Visual Cortex and Deep Networks

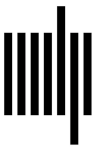
Learning Invariant Representations

By: Tomaso A. Poggio, Fabio Anselmi

Citation:

Visual Cortex and Deep Networks: Learning Invariant Representations
By: Tomaso A. Poggio, Fabio Anselmi
DOI: 10.7551/mitpress/10177.001.0001
ISBN (electronic): 9780262336710
Publisher: The MIT Press
Published: 2016

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.



The MIT Press

Preface

The ventral visual stream is believed to underlie object recognition abilities in primates. Fifty years of modeling efforts, which started with the original Hubel-Wiesel proposal of a hierarchical architecture iterating in different layers the motif of simple and complex cells in V1, led to a series of quantitative models [2, 3] that are increasingly faithful to the biological architecture and are able to mimic properties of cells in different visual areas while achieving humanlike recognition performance under restricted conditions. Deep convolutional learning networks, which are hierarchical and similar to the HMAX model but otherwise do not respect the ventral stream architecture and physiology, have been trained with very large labeled data sets [4]. The population of resulting model neurons mimics well the object recognition performance of the macaque visual cortex (DiCarlo). However, the nature of the computations carried out in the ventral stream is not explained by such models, which can be simulated on a computer but remain otherwise rather opaque.

In this book we develop a mathematical framework describing learning of invariant representations in the ventral stream. Our theory, called i-theory, applies to a broad class of hierarchical networks that pool over transformations. In particular, it applies to deep convolutional learning networks where the transformations are just translations in $\mathbb{R}^2$. The networks associated with the theory do not have nonlinearities other than for pooling. In particular, i-theory applies to networks with pooling nonlinearities (to compute a histogram or associated statistics) such as sigmoidal threshold units or linear rectifiers (see figure 1 and chapter 1).

The following chapters outline a proposal and a theory for the feedforward computation of invariant representations in the ventral stream, that is, a theory of the first 100 msec of visual perception, from the onset of an image to activation of inferior temporal (IT) cortex neurons about 100 msec later. In particular, such representations are likely to underlie rapid categorization, that is, immediate object recognition from flashed images [5, 6]. We emphasize that

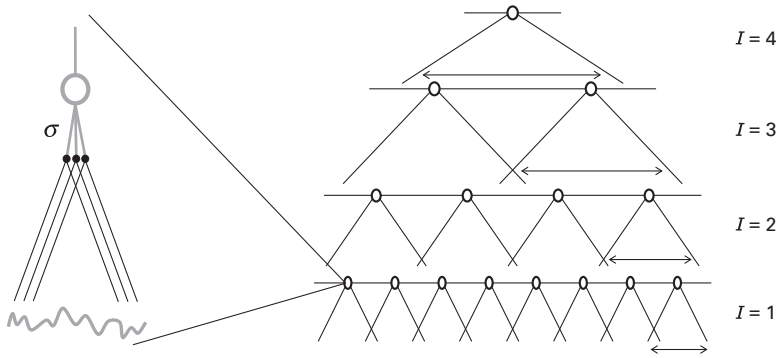


Figure 1

A hierarchy of Hubel-Wiesel modules. In each module *(inset)* the black dots represent simple cells calculating dot products of images with transformed templates (implementing a convolution) and filtering with, for example, a sigmoid function with different thresholds. The outputs of the simple cells are summed by a complex cell (pooling). This set of operations is indicated by the symbol $\bigwedge$.

the theory is not a full theory of vision that also will have to explain the top-down effect and the role of backprojections, but only part of it. The theory is based on the hypothesis that the main computational goal of the ventral stream is to compute neural representations of images that are invariant to transformations commonly encountered in the visual environment and learned from unsupervised experience. i-theory proposes computational explanations for various aspects of the ventral stream architecture and of its neurons. It makes several testable predictions. It also leads to network implementations that show high performance in object recognition benchmarks [7]. As mentioned, the theory is based on the unsupervised, automatic learning of invariant representations. Since invariant representations turn out to be good representations for supervised learning, characterized by small sample complexity, the architecture of the ventral stream may ultimately be dictated by the need to learn, at least during the initial phase of development, from very few labeled examples. We use i-theory to summarize several key aspects of the neuroscience of visual recognition, explain them, and predict others. The book is organized into four chapters, a discussion, and an appendix.

