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 xxxx

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/© 2015 Elsevier B.V. All rights reserved.

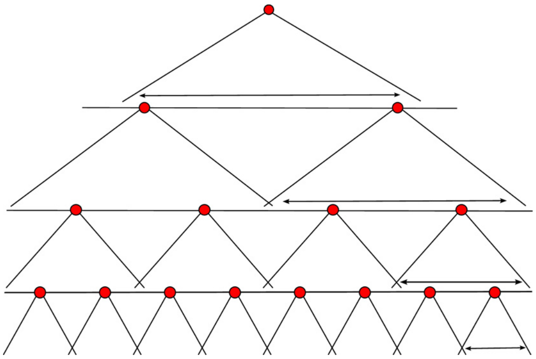


Fig. 1. A hierarchical architecture built from HW-modules. Each red circle represents the signature vector computed by the associated module (the outputs of complex cells) and double arrows represent its receptive fields—the part of the (neural) image visible to the module (for translations this is also 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.

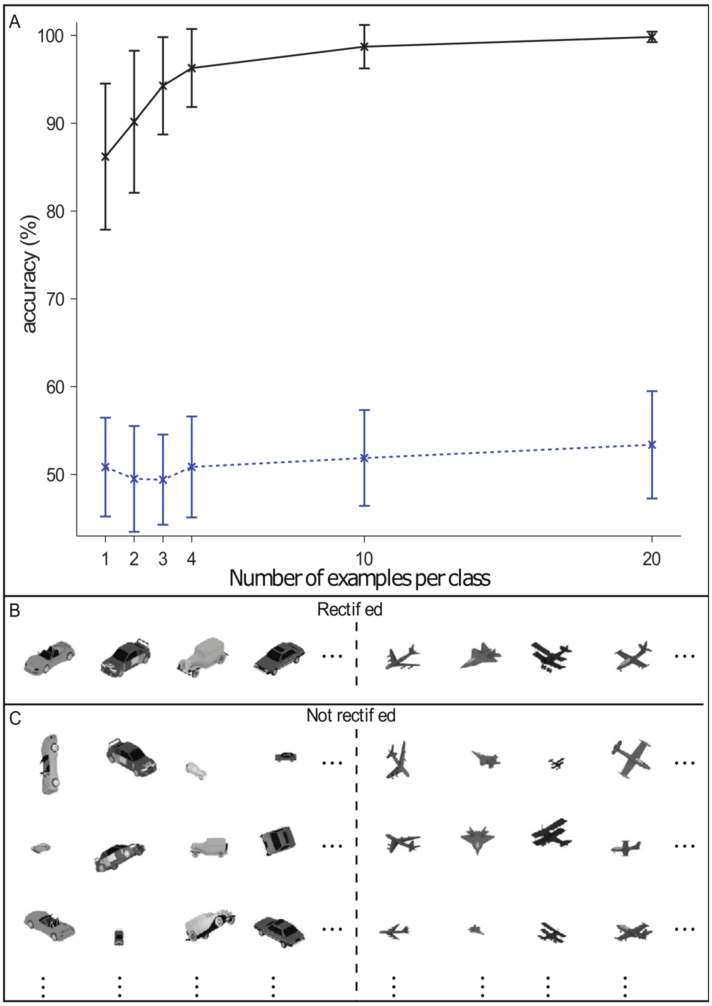


Fig. 2. Sample complexity for the task of categorizing cars vs airplanes from their raw pixel representations (no preprocessing). (A) Performance of a nearest-neighbor classifier (distance metric $= 1 -$ correlation) as a function of the number of examples per class used for training. Each test used 74 randomly chosen images to evaluate the classifier. Error bars represent ±1 standard deviation computed over 100 training/testing splits using different images out of the full set of 440 objects × number of transformation conditions. Solid line: The rectified task. Classifier performance for the case where all training and test images are rectified with respect to all transformations; example images shown in B. Dashed line: The unrectified task. Classifier performance for the case where variation in position, scale, direction of illumination, and rotation around any axis (including rotation in depth) is allowed; example images shown in C. The images were created using 3D models from the Digimation model bank and rendered with Blender.

