210
Accuracy	211
Bias	212
Getting Up and Running with Cloud APIs	217
Training Our Own Custom Classifier	219
Top Reasons Why Our Classifier Does Not Work Satisfactorily	224
Comparing Custom Classification APIs	225
Performance Tuning for Cloud APIs	228

Effect of Resizing on Image Labeling APIs	228
Effect of Compression on Image Labeling APIs	229
Effect of Compression on OCR APIs	230
Effect of Resizing on OCR APIs	230
Case Studies	231
The New York Times	231
Uber	232
Giphy	233
OmniEarth	234
Photobucket	234
Staples	235
InDro Robotics	235
Summary	237
<b>9. Scalable Inference Serving on Cloud with TensorFlow Serving and KubeFlow. . .</b>	<b>239</b>
Landscape of Serving AI Predictions	240
Flask: Build Your Own Server	242
Making a REST API with Flask	242
Deploying a Keras Model to Flask	243
Pros of Using Flask	244
Cons of Using Flask	244
Desirable Qualities in a Production-Level Serving System	245
High Availability	245
Scalability	245
Low Latency	246
Geographic Availability	246
Failure Handling	247
Monitoring	247
Model Versioning	248
A/B Testing	248
Support for Multiple Machine Learning Libraries	248
Google Cloud ML Engine: A Managed Cloud AI Serving Stack	248
Pros of Using Cloud ML Engine	249
Cons of Using Cloud ML Engine	249
Building a Classification API	249
TensorFlow Serving	256
Installation	256
KubeFlow	258
Pipelines	260
Fairing	260

Installation	261
Price Versus Performance Considerations	263
Cost Analysis of Inference-as-a-Service	263
Cost Analysis of Building Your Own Stack	265
Summary	266
<b>10. AI in the Browser with TensorFlow.js and ml5.js. ....</b>	<b>267</b>
JavaScript-Based Machine Learning Libraries: A Brief History	268
ConvNetJS	269
Keras.js	270
ONNX.js	270
TensorFlow.js	272
TensorFlow.js Architecture	273
Running Pretrained Models Using TensorFlow.js	275
Model Conversion for the Browser	277
Training in the Browser	277
Feature Extraction	278
Data Collection	279
Training	280
GPU Utilization	282
ml5.js	283
PoseNet	286
pix2pix	290
Benchmarking and Practical Considerations	295
Model Size	295
Inference Time	296
Case Studies	298
Semi-Conductor	298
TensorSpace	299
Metacar	300
Airbnb's Photo Classification	301
GAN Lab	301
Summary	302
<b>11. Real-Time Object Classification on iOS with Core ML. ....</b>	<b>303</b>
The Development Life Cycle for Artificial Intelligence on Mobile	305
A Brief History of Core ML	306
Alternatives to Core ML	308
TensorFlow Lite	308
ML Kit	309

Fritz	309
Apple's Machine Learning Architecture	309
Domain-Based Frameworks	310
ML Framework	311
ML Performance Primitives	311
Building a Real-Time Object Recognition App	312
Conversion to Core ML	319
Conversion from Keras	319
Conversion from TensorFlow	319
Dynamic Model Deployment	321
On-Device Training	322
Federated Learning	323
Performance Analysis	323
Benchmarking Models on iPhones	324
Measuring Energy Impact	327
Benchmarking Load	331
Reducing App Size	333
Avoid Bundling the Model	334
Use Quantization	334
Use Create ML	336
Case Studies	336
Magic Sudoku	336
Seeing AI	338
HomeCourt	338
InstaSaber + YoPuppet	339
Summary	342
<b>12. Not Hotdog on iOS with Core ML and Create ML.....</b>	<b>343</b>
Collecting Data	345
Approach 1: Find or Collect a Dataset	345
Approach 2: Fatkun Chrome Browser Extension	346
Approach 3: Web Scraper Using Bing Image Search API	349
Training Our Model	350
Approach 1: Use Web UI-based Tools	350
Approach 2: Use Create ML	355
Approach 3: Fine Tuning Using Keras	361
Model Conversion Using Core ML Tools	361
Building the iOS App	361
Further Exploration	363
Summary	363