Chapter 1 describes the general theoretical framework of a computational theory of invariance. In particular, we describe relevant new mathematical results on invariant representations in vision; the appendix provides details and proofs (see also [1, 8, 10]). The starting point is a theorem proving that image

representations that are invariant to translation and scaling and approximately invariant to some other transformations (e.g., face pose and expressions) can considerably reduce the sample complexity of learning. We then describe how an invariant and unique (selective) signature can be computed for each image or image patch. The invariance can be exact in the case of locally compact group transformations (we focus on groups such as the affine group in 2-D and one of its subgroups, the similitude group consisting of translation and uniform scaling) and approximate under nongroup transformations. A module performing filtering and pooling, like the simple and complex cells described by the Hubel-Wiesel (HW) module, can compute such estimates. Each HW module provides a feature vector, which we call a signature, for the part of the visual field that is inside its receptive field. Interestingly, Gabor functions turn out to be optimal templates for maximizing simultaneous invariance to translation and scale. Hierarchies of HW modules inherit their properties, while alleviating the problem of clutter in the recognition of wholes and parts. Finally, the same HW modules at high levels in the hierarchy are shown to be able to compute representations that are approximately invariant to a much broader range of transformations, such as 3-D expression of a face, pose of a body, and viewpoint, by using templates (reflected in the neuron's tuning) that are highly specific for each object class.

Chapter 2 describes how neuronal circuits may implement the operation required by the HW algorithm. It specifically discusses new models of simple and complex cells in V1. It also introduces plausible biophysical mechanisms for tuning and pooling and for learning the wiring based on Hebbian-like unsupervised learning. The rest of the book is devoted to reviewing the application of the theory to the feedforward path of the ventral stream in the primate visual cortex.

Chapter 3 applies the theory to explain the multiresolution, eccentricity-dependent architecture of the retina and V1 as a consequence of the need for simultaneous space and scale invariance. It predicts several still untested properties of the early stages of vision. Chapter 3 also deals with V2 and V4 as higher layers in the hierarchy devoted to progressively increase invariance to shift and scaling while minimizing interference from clutter.

Chapter 4 is about the final IT stage, where class-specific representations that are quasi-invariant to nongeneric transformations are computed from a shift and scale invariant representation obtained from V4. It also discusses the modular organization of anterior IT in terms of the theory. In particular, it proposes an explanation of the architecture and of some puzzling properties of the face patches system.

Chapter 5 discusses predictions to be tested and other open problems.

The book integrates work on invariance in object recognition done in our group since 2011. The results have appeared in several different technical reports and papers available online, for which we are indebted to our collaborators Lorenzo Rosasco, Joel Leibo, Jim Mutch, Andrea Tacchetti, Qianli Liao, Leyla Isik, and Cheston Tan (see [1, 7–9, 11–15]). This is the first time we attempt to organize them in book form. The book is intended to address a broad audience from neuroscientists to mathematicians. Readers not interested in mathematical details can skip the more formal appendix without compromising understanding of the main text.

Further developments, mentioned in chapter 1 (end of section 1.7) are already under way. They will be fully explored in a planned sequel to this book.

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.

Library of Congress Cataloging-in-Publication Data

Names: Poggio, Tomaso, author. | Anselmi, Fabio, author.
Title: Visual cortex and deep networks : learning invariant representations / Tomaso A. Poggio and Fabio Anselmi.
Description: Cambridge, MA : MIT Press, [2016] | Series: Computational neuroscience | Includes bibliographical references and index.
Identifiers: LCCN 2016005774 | ISBN 9780262034722 (hardcover : alk. paper)
Subjects: LCSH: Visual cortex. | Vision. | Neural networks (Neurobiology) | Perceptual learning. | Computational neuroscience.
Classification: LCC QP383.15 .P64 2016 | DDC 612.8—dc23 LC record available at http://lccn.loc.gov/2016005774

10 9 8 7 6 5 4 3 2 1