3. Invariance and selectivity

Consider the problem of recognizing an image, or an image patch, independently of whether it has been transformed by the action of a group like the affine group in $\mathbb{R}^2$. We would like to associate to each object/image I a *signature*, i.e. a vector which is *selective* and *invariant* with respect to a group of transformations G. Note that our analysis, as we will see later, is not restricted to the case of group transformations. For now, we consider groups that are compact and, for simplicity, finite (of cardinality $|G|$). We indicate, with slight abuse of notation, a generic group element and its (unitary) representation with the same symbol g, and its action on an image as $gI(x) = I(g^{-1}x)$ (e.g. a translation, $g_\xi I(x) = I(x - \xi)$). A natural mathematical object to consider is the *orbit* O_I—the set of images gI generated from a single image I under the action of the group. We say that two images are equivalent when they belong to the same orbit: $I \sim I'$ if $\exists g \in G$ such that $I' = gI$. This equivalence relation formalizes the idea that an orbit is invariant and selective. Indeed, if two orbits have a point in common they are identical everywhere. Conversely, two orbits are different if none of the images in one orbit coincide with any image in the other [7].

How can two orbits be characterized and compared? There are several possible approaches. A distance between orbits can be defined in terms of a metric on images, but its computation is not obvious (especially by neurons). We follow here a different strategy: intuitively two empirical orbits are the same irrespective of the ordering of their points. This suggests considering the probability distribution P_I induced by the group's action on an image I (gI can be seen as a realization of a random variable; to have an intuition we can think to P_I as the pixels grey level distribution of the image over the transformations g. The distribution of the transformations is assumed to be uniform, the uniform Haar measure over the group). It is possible to prove (see [6] for further details) that if two orbits coincide then their associated distributions under the group G are identical, that is

$$I \sim I' \iff O_I = O_{I'} \iff P_I = P_{I'}. \tag{1}$$

The distribution P_I is thus invariant and selective, but it also inhabits a high-dimensional space and is therefore difficult to estimate. In particular, it is unclear how neurons or neuron-like elements could estimate it.

As argued later, neurons can effectively implement high-dimensional inner products, $\langle \cdot, \cdot \rangle$, between inputs and stored "templates" which are neural images, followed by a non-linear operation (e.g. a threshold sigmoid) and a pooling operation. The results proven in [6] say (informally) that an invariant and selective signature of an image I can be obtained as

$$\mu_h^k(I) = \frac{1}{|G|}\sum_{i=1}^{|G|} \eta_h\big(\langle I, g_i t^k \rangle\big) \tag{2}$$

where η_h, $h = 1, \cdots, H$ is a set of nonlinear functions, t^k, $k = 1, \cdots, K$ is a set of randomly chosen images called templates and we suppose G a discrete finite group. We call $\vec{\mu}(I) \in \mathbb{R}^{HK}$ the signature of image I and it can be proven to be a proxy of the probability distribution P_I. In particular it has the following properties:

Invariance theorem. *The distributions represented by equation (2) are invariant, that is for each h, k*

$$\mu_h^k(I) = \mu_h^k(gI) \tag{3}$$

for any g in G, where G is the (compact) group of transformations labeled g_i in equation (1).

The result follows simply observing that, for fixed h, k, I and t^k, $\eta_h(\langle I, gt^k \rangle) \equiv f(g)$ is a function on the group and Eq. (2) is the group average over f.

The signature in Eq. (2) is also selective since it is a proxy of the probability distribution P_I. This result follows in two steps: 1) Eq. (2) gives a proxy of the probability distribution of $\langle I, gt^k \rangle$ (the situation is particularly simple when the nonlinear functions are indicator functions of width Δ centered in h; in this case μ_h^k is a bin of the histogram of the distribution of $\langle I, gt^k \rangle$) 2) the theorem below (based on the Cramer–Wold theorem, [8], see [6] for further details) assures that a probability distribution P_I can be almost uniquely characterized by K one-dimensional probability distributions $P_{\langle I, t^k \rangle}$ induced by the results of projections $\langle I, t^k \rangle$. More precisely we have:

Selectivity theorem. *For (compact) groups of transformations, the distributions represented by equation (1) can achieve any desired selectivity for an image among N images, in the sense that they can ϵ-approximate the true distance between each pair with probability $1 - \delta$, provided that*

$$K > \frac{c}{\epsilon^2} \ln \frac{N}{\delta} \tag{4}$$

where c is a universal constant.

