# Visual Cortex and Deep Networks

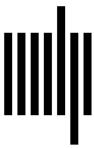
## Learning Invariant Representations

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

# 5 Discussion

## 5.1 i-Theory: Main Ideas

*Neurally plausible computation for learning invariant representations*. i-theory deals with a key computational problem in object recognition: identifying or categorizing an object after looking at a single example of it or of an exemplar of its class. To paraphrase Geman [41], the difficulty in understanding how biological organisms learn—in this case how they recognize—is not the usual $n \to \infty$ but $n \to 0$. The invariance algorithm we study is suggested by known properties of neurons, in particular their capability of performing almost automatically an approximation of the dot product of a vector of inputs with another stored vector (of synaptic weights). We introduce the notion of a *signature* of an object as a set of similarity measurements with respect to the group transformations of a small set of template images. We prove that the signature vector is invariant for the same group of transformations. This result provides a solution for the problem of recognizing an object after training with a single image of it—under conditions that are idealized but hopefully capture a good approximation of reality.

The study of this algorithm and its neural implementation leads to a number of predictions and explanations related to aspects of the ventral stream. The main ones are as follows:

- A two-stage architecture comprising a retinotopic set of areas with cell tuning that is independent of the object class and a set of nonretinotopic, class-specific modules
- Increasing receptive field (RF) sizes and invariance in the areas of the retinotopic stage
- Object-specific invariances in the class-specific modules

- Gabor-like RFs in V1 because of optimal simultaneous invariance in scale and position
- A new model of eccentricity dependence of RFs in cortical areas because of the need for position and scale invariance

## 5.2 i-Theory, Deep Learning Networks, and the Visual Cortex

Since the work on i-theory began, a class of learning algorithms called deep convolutional learning networks (DCLNs)—a new name for multilayer perceptrons (MLPs)—has achieved good performance in many applications. Historically, hardwired invariance to translation was first introduced in the Neocognitron by Fukushima [2] and later in LeNet [109] and in HMAX [47]; HMAX also had invariance to scale. These architectures are early examples of convolutional networks. There is no significant difference between i-theory and DCLNs. The basic operation in both cases is obtaining the inner product of the input $I$ with a group transformation $gt$ of a template. A pooling stage follows. Thus the basic operation in a convolutional layer is

$$\sum_g \sigma(\langle I, gt\rangle + b), \quad t \in \mathcal{T}, b \in \mathbb{R}. \tag{5.1}$$

If $\sigma$ is a step function, the pooling can compute a histogram. However, several other nonlinearities are admissible, being invariant and selective (see chapter 1 and [9]). The nonlinearity used in present-day DCLNs is a "ramp," that is, a rectification nonlinearity.

$$\sum_g |\langle I, gt\rangle + b|_+, \quad t \in \mathcal{T}, b \in \mathbb{R}. \tag{5.2}$$

Note that nonconvolutional layers are a special case of the preceding equations when the only element in $G$ is the identity. The rectification is neurally quite plausible, and it should be included in a theory of the visual cortex. As mentioned, the results of i-theory continue to hold:

- Linear combinations of rectified dot products are equivalent to kernels measuring the similarity between $I$ and $gt$. Thus the effect of rectifying nonlinearities is to define a new dot product in the feature space of the kernel.
- Group averages of nonlinear functions of the dot products are invariant and selective (no need for explicit histograms).

The *pooling* operation

$$\sum_g |\langle I, gt\rangle + b|_+, \quad t \in \mathcal{T}, b \in \mathbb{R}. \tag{5.3}$$

can also take other forms such as

$$\max_g | \langle I, gt \rangle + b|_+. \tag{5.4}$$

An alternative form of pooling, similar but biologically more plausible, is provided by softmax pooling:

$$\sum_g \frac{(\langle I, gt \rangle)^n}{\sum_{g'} (1 + \langle I, g't \rangle)^{n-1}}, \quad t \in \mathcal{T}, \tag{5.5}$$

Present-day DCNs seem to use eq. (5.5) for the pooling layers and eq. (5.2) for the nonconvolutional layers. The latter is the degenerate case of eq. (5.5) (when $G$ contains only the identity element).

