Theoretical Computer Science 633 (2016) 112–121

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Unsupervised learning of invariant representations

Fabio Anselmi [a,b], Joel Z. Leibo [a], Lorenzo Rosasco [a,b,c], Jim Mutch [a], Andrea Tacchetti [a], Tomaso Poggio [a,b,*]

[a] Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[b] Istituto Italiano di Tecnologia, Laboratory for Computational and Statistical Learning, Genova, 16163, Italy
[c] Universita degli studi di Genova, Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi, Italy

ARTICLE INFO

Article history:
Received 3 December 2014
Received in revised form 6 April 2015
Accepted 22 June 2015
Available online 25 June 2015

Keywords:
Invariance
Cortex
Hierarchy
Convolutional networks

ABSTRACT

The present phase of Machine Learning is characterized by supervised learning algorithms relying on large sets of labeled examples ($n \to \infty$). The next phase is likely to focus on algorithms capable of learning from very few labeled examples ($n \to 1$), like humans seem able to do. We propose an approach to this problem and describe the underlying theory, based on the unsupervised, automatic learning of a "good" representation for supervised learning, characterized by small sample complexity. We consider the case of visual object recognition, though the theory also applies to other domains like speech. The starting point is the conjecture, proved in specific cases, that image representations which are invariant to translation, scaling and other transformations can considerably reduce the sample complexity of learning. We prove that an invariant and selective signature can be computed for each image or image patch: the invariance can be exact in the case of group transformations and approximate under non-group transformations. A module performing filtering and pooling, like the simple and complex cells described by Hubel and Wiesel, can compute such signature. The theory offers novel unsupervised learning algorithms for "deep" architectures for image and speech recognition. We conjecture that the main computational goal of the ventral stream of visual cortex is to provide a hierarchical representation of new objects/images which is invariant to transformations, stable, and selective for recognition—and show how this representation may be continuously learned in an unsupervised way during development and visual experience.

1. Introduction

It is known that Hubel and Wiesel's original proposal [1] for visual area V1—of a module consisting of complex cells (C-units) combining the outputs of sets of simple cells (S-units) with identical orientation preferences but differing retinal positions—can be used to construct translation-invariant detectors. This is the insight underlying many networks for visual recognition, including HMAX [2] and convolutional neural nets [3,4]. We show here how the original idea can be developed into a comprehensive theory of visual recognition that is relevant for computer vision and possibly for the visual cortex.

The first step in the theory is the conjecture that a representation of images and image patches, with a feature vector that is invariant to a broad range of transformations—such as translation, scale, viewpoint, pose of a body and expression of a face—makes it possible to recognize objects from only a few labeled examples. The second step is proving that hierar-

* Corresponding author at: 46-5177B, 43 Vassar Street, Cambridge, MA 02139, United States. Tel.: +1 617 253 5230.
E-mail address: tp@ai.mit.edu (T. Poggio).

http://dx.doi.org/10.1016/j.tcs.2015.06.048
0304-3975/

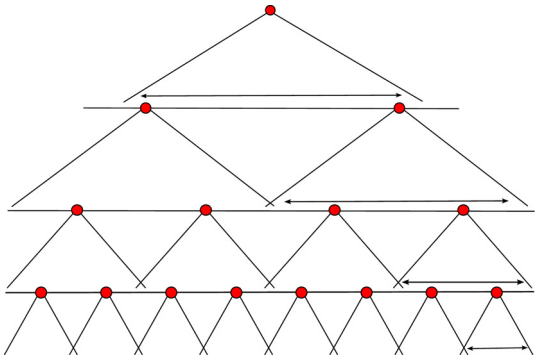

Fig. 1. A hierarchical architecture built from HW-modules. Each red circle represents the signature vector computed by the associated module (the outputs of complex cells) and double arrows represent its receptive fields—the part of the (neural) image visible to the module (for translations this is also the pooling range). The "image" is at level 0, at the bottom. The vector computed at the top of the hierarchy consists of invariant features for the whole image and is usually fed as input to a supervised learning machine such as a classifier; in addition signatures from modules at intermediate layers may also be inputs to classifiers for objects and parts.

chical architectures of Hubel–Wiesel ('HW') modules (indicated by $\bigwedge$ in Fig. 1) can provide such invariant representations while maintaining selective information about the original image. Each $\bigwedge$-module provides a feature vector, which we call a *signature*, for the part of the visual field that is inside its "receptive field". The signature is invariant to 2D affine transformations within its receptive field. The hierarchical architecture, since it computes a set of signatures for different parts of the image, is proven to be invariant to a rather general family of locally affine transformations, including (globally) affine transformations.

2. Invariant representations and sample complexity

One could argue that the most important aspect of intelligence is the ability to learn. How do present supervised learning algorithms compare with brains? One of the most obvious differences is the ability of people and animals to learn from very few labeled examples. A child, or a monkey, can learn a recognition task from just a few examples. The main motivation of this paper is the conjecture that the key to reducing the sample complexity of object recognition is invariance to transformations. Images of the same object usually differ from each other because of simple transformations such as translation, scale (distance) or more complex deformations such as viewpoint (rotation in depth) or change in pose (of a body) or expression (of a face).

The conjecture is supported by previous theoretical work showing that *almost all the complexity* in recognition tasks is often due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the object [5]. It implies that in many cases, recognition—i.e. both identification, e.g. of a specific car relative to other cars—as well as categorization, e.g. distinguishing between cars and airplanes—would require fewer examples *if* the images of objects were "rectified" with respect to all transformations, or equivalently, if the image representation itself was invariant. The conjecture is proved, using a dimensionality reduction argument, for the special case of translation (and any Abelian group–see [6] for more details):

Sample complexity for translation invariance

Consider a space of images of dimensions $p \times p$ which may appear in any position within a window of size $rp \times rp$. The natural image representation yields a sample complexity (for a linear classifier) of order $m_{image} = O(r^2 p^2)$; the invariant representation yields a sample complexity of order $m_{inv} = O(p^2)$.

The case of identification is obvious since the difficulty in recognizing exactly the same object, e.g. an individual face, is only due to transformations. In the case of categorization, consider the suggestive evidence from the classification task in Fig. 2. The figure shows that if an oracle factors out all transformations in images of many different cars and airplanes, providing "rectified" images with respect to viewpoint, illumination, position and scale, the problem of categorizing cars vs airplanes becomes easy: it can be done accurately with very few labeled examples. In this case, good performance can be obtained from a single training image of each class, using a simple classifier. In other words, the sample complexity of the problem seems to be very low. We propose that the ventral stream in visual cortex tries to approximate such an oracle, providing a quasi-invariant signature for images and image patches.

Note that this does not amount to a claim that all vision tasks demand, or would even benefit from, invariance to geometric transformations. Of course some tasks require signatures that are selective for (say) pose, but invariant to identity. However, in those cases, the computational problem is considerably easier since resemblance in the input space matches much more closely the desired outcome.