Thus, selectivity could be achieved (up to ϵ) via empirical proxy of the one-dimensional distribution $P_{\langle I, t^k \rangle}$ of projections of the image onto a finite number of templates t^k, $k = 1, \ldots, K$ under the action of the group. Note that number K of projection is in general infinite. A probability function in d variables (the image dimensionality) induces a unique set of 1D projections which is selective; empirically a small number of projections is usually sufficient to discriminate among a finite number of different probability distributions. Note that the bound of the number of templates in (4) is very general. A better bound can be obtained if we restrict to a the set of specific images.

4. Memory-based learning of invariance

Notice that the computation of a proxy of $P_{\langle I, t^k \rangle}$ requires the observation of the image *and* "all" its transforms gI. Ideally, however, we would like to compute an invariant signature for a new object seen only once. For example, we can recognize a new face at different distances after just one observation. It is remarkable that this is also made possible by the projection step. The key is the observation that $\langle gI, t^k \rangle = \langle I, g^{-1} t^k \rangle$ (this is true in the case of unitary groups; however any differentiable transformation can be turned unitary by dividing by the modulus of the determinant of its Jacobian, which, in the most generic case, will be a function of the spatial coordinates). The same one-dimensional distribution is obtained from the projections of the image and all its transformations onto a fixed template, as from the projections of the image onto all the transformations of the same template. Indeed, the distributions of the variables $\langle I, g^{-1} t^k \rangle$ and $\langle gI, t^k \rangle$ are the same.

Thus it is possible for the system to store for each template t^k all its transformations gt^k for all $g \in G$ and later obtain an invariant signature for new images without any explicit knowledge of the transformations g or of the group to which they belong. *Implicit knowledge of the transformations*, in the form of the stored transformed templates, allows the system to be *automatically invariant to those transformations for new inputs.*

Finally note that a visual system need not recover the actual probabilities from the empirical proxy μ_n^k in order to compute a selective signature. The set of $\mu_h^k(I)$ values is sufficient, since it identifies the associated orbit. Crucially, mechanisms capable of computing invariant representations under affine transformations for future objects can be learned and maintained in an unsupervised, automatic way by storing and updating sets of transformed templates which are *unrelated to those future objects*.

5. A theory of pooling

The above argument requires an effective normalization of the elements of the inner product (e.g. $\langle I, g_i t^k\rangle \mapsto \frac{\langle I, g_i t^k\rangle}{\|I\|\|g_i t^k\|}$) for the property $\langle gI, t^k\rangle = \langle I, g^{-1}t^k\rangle$ to be valid. Notice that invariant signatures can be computed in several ways from one-dimensional probability distributions. Instead of the $\mu_h^k(I)$ components directly representing the empirical distribution, the moments $m_h^k(I) = 1/|G| \sum_{i=1}^{|G|} (\langle I, g_i t^k\rangle)^h$ of the same distribution can be used [9] (this corresponds to the choice $\eta_h(\cdot) \equiv (\cdot)^h$). Under weak conditions, the set of *all* moments uniquely characterizes the one-dimensional distribution $P_{\langle I, t^k\rangle}$ (and thus P_I). $h = 1$ corresponds to pooling via sum/average (and is the only pooling function that does not require a nonlinearity); $h = 2$ corresponds to "energy models" of complex cells (the models have its origin in the observation that receptive fields of adjacent cells are often quadrature pairs and their sum gives a measure of the motion energy [10]) and $h = \infty$ is related to max-pooling ($lim_{h\to\infty}(m_h^k(I))^{1/h}$). In our simulations, just one of these moments usually provides sufficient selectivity to a hierarchical architecture. Other nonlinearities are also possible [11]. The arguments of this section begin to provide a theoretical understanding of "pooling", giving insight into the search for the "best" choice in any particular setting—something which is normally done empirically (e.g. [12]). According to this theory, these different pooling functions are all invariant, each one capturing part of the full information contained in the PDFs.

