A Guide to Convolutional Neural Networks for Computer Vision

Synthesis Lectures on Computer Vision

Editors
Gérard Medioni, ***University of Southern California***
Sven Dickinson, ***University of Toronto***

Synthesis Lectures on Computer Vision is edited by Gérard Medioni of the University of Southern California and Sven Dickinson of the University of Toronto. The series publishes 50–150 page publications on topics pertaining to computer vision and pattern recognition. The scope will largely follow the purview of premier computer science conferences, such as ICCV, CVPR, and ECCV. Potential topics include, but not are limited to:

- Applications and Case Studies for Computer Vision
- Color, Illumination, and Texture
- Computational Photography and Video
- Early and Biologically-inspired Vision
- Face and Gesture Analysis
- Illumination and Reflectance Modeling
- Image-Based Modeling
- Image and Video Retrieval
- Medical Image Analysis
- Motion and Tracking
- Object Detection, Recognition, and Categorization
- Segmentation and Grouping
- Sensors
- Shape-from-X
- Stereo and Structure from Motion
- Shape Representation and Matching

- Statistical Methods and Learning
- Performance Evaluation
- Video Analysis and Event Recognition

A Guide to Convolutional Neural Networks for Computer Vision
Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun
2018

Covariances in Computer Vision and Machine Learning
Hà Quang Minh and Vittorio Murino
2017

Elastic Shape Analysis of Three-Dimensional Objects
Ian H. Jermyn, Sebastian Kurtek, Hamid Laga, and Anuj Srivastava
2017

The Maximum Consensus Problem: Recent Algorithmic Advances
Tat-Jun Chin and David Suter
2017

Extreme Value Theory-Based Methods for Visual Recognition
Walter J. Scheirer
2017

Data Association for Multi-Object Visual Tracking
Margrit Betke and Zheng Wu
2016

Ellipse Fitting for Computer Vision: Implementation and Applications
Kenichi Kanatani, Yasuyuki Sugaya, and Yasushi Kanazawa
2016

Computational Methods for Integrating Vision and Language
Kobus Barnard
2016

Background Subtraction: Theory and Practice
Ahmed Elgammal
2014

Vision-Based Interaction
Matthew Turk and Gang Hua
2013

Camera Networks: The Acquisition and Analysis of Videos over Wide Areas
Amit K. Roy-Chowdhury and Bi Song
2012

Deformable Surface 3D Reconstruction from Monocular Images
Mathieu Salzmann and Pascal Fua
2010

Boosting-Based Face Detection and Adaptation
Cha Zhang and Zhengyou Zhang
2010

Image-Based Modeling of Plants and Trees
Sing Bing Kang and Long Quan
2009

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2018

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

A Guide to Convolutional Neural Networks for Computer Vision

Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun

ISBN: 978-3-031-00693-7 paperback
ISBN: 978-3-031-01821-3 ebook
ISBN: 978-3-031-00078-2 hardcover

DOI 10.1007/978-3-031-01821-3

A Publication in the Springer Nature series
SYNTHESIS LECTURES ON COMPUTER VISION

Lecture #15
Series Editors: Gérard Medioni, *University of Southern California*
Sven Dickinson, *University of Toronto*
Series ISSN
Print 2153-1056 Electronic 2153-1064

A Guide to Convolutional Neural Networks for Computer Vision

Salman Khan
Data61-CSIRO and Australian National University

Hossein Rahmani
The University of Western Australia, Crawley, WA

Syed Afaq Ali Shah
The University of Western Australia, Crawley, WA

Mohammed Bennamoun
The University of Western Australia, Crawley, WA

SYNTHESIS LECTURES ON COMPUTER VISION #15

ABSTRACT

Computer vision has become increasingly important and effective in recent years due to its wide-ranging applications in areas as diverse as smart surveillance and monitoring, health and medicine, sports and recreation, robotics, drones, and self-driving cars. Visual recognition tasks, such as image classification, localization, and detection, are the core building blocks of many of these applications, and recent developments in Convolutional Neural Networks (CNNs) have led to outstanding performance in these state-of-the-art visual recognition tasks and systems. As a result, CNNs now form the crux of deep learning algorithms in computer vision.

This self-contained guide will benefit those who seek to both understand the theory behind CNNs and to gain hands-on experience on the application of CNNs in computer vision. It provides a comprehensive introduction to CNNs starting with the essential concepts behind neural networks: training, regularization, and optimization of CNNs. The book also discusses a wide range of loss functions, network layers, and popular CNN architectures, reviews the different techniques for the evaluation of CNNs, and presents some popular CNN tools and libraries that are commonly used in computer vision. Further, this text describes and discusses case studies that are related to the application of CNN in computer vision, including image classification, object detection, semantic segmentation, scene understanding, and image generation.

This book is ideal for undergraduate and graduate students, as no prior background knowledge in the field is required to follow the material, as well as new researchers, developers, engineers, and practitioners who are interested in gaining a quick understanding of CNN models.