## 5.3 Predictions and Explanations

From the point of view of neuroscience, i-theory provides a number of predictions and explanations. Following are questions that the theory tries to answer:

- What is visual cortex computing?
- What are the functions and circuits of simple and complex cells?
- Why is there Gabor-like tuning in simple cells?
- Why is there generic Gabor-like tuning in early areas and specific selective tuning higher up?
- What is the computational reason for the eccentricity-dependent size of RFs in V1, V2, V4?

Some predictions are the following:

- Simple and complex cells should be found in all visual and auditory areas, not only in V1. Our definitions of simple cells and complex cells are different from the traditional ones used by physiologists. In i-theory complex cells represent invariant measurements associated with statistics of the outputs of simple cells. The theory implies that under some conditions exact or approximate invariance to all geometric image transformations can be learned, either during development or in adult life. It is, however, also consistent with the possibility that basic invariances may be genetically encoded by evolution but refined and maintained by unsupervised visual experience. A single-cell model for simple and complex cells follows from the theory as an interesting possibility.

- The output of V2, V4, PIT should be capable of accessing memory either via connections that bypass higher areas or indirectly via equivalent neurons in higher areas (because of the argument about clutter).
- Areas V1, V2, V4, and possibly PIT are mainly dedicated to compute signatures that are invariant to translation, scale, and their combinations as experienced in past visual experience.
- Inferotemporal cortex is a complex of parallel class-specifc modules for a large number of large and small object classes. These modules receive position- and scale-invariant inputs (invariance in the inputs greatly facilitates unsupervised learning of class-specific transformations). From the perspective of the theory, the logothetis data [97] concern single-object modules and strongly support the prediction that exposure to a transformation leads to neuronal tuning to several frames of it.
- Each cell's tuning properties are shaped by visual experience of image transformations during developmental and adult plasticity.
- The mix of transformations (seen from the retina) learned in each area influences the tuning properties of the cells: oriented bars in V1 and V2, radial and spiral patterns in V4, class-specific tuning in anterior IT (e.g., face-tuned cells).
- During evolution areas above V1 should appear later than V1, reflecting increasing object categorization abilities and the need for invariance beyond translation.
- An architecture based on signatures that are invariant (from an area at some level) to affine transformations may underlie *perceptual constancy* against small eye movements and other small motions. There may be physiological evidence [110] suggesting invariance of several minutes of arc at the level of V1 and above.
- Invariance to affine transformations (and others) can provide the seed for evolutionary development of conceptual invariances.
- The *transfer of invariance* accomplished by the machinery of the set of transformed templates may be used to implement high-level abstractions.

## 5.4 Remarks

### Object-Based versus 3-D versus View-Based Recognition

We mention here an old controversy about whether visual recognition is based on views or on 3-D primitive shapes called geons. According to i-theory, image views retain the main role, but ideas related to 3-D shape may also be valid. The psychophysical experiments of Edelman and Bülthoff [29, 96] concluded

that generalization for rotations in depth was limited to a few degrees ($\approx \pm 30$ degrees) around a view independently of whether 2-D or 3-D information was provided to the human observer (psychophysics in monkeys [96, 97] yielded similar results). The experiments were carried out using paperclip objects with random 3-D structures (or similar but smoother object). For this type of objects class-specific learning is impossible (they do not satisfy the second condition in the class-specific theorem), and thus i-theory predicts the result obtained by Edelman and Bülthoff. For other objects, however, such as faces, the generalization that can be achieved from a single view by i-theory can span a much larger range than $\pm 30$ degrees, indirectly exploiting 3-D-like information from templates of the same class.