6. Implementations

The theory has strong empirical support from several specific implementations which have been shown to perform well on a number of databases of natural images. Support is provided by HMAX, an architecture in which pooling is done with a max operation and invariance to translation and scale is mostly hardwired, though it could also be learned. Its performance on a variety of tasks is discussed in [6] (see also [13–17]). Good performance is also achieved by other very similar architectures [18]. This class of existing models inspired the present theory, and may now be seen as special cases of it. Using the principles of invariant recognition the theory makes explicit, it is possible to develop models that incorporate invariance to more complex transformations that cannot be solved by the architecture of the network and thus must be learned from examples of objects undergoing transformations. These include non-affine and even non-group transformations, allowed by the hierarchical extension of the theory (see below). Performance for one such model is shown in Fig. 3 (see caption for details).

7. Extensions of the theory

7.1. Invariance implies localization and sparsity

The core of the theory applies without qualification to compact groups such as rotations of the image in the image plane or 3D rotations of 3D objects in 3D space. Translation and scaling are however only locally compact, and in any case, each of the modules of Fig. 1 observes only a part of the transformation's full range. Each $\bigwedge$-module has a finite pooling range, corresponding to a finite "window" over the orbit associated with an image. *Exact invariance* for each module, in the case of translations or scaling transformations, is equivalent to a condition of *localization/sparsity* of the dot product between image and template (see [6] for details). In the simple case of the translation group in one dimension the condition is (for simplicity I and t have support center in zero; a similar condition can be written for the scale group in one dimension: in this last case for the condition to make sense we require the image to be bandpass):

$$\langle I, g_x t^k\rangle = 0 \quad |x| > a. \tag{5}$$

Since this condition is a form of sparsity of the generic image I w.r.t. a dictionary of templates t^k (under a group), this result may provide a computational justification for *sparse* encoding in sensory cortex [19]. Strictly speaking the condition is valid when the object has localized support in the pooling region (is an isolated object); however it holds approximately whenever $\langle I, g_x t\rangle$ has a fast decay with the transformation (e.g. wavelet coefficients).

It turns out that localization yields the following surprising result (see [6] for further details):

Please cite this article in press as: F. Anselmi et al., Unsupervised learning of invariant representations, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.06.048

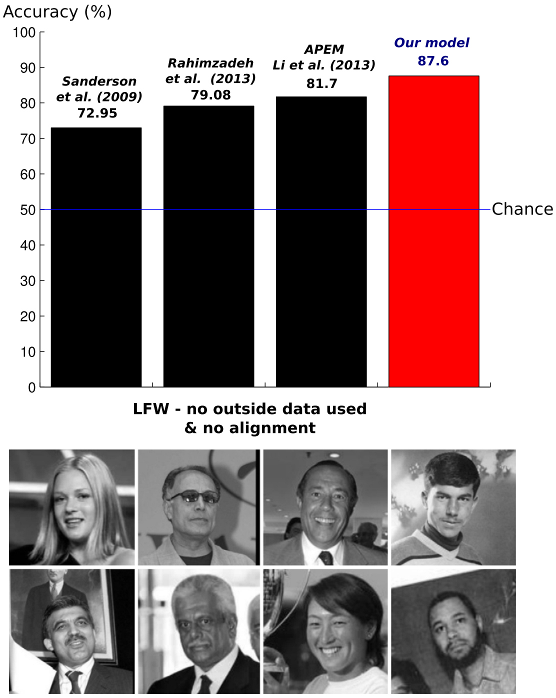


Fig. 3. Performance of a recent model [15] (inspired by the present theory) on Labeled Faces in the Wild, a same/different person task for faces seen in different poses and in the presence of clutter. A layer which builds invariance to translation, scaling, and limited in-plane rotation is followed by another which pools over variability induced by other transformations.

Optimal invariance theorem

Gabor functions of the form (here in 1D) $t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}$ are the templates that are simultaneously maximally invariant for translation and scale (at each x and ω).