<b>13. Shazam for Food: Developing Android Apps with TensorFlow Lite and ML Kit . . .</b>	<b>365</b>
The Life Cycle of a Food Classifier App	366
An Overview of TensorFlow Lite	368
TensorFlow Lite Architecture	371
Model Conversion to TensorFlow Lite	372
Building a Real-Time Object Recognition App	373
ML Kit + Firebase	382
Object Classification in ML Kit	384
Custom Models in ML Kit	384
Hosted Models	385
A/B Testing Hosted Models	391
Using the Experiment in Code	397
TensorFlow Lite on iOS	397
Performance Optimizations	397
Quantizing with TensorFlow Lite Converter	398
TensorFlow Model Optimization Toolkit	398
Fritz	399
A Holistic Look at the Mobile AI App Development Cycle	402
How Do I Collect Initial Data?	402
How Do I Label My Data?	403
How Do I Train My Model?	403
How Do I Convert the Model to a Mobile-Friendly Format?	403
How Do I Make my Model Performant?	404
How Do I Build a Great UX for My Users?	404
How Do I Make the Model Available to My Users?	404
How Do I Measure the Success of My Model?	405
How Do I Improve My Model?	405
How Do I Update the Model on My Users' Phones?	406
The Self-Evolving Model	406
Case Studies	408
Lose It!	408
Portrait Mode on Pixel 3 Phones	410
Speaker Recognition by Alibaba	411
Face Contours in ML Kit	411
Real-Time Video Segmentation in YouTube Stories	412
Summary	413
<b>14. Building the Purrfect Cat Locator App with TensorFlow Object Detection API. . .</b>	<b>415</b>
Types of Computer-Vision Tasks	417
Classification	417

Localization	417
Detection	417
Segmentation	418
Approaches to Object Detection	420
Invoking Prebuilt Cloud-Based Object Detection APIs	421
Reusing a Pretrained Model	423
Obtaining the Model	423
Test Driving Our Model	424
Deploying to a Device	425
Building a Custom Detector Without Any Code	427
The Evolution of Object Detection	432
Performance Considerations	433
Key Terms in Object Detection	435
Intersection over Union	435
Mean Average Precision	436
Non-Maximum Suppression	436
Using the TensorFlow Object Detection API to Build Custom Models	437
Data Collection	437
Labeling the Data	441
Preprocessing the Data	445
Inspecting the Model	446
Training	448
Model Conversion	450
Image Segmentation	452
Case Studies	453
Smart Refrigerator	453
Crowd Counting	454
Face Detection in Seeing AI	456
Autonomous Cars	457
Summary	458
<b>15. Becoming a Maker: Exploring Embedded AI at the Edge.....</b>	<b>459</b>
Exploring the Landscape of Embedded AI Devices	460
Raspberry Pi	462
Intel Movidius Neural Compute Stick	464
Google Coral USB Accelerator	465
NVIDIA Jetson Nano	466
FPGA + PYNQ	468
Arduino	472
A Qualitative Comparison of Embedded AI Devices	474

Hands-On with the Raspberry Pi	476
Speeding Up with the Google Coral USB Accelerator	478
Port to NVIDIA Jetson Nano	480
Comparing the Performance of Edge Devices	483
Case Studies	484
JetBot	484
Squatting for Metro Tickets	486
Cucumber Sorter	487
Further Exploration	488
Summary	489
<b>16. Simulating a Self-Driving Car Using End-to-End Deep Learning with Keras. . . . .</b>	<b>491</b>
A Brief History of Autonomous Driving	492
Deep Learning, Autonomous Driving, and the Data Problem	493
The “Hello, World!” of Autonomous Driving: Steering Through a	
Simulated Environment	496
Setup and Requirements	496
Data Exploration and Preparation	498
Identifying the Region of Interest	501
Data Augmentation	503
Dataset Imbalance and Driving Strategies	504
Training Our Autonomous Driving Model	509
Drive Data Generator	510
Model Definition	512
Deploying Our Autonomous Driving Model	518
Further Exploration	521
Expanding Our Dataset	522
Training on Sequential Data	522
Reinforcement Learning	523
Summary	523
<b>17. Building an Autonomous Car in Under an Hour: Reinforcement Learning with</b>	
<b>    AWS DeepRacer. . . . .</b>	<b>525</b>
A Brief Introduction to Reinforcement Learning	526
Why Learn Reinforcement Learning with an Autonomous Car?	527
Practical Deep Reinforcement Learning with DeepRacer	529
Building Our First Reinforcement Learning	532
Step 1: Create Model	533
Step 2: Configure Training	534
Step 3: Model Training	541

Step 4: Evaluating the Performance of the Model	542
Reinforcement Learning in Action	544
How Does a Reinforcement Learning System Learn?	544
Reinforcement Learning Theory	548
Reinforcement Learning Algorithm in AWS DeepRacer	551
Deep Reinforcement Learning Summary with DeepRacer as an Example	552
Step 5: Improving Reinforcement Learning Models	553
Racing the AWS DeepRacer Car	558
Building the Track	558
AWS DeepRacer Single-Turn Track Template	559
Running the Model on AWS DeepRacer	559
Driving the AWS DeepRacer Vehicle Autonomously	560
Further Exploration	563
DeepRacer League	563
Advanced AWS DeepRacer	563
AI Driving Olympics	564
DIY Robocars	564
Roborace	565
Summary	566
<b>A. A Crash Course in Convolutional Neural Networks.....</b>	<b>567</b>
<b>Index.....</b>	<b>577</b>