### Genes or Learning

i-Theory shows how the tuning of the simple cells in V1 and other areas could be learned in an unsupervised way. It is possible, however, that the tuning (or better, the ability to quickly develop it in interaction with the environment) may have been partly compiled during evolution into the genes. Note that this hypothesis implies that most of the time the specific function is not fully encoded in the genes; genes facilitate learning but do not replace it completely. It has to be expected in the nature versus nurture debate that nature usually needs nurture, and nurture is made easier by nature.

### Computational Structure of the Hubel-Wiesel Module

The HW module computes the cumulative distribution function of $\left\langle I, g_i t^k \right\rangle$ over all $g_i \in G$ or equivalent statistics. The computation consists of

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\left\langle I, g_i t^k \right\rangle + h\Delta). \tag{5.6}$$

The main forms of the nonlinearity $\sigma$ are either a threshold function or a power $n = 1, \ldots, \infty$ of its argument. Several known networks are special cases of this module. One interesting case is when $G$ is the translation group and $\sigma(\cdot) = \|\cdot\|_\infty$; then the equation is related (for $H = 0$) to a unit in a convolutional network with max pooling. In another noteworthy case (we always assume that $I$ and $t^k$ are normalized) the equation is very similar to the radial basis function network proposed by Poggio and Edelman [98] for view classification. In this spirit, note that the equation for a unit in a convolutional network is

$$\frac{1}{|G|} \sum_{i=1}^{|G|} c_i \sigma(\langle I, g_i t \rangle + h\Delta), \tag{5.7}$$

where $I$ is the input vector, $c_i, t, \Delta$ are parameters to be learned in supervised mode from labeled data, and $g_i t(x) = t(x - i\delta_x)$. Thus units in convolutional

networks are equivalent to units of the i-theory (because of weight-sharing, all $c_i$ are the same) but only when $G$ is the translation group (in i-theory $G$ is the full affine group for the first layers and can be a nongroup such as the transformation induced by rotations in depth).

### Invariance and Deep Learning Networks

i-theory provides a general theory for DLNs while offering two extensions: (1) it ensures, within the same algorithm, invariances to other groups beyond translation and in an approximate way to certain nongroup transformations, and (2) it provides a way to learn arbitrary quasi invariances from implicitely supervised learning.

### Invariance in 2-D and 3-D Vision

We assume that images as well as templates are in 2-D. This is the case if possible sources of 3-D information such as stereopsis and motion are eliminated. Interestingly, it seems that stereopsis does not facilitate recognition; this suggests that 3-D information, even when available, is not used by the human visual system (see [111]).

### Relations to the Scattering Transform

There are connections between the scattering transform and i-theory but also several differences. There is no obvious correspondence between operations in the scattering transform and simple and complex cells in the ventral stream in contrast to convolutional networks and i-theory networks. i-theory provides an algorithm in which invariances are learned from implicitly supervised experience of transformations of a random set of objects/images; in the scattering transform invariances are hardwired. i-theory proves that Gabor-like templates are optimal for simultaneous invariance to scale and shift and that such invariance requires a multiresolution inverted truncated pyramid, which turns out to be reflected in the architecture of the visual cortex, starting with the eccentricity-dependent organization of the retina; in the scattering transform, Gabor wavelets are assumed at the start.

### Explicit or Implicit Gating of Object Classes

The second stage of the recognition architecture consists of a large set of object class-specific modules of which probably the most important is the face network. It is natural to think that signals from lower areas should be gated, in order to route access only to the appropriate module. In fact, Tsao and Living Stone [112] postulated a gate mechanism for the network of face patches. The structure of the modules, however, suggests that the modules themselves provide a gating function even if their primary computational function is invariance. This is especially

clear in the case of the module associated with a single object, (the object class consists of a single object, as in the case of a paperclip). The input to the module is subject to dot products with each of the stored views of the object. If none matches well enough, the output of the module will be close to zero, effectively gating off the signal and switching off subsequent stages of processing.