Since a frame of Gabor wavelets follows from natural requirements of completeness, this may be related to the choice of wavelets for the Scattering Transform approach of Mallat based on wavelets [20].

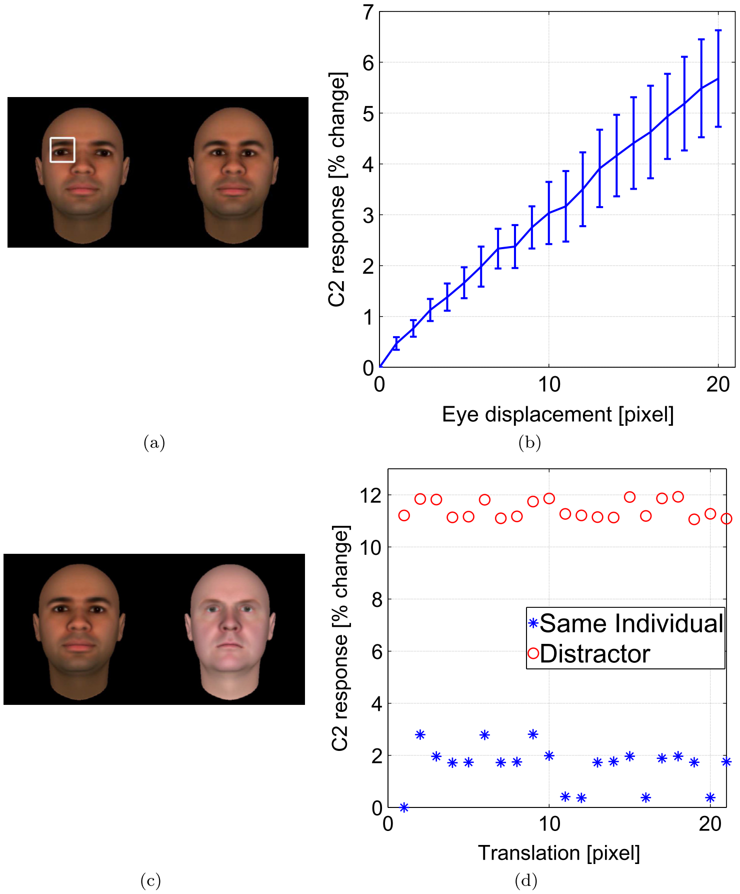
A similar equation (5), if relaxed to hold approximately becomes a *sparsity condition for the class of images $I \in C \subseteq R^d$ w.r.t. the dictionary t^k under the group G* when restricted to a subclass C of similar images. This property (which is related to the notion of "incoherence" in compressive sensing [21]) requires that I and t^k have a representation with sharply peaked correlation and autocorrelation. When the condition is satisfied, the basic HW-module equipped with such templates can provide approximate invariance to non-group transformations such as rotations in depth of a face or its changes of expression (see [22] or [6] for further details).

In summary, the localization condition can be satisfied in two different *regimes*. The first one, exact and valid for generic I, yields optimal Gabor templates. The second regime, approximate and valid for specific subclasses of C of images, yields highly tuned templates, specific for the subclass. Note that this argument suggests generic, Gabor-like templates in the first layers of the hierarchy and highly specific templates at higher levels. Note also that incoherence increases with increasing dimensionality.

7.2. Hierarchical architectures

We have focused so far on the basic HW-module. Architectures consisting of such modules can be single-layer as well as multi-layer (see Fig. 1). In our theory, the key property of hierarchical architectures of repeated HW-modules—allowing the recursive use of modules in multiple layers—is the property of *covariance*. By a covariant response at layer ℓ we mean that the distribution of the values of each projection is the same if we consider the image or the template transformations, i.e. $distr(\langle \mu_\ell(gI), \mu_\ell(t^k) \rangle) = distr(\langle \mu_\ell(I), \mu_\ell(gt^k) \rangle), \ \forall k$.