KEYWORDS

deep learning, computer vision, convolution neural networks, perception, back-propagation, feed-forward networks, image classification, action recognition, object detection, object tracking, video processing, semantic segmentation, scene understanding, 3D processing

SK: To my parents and my wife Nusrat

HR: To my father Shirzad, my mother Rahimeh, and my wife Shahla

AS: To my parents, my wife Maleeha, and our children Abiya, Maryam, and Muhammad. Thanks for always being there for me.

MB: To my parents: Mostefa and Rabia Bennamoun and to my nuclear family: Leila, Miriam, Basheer, and Rayaane Bennamoun

Contents

Preface . . . . . **xvii**

Acknowledgments . . . . . **xix**

1 Introduction **1**

1.1 What is Computer Vision? 1
1.1.1 Applications 1
1.1.2 Image Processing vs. Computer Vision 3
1.2 What is Machine Learning? 4
1.2.1 Why Deep Learning? 6
1.3 Book Overview 6

2 Features and Classifiers **11**

2.1 Importance of Features and Classifiers 11
2.1.1 Features 11
2.1.2 Classifiers 13
2.2 Traditional Feature Descriptors 13
2.2.1 Histogram of Oriented Gradients (HOG) 14
2.2.2 Scale-invariant Feature Transform (SIFT) 16
2.2.3 Speeded-up Robust Features (SURF) 19
2.2.4 Limitations of Traditional Hand-engineered Features 21
2.3 Machine Learning Classifiers 21
2.3.1 Support Vector Machine (SVM) 22
2.3.2 Random Decision Forest 26
2.4 Conclusion 29

3 Neural Networks Basics **31**

3.1 Introduction 31
3.2 Multi-layer Perceptron 32
3.2.1 Architecture Basics 32
3.2.2 Parameter Learning 32
3.3 Recurrent Neural Networks 36

3.3.1 Architecture Basics . . . 37
3.3.2 Parameter Learning . . . 38
3.4 Link with Biological Vision . . . 39
3.4.1 Biological Neuron . . . 39
3.4.2 Computational Model of a Neuron . . . 40
3.4.3 Artificial vs. Biological Neuron . . . 41

4 Convolutional Neural Network . . . 43
4.1 Introduction . . . 43
4.2 Network Layers . . . 44
4.2.1 Pre-processing . . . 44
4.2.2 Convolutional Layers . . . 45
4.2.3 Pooling Layers . . . 53
4.2.4 Nonlinearity . . . 54
4.2.5 Fully Connected Layers . . . 56
4.2.6 Transposed Convolution Layer . . . 57
4.2.7 Region of Interest Pooling . . . 60
4.2.8 Spatial Pyramid Pooling Layer . . . 61
4.2.9 Vector of Locally Aggregated Descriptors Layer . . . 63
4.2.10 Spatial Transformer Layer . . . 63
4.3 CNN Loss Functions . . . 65
4.3.1 Cross-entropy Loss . . . 66
4.3.2 SVM Hinge Loss . . . 66
4.3.3 Squared Hinge Loss . . . 67
4.3.4 Euclidean Loss . . . 67
4.3.5 The ℓ^1 Error . . . 67
4.3.6 Contrastive Loss . . . 67
4.3.7 Expectation Loss . . . 68
4.3.8 Structural Similarity Measure . . . 68

5 CNN Learning . . . 69
5.1 Weight Initialization . . . 69
5.1.1 Gaussian Random Initialization . . . 69
5.1.2 Uniform Random Initialization . . . 70
5.1.3 Orthogonal Random Initialization . . . 70
5.1.4 Unsupervised Pre-training . . . 70
5.1.5 Xavier Initialization . . . 70

5.1.6 ReLU Aware Scaled Initialization 71
5.1.7 Layer-sequential Unit Variance 71
5.1.8 Supervised Pre-training 71
5.2 Regularization of CNN 73
5.2.1 Data Augmentation 73
5.2.2 Dropout 75
5.2.3 Drop-connect 75
5.2.4 Batch Normalization 76
5.2.5 Ensemble Model Averaging 77
5.2.6 The ℓ^2 Regularization 77
5.2.7 The ℓ^1 Regularization 78
5.2.8 Elastic Net Regularization 78
5.2.9 Max-norm Constraints 79
5.2.10 Early Stopping 79
5.3 Gradient-based CNN Learning 80
5.3.1 Batch Gradient Descent 80
5.3.2 Stochastic Gradient Descent 81
5.3.3 Mini-batch Gradient Descent 81
5.4 Neural Network Optimizers 81
5.4.1 Momentum 81
5.4.2 Nesterov Momentum 82
5.4.3 Adaptive Gradient 83
5.4.4 Adaptive Delta 84
5.4.5 RMSprop 85
5.4.6 Adaptive Moment Estimation 85
5.5 Gradient Computation in CNNs 87
5.5.1 Analytical Differentiation 87
5.5.2 Numerical Differentiation 87
5.5.3 Symbolic Differentiation 88
5.5.4 Automatic Differentiation 89
5.6 Understanding CNN through Visualization 92
5.6.1 Visualizing Learned Weights 92
5.6.2 Visualizing Activations 93
5.6.3 Visualizations based on Gradients 95