### Invariance to $X$ and Estimation of $X$

Our description of i-theory focuses on the problem of recognition as estimating identity or category invariantly to a transformation ($X$) such as translation or scale or pose. Often however, the complementary problem, of estimating $X$, for instance, pose, is also important. Empirically a neural population of cells in IT is capable of supporting both representations; this multiplexed representation was shown in IT recordings [113] and model simulations [38]. As human observers, we are certainly able to estimate position, rotation, illumination of an object without eye movements. Different HW complex cells pooling over the same simple cells in different way—pooling over identities for each pose or pooling over pose for each identity—can provide the different types of information. Anselmi et al. [9, fig. 45] show simulations of recognizing a specific body invariantly to pose and estimating pose of a body invariantly to the identity.

### PCA versus ICA

Independent component analysis (ICA) [114] and similar unsupervised mechanisms describe plasticity rules similar to the basic Oja flow. They can generate Gabor-like receptive fields, and they may not need the assumption of different sizes of Gaussian distributions of lateral geniculate nucleus (LGN) synapses. We used principal component analysis (PCA) simply because its properties are easier to analyze and should be indicative of the properties of similar Hebbian-like mechanisms.

### Parsing a Scene

Full parsing of a scene cannot be done in a single feedforward processing step in the ventral stream. It requires task-dependent top-down control; in general, multiple fixations; and therefore observation times longer than $\approx 100$ msec. This follows also from the limited high-resolution region of the inverted pyramid model of the visual system, which i-theory predicts as a consequence of simultaneous invariance to shift and scale. In any case, full parsing of a scene is beyond what a purely feedforward model can provide.

### Feedforward and Feedback

We have reviewed a forward theory of recognition and some of the related evidence. i-theory does not address top-down or recurrent or horizontal

connectivity and their computational role. It makes it easier, however, to consider plausible hypotheses. The inverted pyramid architecture that follows from scale and position invariance requires for everyday vision a tight loop between different fixations in which an efficient control module drives eye movements by combining task requirements with memory access. Within a single fixation, however, the space-scale inverted pyramid cannot be shifted in space. A key function that can be controlled from the top is access to memory of lower-level signatures. For instance, a reverse hierarchy strategy may use the top-level signature to provide a high-level categorical scene interpretation and then to select appropriate routing of specific local low-level signatures to the memory and classification stage. In addition, the parameters of pooling could be controlled in a feedback mode, including the choice of which scales to use, depending on the results of classification or memory access. The most obvious limitation of feedforward architectures is recognition in clutter, and the most obvious way around the problem is the attentional masking of large parts of the image under top-down control. Further, a realistic implementation of the present theory requires top-down control signals and circuits, supervising learning, and possibly fetching signatures from different areas and at different locations in a task-dependent way. A simple idea is that signatures from all layers access the associative memory or classifier module and then control iterations in visual recognition and processing. Of course, at lower layers there are many signatures, each one in different complex cell layer locations, while at the top layer there are only a small number of signatures—in the limit only one. An even more interesting hypothesis is that backprojections update local signatures at lower levels, depending on the scene class currently detected at the top (an operation similar to the top-down pass of Borenstein and Ullman [115]). Thus the output of the feedforward pass may be used to retrieve labels and routines associated with the image. Backprojections may implement an attentional focus of processing to reduce clutter effects and also to run visual routines [38] at various levels of the hierarchy.

### Motion Helps Learning Isolated Templates

Ideally templates and their transformations should be learned without clutter. It can be argued that if the background changes between transformed images of the same template, the averaging effect intrinsic to pooling will mostly average out the effect of clutter during the unsupervised learning stage. Although this is correct, and we have computer simulations that provide empirical support to the argument, it is interesting to speculate that motion could provide a simple way to eliminate most of the clutter. Sensitivity to motion is one of the earliest visual computations to appear in the course of evolution and one of the

most primitive. Stationary images on the retina tend to fade away. Detection of relative movement is a strong perceptual cue in primate vision as well as in insect vision, probably with similar normalization-like mechanisms [116, 117]. Motion induced by the transformation of a template may then serve two important roles:

- to bind together images of the same template while transforming: continuity of motion is implicitly used to ensure that identity is preserved
- to eliminate background and clutter by effectively using relative motion between the target object and the background.