One-layer networks can achieve invariance to *global* transformations of the whole image while providing a selective global signature which is stable with respect to small perturbations of the image (see [6] and [11] for details). The three main reasons for a hierarchical architecture such as Fig. 1 are (a) the need to compute representations that are not affected by clutter, (b) the need to compute an invariant representation not only for the whole image but especially for all parts of it, which may contain objects and object parts, and (c) invariance to global transformations that are not affine, but are locally affine, that is, affine within the pooling range of some of the modules in the hierarchy (any differentiable transformation, no matter how complex, can be seen locally as an affine transformation).



(a) (b)

(c) (d)

Fig. 4. Empirical demonstration of the properties of invariance, stability and discriminability of the hierarchical architecture in a specific 2 layer implementation (HMAX). (a) Shows the reference image on the left and a deformation of it (the eyes are closer to each other) on the right; (b) shows the relative change in signature provided by 128 HW-modules at layer 2 (C_2) whose receptive fields contain the whole face. This signature vector is (Lipschitz) stable with respect to the deformation. Error bars represent ±1 standard deviation. Two different images (c) are presented at various location in the visual field. (d) Shows the relative change of the signature vector for different values of translation. The signature vector is invariant to global translation and discriminative between the two faces. In this example the HW-module represents the top of a hierarchical, convolutional architecture. The images were 200 × 200 pixels.

Of course, one could imagine local and global one-layer architectures used in the same visual system without a hierarchical configuration, but there are further reasons favoring hierarchies including compositionality and reusability of parts. In addition to the issues of clutter, sample complexity and connectivity, one-stage architectures are unable to capture the hierarchical organization of the visual world where scenes are composed of objects which are themselves composed of parts. Objects can move in a scene relative to each other without changing their identity and often changing the scene only in a minor way. The same is often true for parts within an object. Thus global and local signatures from all levels of the hierarchy must be able to access memory in order to enable the categorization and identification of whole scenes as well as of patches of the image corresponding to objects and their parts. Fig. 4 show examples of invariance and stability for wholes and parts. In the architecture of Fig. 1, each $\bigwedge$-module provides discriminability, invariance and stability at different levels, over increasing ranges from bottom to top. Thus these architectures match the hierarchical structure of the visual world and enable retrieval of items from memory at various levels of size and complexity. These results are part of a general theory of hierarchical architectures which is beginning to take form (see [11,20,23,24]) around the basic function of computing invariant representations. The property of compositionality discussed above is related to the efficacy of hierarchical architectures vs. one-layer architectures in dealing with the problem of partial occlusion and the more difficult problem of clutter in object recognition. Hierarchical architectures are better at recognition in clutter than one-layer networks [25] because they provide signatures for image patches of several sizes and locations. However, hierarchical feedforward architectures cannot fully solve the problem of clutter. More complex (e.g. recurrent) architectures are likely needed for human-level recognition in clutter (see for instance [26–28]) and for other aspects of human vision. It is likely that much of the circuitry of visual cortex is required by these recurrent computations, not considered in this paper.

Please cite this article in press as: F. Anselmi et al., Unsupervised learning of invariant representations, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.06.048

8. Visual cortex

The theory described above effectively maps the computation of an invariant signature onto well-known capabilities of cortical neurons. A key difference between transistors—the basic components of our digital computers—and neurons is the number of connections: 3 wires vs. 10^3–10^4 synapses per cortical neuron. Taking into account basic properties of synapses, it follows that a single neuron can compute high-dimensional (10^3–10^4) inner products between input vectors and the stored vector of synaptic weights [29].

Consider an HW-module of "simple" and "complex" cells [1] looking at the image through a window defined by their receptive fields. Suppose that images of objects in the visual environment undergo affine transformations. During development—and more generally, during visual experience—a set of $|G|$ simple cells store in their synapses an image patch t^k and its transformations $g_1 t^k, \cdots, g_{|G|} t^k$–one per simple cell. This is done, possibly at separate times, for K different image patches t^k (templates), $k = 1, \cdots, K$. Each gt^k for $g \in G$ is a sequence of frames, literally a movie of an image patch t^k transforming. There is a very *simple, general, and powerful way to learn* such unconstrained transformations. Unsupervised learning is the main mechanism: for a "complex" cell to pool over several simple cells, the key is a modified Hebbian rule based on temporal association (Foldiak-type rule, [30]): *cells that fire in temporal contiguity are wired together*. At the level of complex cells this rule determines *classes of equivalence* among simple cells—reflecting observed *time correlations in the real world, that is, transformations* of the image. Time continuity allows associative labeling of stimuli based on their temporal contiguity.