6 Examples of CNN Architectures . . . 101

6.1 LeNet . . . 101
6.2 AlexNet . . . 102
6.3 Network in Network . . . 103
6.4 VGGnet . . . 104
6.5 GoogleNet . . . 104
6.6 ResNet . . . 106
6.7 ResNeXt . . . 108
6.8 FractalNet . . . 112
6.9 DenseNet . . . 114

7 Applications of CNNs in Computer Vision . . . 117

7.1 Image Classification . . . 117
7.1.1 PointNet . . . 117
7.2 Object Detection and Localization . . . 120
7.2.1 Region-based CNN . . . 120
7.2.2 Fast R-CNN . . . 122
7.2.3 Regional Proposal Network (RPN) . . . 123
7.3 Semantic Segmentation . . . 127
7.3.1 Fully Convolutional Network (FCN) . . . 127
7.3.2 Deep Deconvolution Network (DDN) . . . 130
7.3.3 DeepLab . . . 133
7.4 Scene Understanding . . . 135
7.4.1 DeepContext . . . 135
7.4.2 Learning Rich Features from RGB-D Images . . . 139
7.4.3 PointNet for Scene Understanding . . . 141
7.5 Image Generation . . . 141
7.5.1 Generative Adversarial Networks (GANs) . . . 142
7.5.2 Deep Convolutional Generative Adversarial Networks (DCGANs) . 145
7.5.3 Super Resolution Generative Adversarial Network (SRGAN) . . . 147
7.6 Video-based Action Recognition . . . 150
7.6.1 Action Recognition From Still Video Frames . . . 150
7.6.2 Two-stream CNNs . . . 153
7.6.3 Long-term Recurrent Convolutional Network (LRCN) . . . 155

8 Deep Learning Tools and Libraries 159

8.1 Caffe 159
8.2 TensorFlow 160
8.3 MatConvNet 160
8.4 Torch7 161
8.5 Theano 161
8.6 Keras 162
8.7 Lasagne 162
8.8 Marvin 164
8.9 Chainer 164
8.10 PyTorch 165

9 Conclusion 169

Bibliography 173

Authors' Biographies 185

Preface

The primary goal of this book is to provide a comprehensive treatment to the subject of convolutional neural networks (CNNs) from the perspective of computer vision. In this regard, this book covers basic, intermediate and well as advanced topics relating to both the theoretical and practical aspects.

This book is organized into nine chapters. The first chapter introduces the computer vision and machine learning disciplines and presents their highly relevant application domains. This sets up the platform for the main subject of this book, "Deep Learning", which is first defined towards the later part of first chapter. The second chapter serves as a background material, which presents popular hand-crafted features and classifiers which have remained popular in computer vision during the last two decades. These include feature descriptors such as Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Speeded-Up Robust Features (SURF), and classifiers such as Support Vector Machines (SVM), and Random Decision Forests (RDF).

Chapter 3 describes neural networks and covers preliminary concepts related to their architecture, basic building blocks, and learning algorithms. Chapter 4 builds on this and serves as a thorough introduction to CNN architecture. It covers its layers, including the basic ones (e.g., sub-sampling, convolution) as well as more advanced ones (e.g., pyramid pooling, spatial transform). Chapter 5 comprehensively presents techniques to learn and regularize CNN parameters. It also provides tools to visualize and understand the learned parameters.

Chapter 6 and onward are more focused on the practical aspects of CNNs. Specifically, Chapter 6 presents state-of-the-art CNN architectures that have demonstrated excellent performances on a number of vision tasks. It also provides a comparative analysis and discusses their relative pros and cons. Chapter 7 goes in further depth regarding applications of CNNs to core vision problems. For each task, it discusses a set of representative works using CNNs and reports their key ingredients for success. Chapter 8 covers popular software libraries for deep learning such as Theano, Tensorflow, Caffe, and Torch. Finally, in Chapter 9, open problems and challenges for deep learning are presented along with a succinct summary of the book.

The purpose of the book is not to provide a literature survey for the applications of CNNs in computer vision. Rather, it succinctly covers key concepts and provides a bird's eye view of recent state-of-the-art models designed for practical problems in computer vision.

Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun
January 2018

Acknowledgments

We would like to thank Gerard Medioni and Sven Dickinson, the editors of this Synthesis Lectures on Computer Vision series, for giving us an opportunity to contribute to this series. We greatly appreciate the help and support of Diane Cerra, Executive Editor at Morgan & Claypool, who managed the complete book preparation process. We are indebted to our colleagues, students, collaborators, and co-authors we worked with during our careers, who contributed to the development of our interest in this subject. We are also deeply thankful to the wider research community, whose work has led to major advancements in computer vision and machines learning, a part of which is covered in this book. More importantly, we want to express our gratitude toward the people who allowed us to use their figures or tables in some portions of this book. This book has greatly benefited from the constructive comments and appreciation by the reviewers, which helped us improve the presented content. Finally, this effort would not have been possible without the help and support from our families.

We would like to acknowledge support from Australian Research Council (ARC), whose funding and support was crucial to some of the contents of this book.

Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun
January 2018