The required mechanisms are probably available in the retina and early visual cortex.

### Different Levels of Understanding

i-theory is at several different levels addressing the computational goal of the ventral stream; the algorithms used, down to the architecture of visual cortex; its hierarchical architecture; and the neural circuits underlying tuning of cells. This is unlike most other models or theories.

Finally, we should mention that some of the existing models between neuroscience and machine learning, such as HMAX [47, 50, 118] and other convolutional neural networks [2, 34, 119], are special cases of i-theory. Despite significant advances in sensory neuroscience over the last five decades, a true understanding of the basic functions of the ventral stream in visual cortex has proved elusive. Thus it is interesting that our theory follows from a novel hypothesis about the main computational function of the ventral stream: the representation of new objects/images in terms of a signature that is invariant to transformations learned during visual experience, thereby allowing recognition from very few labeled examples—in the limit, just one. This view of the cortex may also represent a novel theoretical framework for the next major challenge in learning theory beyond the supervised learning setting, which is now relatively mature: the problem of representation learning, formulated here as the unsupervised learning of invariant representations that significantly reduce the sample complexity of the supervised learning stage.

## 5.5 Ideas for Research

Following are some potential weaknesses of i-theory and some comments.

*"The theory is too good to be true."* One of the main problems of the theory is that it seems too elegant (in the sense of physics) for biology.

*Backprojections are not taken into account*, and they are an obvious feature of the cortical anatomy, which any real theory should explain. Backprojections and top-down controls are, however, implied by the present theory. Since feedforward architectures for recognition do not perform well in the presence of clutter, attentional masking may be a possible solution. Attentional focus driven by cortical backprojections will reduce clutter effects and activate visual routines at various levels of the hierarchy [38]. A Simple extension of the theory is an associative memory (storing signatures) that controls lower levels of the hierarchy via backprojections.

*Subcortical projections*, such as projections to and from the pulvinar, *are not predicted by the theory*. The present theory is still in the "cortical chauvinism" camp. Hopefully somebody will rescue it.

Cortical areas are organized in a series of layers with specific types of cells and corresponding arborizations and connectivities. *The theory does not say anything at this point about this level of the circuitry.*

Some directions for future research follow.

**Associative Architecture for Retrieval**

In past work on HMAX we assumed that hierarchical architectures perform preprocessing of an image to compute a vector (signature), which is then provided as input to a classifier. This view is extended by assuming that *signature vectors*, not only at the top of the hierarchy but at every complex cell level, are input to an associative memory. In this way a number of properties of the image (and associations) can be recalled. Parenthetically we note that old *associative memories* can be regarded as vector-valued classifiers.

- A neuron and its $10^3$–$10^4$ synapses provide the basic computational operation: a *dot product*. The input $I$ to the neuron gives the scalar $\langle I, t \rangle$ as the output, where $t$ is the vector of synaptic weights.
- A set of $K$ neurons (simple cells) computes the matrix operation $M^i f$, where

$$M^i = \begin{pmatrix} g_0 t^i \\ \cdots \\ g_K t^i \end{pmatrix}.$$

- Then the output of each complex cell $c_i$ is the average over $i$ of $|\langle M^i, I \rangle|$, which can be written as the dot product $c = e^T |\langle M^i, I \rangle|$, where $e = (1, \cdots, 1)^T$.
- The signature vector $c$ is used to access an associative memory represented as a matrix $A$. Let us denote a specific vector $c$ as $c^j$, and assume that the elements of the matrix $M$ have stored by associating the $j$ signature with a set

of properties given by the vector $p^j$: $M_{k,l} = \sum_j p_k^j c_l^j$. If $c^n$ is noiselike, then $Mc^n \sim p^n$.