Later, when an image is presented, the simple cells compute $\langle I, g_i t^k \rangle$ for $i = 1, \ldots, |G|$. The next step, as described above, is to give a proxy of the one-dimensional probability distribution of such a projection, that is, the distribution of the outputs of the simple cells. It is generally assumed that complex cells pool the outputs of simple cells. Thus a complex cell could compute

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma\left(\langle I, g_i t^k \rangle + h\Delta\right) \qquad (6)$$

where σ is a smooth version of the step function ($\sigma(x) = 0$ for $x \leq 0$, $\sigma(x) = 1$ for $x > 0$) and $h = 1, \ldots, H$ (this corresponds to the choice $\eta_h(\cdot) \equiv \sigma(\cdot + h\Delta)$). Each of these H complex cells would estimate one bin of an approximated CDF (cumulative distribution function) for $P_{\langle I, t^k \rangle}$. Since a distribution can be represented exactly by all its moments and approximatively (in general) by a few of them, the complex cells could compute, instead of an empirical CDF, one or more of its moments, $\eta_h(\cdot) \equiv (\cdot)^h$; as explained before $h = 1$ is the mean, $h = 2$ corresponds to an energy model of complex cells; very large h is related to a *max* operation. Conventional wisdom mostly interprets the available physiological data to suggest that simple/complex cells in V1 may be described in terms of energy models. Our theory suggests that in addition to energy models some of the complex cells may represent other moments of the distribution or even different linear combinations of them. A number of other models are clearly allowed by the theory since many functions that depend on the CDF—such as a linear combination of its moments—are invariant; a sufficient set of them can contain sufficient information for discrimination.

As described above, a template and its transformed versions may be learned from unsupervised visual experience through Hebbian plasticity. Hebbian plasticity, as formalized by Oja, can yield *Gabor-like tuning*—i.e. the templates that provide optimal invariance to translation and scale, since the statistics of natural images is translation invariant and approximately scale invariant. Remarkably, our analysis and empirical studies [11] find that quantitative properties of the associated Gabor-like tuning fits experimental data in different species.

The localization condition (Equation (5)) can also be satisfied by images and templates that are similar to each other. The result is invariance to class-specific transformations. This part of the theory is consistent with the existence of class-specific modules in primate cortex such as a face module and a body module [31,32,16]. It is intriguing that *the same localization condition* suggests *general Gabor-like templates for generic images* in the first layers of a hierarchical architecture and *specific, sharply tuned templates* for the last stages of the hierarchy. This theory also fits physiology data concerning Gabor-like tuning in V1 and possibly in V4 (see [11]). It can also be shown that the theory, together with the hypothesis that storage of the templates takes place via Hebbian synapses, also predicts properties of the tuning of neurons in the face patch AL of macaque visual cortex [11,17].

From the point of view of neuroscience, the theory makes a number of predictions. One is that the machinery implementing selectivity and invariance should be similar across all visual and auditory areas. One more speculative prediction concerns complex cell responses: they may correspond to invariant measurements associated with histograms of the outputs of simple cells or of moments of their response distribution. Note also that this neural interpretation is also valid if "simple cells" are dendritic compartments rather than cells themselves. 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 and possibly refined and maintained by unsupervised visual experience.

9. Discussion

The goal of this paper is to introduce a new theory of learning invariant representations for object recognition which cuts across levels of analysis [11,33]. Some of the existing models between neuroscience and machine learning, such as HMAX [2,34,35] and Convolutional Neural Networks [3,4,36,37], are special and limited cases of the 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 proven elusive. Thus it is interesting that the theory of this paper 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. A main contribution of our work to machine learning is a novel theoretical framework for the next major challenge in learning theory beyond supervised learning: the problem of *representation learning*.