### Retrieving from an Associative Memory: Optimal Sparse Encoding and Recall

There are estimates of optimal properties of codes for associative memories, including optimal sparseness (see [120, 121]). It would be interesting to connect these results to estimated capacity of visual memory [121].

### Weak Labeling by Association of Video Frames

Assume that the top associative module associates images in a video that are contiguous in time (except when there are clear transitions). This idea (Kai Yu, personal communication) relies on smoothness in time to label via association. It is biological semisupervised learning, essentially identical to a Földiák-type rule. It is thus very much in tune with our proposal of the simple complex memory-based module for learning invariances to transformations and with ideas about an associative memory module at the very top.

### Invariance and Perception

Other invariances in visual perception may be analyzed in a parallel way. An example is color constancy. Invariance to illumination (and color opponent cells) may emerge during development in a similar way as invariance to affine transformations. Thus we have a *color constancy* conjecture. The theory of chapter 1 should be able to learn invariance to illumination by observing during development transformations in the appearance of the same scene under changes of the illuminant—direction and spectral composition. A natural conjecture emerging from the approach in chapter 2 is that eigenvectors of the covariance matrix of such transformations of natural images may provide the spatial-chromatic tuning of different types of color opponent cells in V1 and other areas.

The idea that the key computational goal of the visual cortex is to learn and exploit invariances extends to other sensory modalities such as hearing of sounds and speech. It is tempting to think of music as an abstraction (in the sense of information compression and PCA) of the transformations of sounds. Classical (Western) music would then emerge from the transformations of human speech (the roots of Western classical music were based in the human voice—Gregorian chants).

### Visual Concepts

- *Concept of Parallel Lines*. Consider an architecture using signatures. Assume it has learned sets of templates that guarantee invariance to all affine

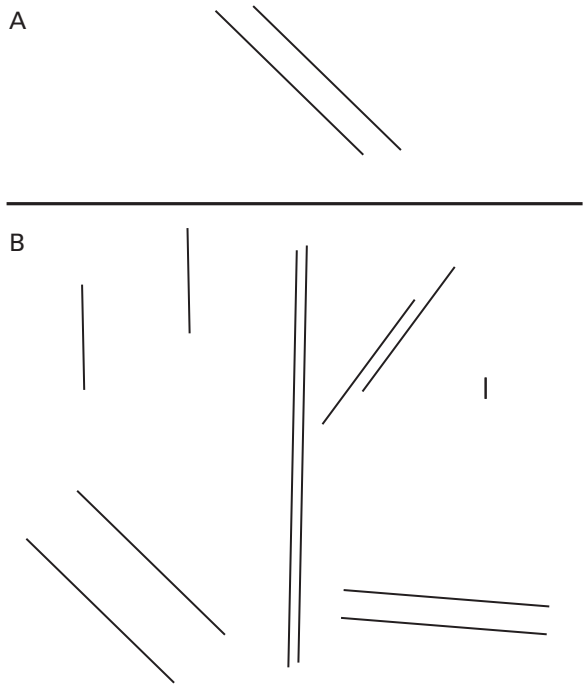



**Figure 5.1**

For a system that is invariant to affine transformations, a single training example A allows recognition of all other instances of parallel lines never seen before.

transformations. The claim is that *the architecture will appear to have learned the concept of parallel lines from a single specific example of two parallel lines* (see figure 5.1). According to the theorems presented here, the signature of the single image of the parallel lines will be invariant to affine transformations (within some range).

- *Number of items in an image*. A classifier that learns the number 5 in a way that is invariant to scale should be able to recognize five objects independent of class of objects.
- *Line drawings conjecture*. The memory-based module described here should be able to generalize from real images to line drawings when exposed to illumination-dependent transformations of images. This may need to happen at more than one level in the system, starting with the very first layer (e.g., V1). Generalizations with respect to recognition of objects invariant to shadows may also be possible.


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