Acknowledgements

We would like to thank the McGovern Institute for Brain Research for their support. We would also like to thank for having read earlier versions of the manuscript Yann LeCun, Ethan Meyers, Andrew Ng, Bernhard Schoelkopf and Alain Yuille. We also thank Michael Buice, Charles Cadieu, Robert Desimone, Leyla Isik, Gabriel Kreiman, Lakshminarayanan Mahadevan, Stephane Mallat, Pietro Perona, Maximilian Riesenhuber, Ryan Rifkin, Terrence J. Sejnowski, Steve Smale, Haim Sompolinsky for useful comments. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by the Eugene McDermott Foundation. At the time of our writing, the working monograph [11] contains the most up-to-date account of the theory. The current monograph evolved from one that first appeared in July 2011 [11].

References

[1] D. Hubel, T. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106, http://jp.physoc.org/content/160/1/106.full.pdf.

[2] M. Riesenhuber, T. Poggio, Models of object recognition, Nat. Neurosci. 3 (11) (2000) 1199–1204, http://www.nature.com/neuro/journal/v3/n11s/full/nn1100_1199.html.

[3] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybernet. 36 (4) (1980) 193–202, http://www.springerlink.com/content/r6g5w3tt54528137.

[4] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, http://www.mitpressjournals.org/doi/abs/10.1162/neco.1989.1.4.541.

[5] T. Lee, S. Soatto, Video-based descriptors for object recognition, Image Vis. Comput. 29 (2012) 639–652.

[6] F. Anselmi, J. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio, Unsupervised learning of invariant representations with low sample complexity: the magic of sensory cortex or a new framework for machine learning?, Center for Brains, Minds and Machines (CBMM) Memo No. 1, arXiv:1311.4158v5.

[7] H. Schulz-Mirbach, Constructing invariant features by averaging techniques, in: Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994, Conference B: Computer Vision and Image Processing, vol. 2, 1994, pp. 387–390.

[8] H. Cramer, H. Wold, Some theorems on distribution functions, J. London Math. Soc. 4 (1936) 290–294.

[9] A. Koloydenko, Symmetric measures via moments, Bernoulli 14 (2) (2008) 362–390.

[10] E. Adelson, J. Bergen, Spatiotemporal energy models for the perception of motion, J. Opt. Soc. Amer. A 2 (2) (1985) 284–299, http://www.opticsinfobase.org/abstract.cfm?URI=josaa-2-2-284.

[11] F. Anselmi, J. Leibo, J. Rosasco, L. Mutch, A. Tacchetti, T. Poggio, Magic materials: a theory of deep hierarchical architectures for learning sensory representations, CBCL paper, http://cbcl.mit.edu/publications/ai-publications/2013/Magic_working_paper_May6_2013.pdf.

[12] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition?, in: IEEE International Conference on Computer Vision, 2009, pp. 2146–2153, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5459469.

[13] S. Voinea, C. Zhang, G. Evangelopoulos, L. Rosasco, T. Poggio, Word-level invariant representations from acoustic waveforms, in: 15th Annual Conf. of the International Speech Communication Association, INTERSPEECH 2014, Singapore, 2014.

[14] C. Zhang, S. Voinea, G. Evangelopoulos, L. Rosasco, T. Poggio, Phone classification by a hierarchy of invariant representation layers, in: 15th Annual Conf. of the International Speech Communication Association, INTERSPEECH 2014, Singapore, 2014.

[15] Q. Liao, J.Z. Leibo, T. Poggio, Learning invariant representations and applications to face verification, in: Advances in Neural Information Processing Systems, NIPS, Lake Tahoe, NV, 2013.

[16] J.Z. Leibo, J. Mutch, T. Poggio, Why the brain separates face recognition from object recognition, in: Advances in Neural Information Processing Systems, NIPS, Granada, Spain, 2011, http://cbcl.mit.edu/publications/ps/Leibo_Mutch_Poggio_face_invar_v08_cam_rdy_letter_Dec2011.pdf.

[17] J.Z. Leibo, F. Anselmi, J. Mutch, A.F. Ebihara, W. Freiwald, T. Poggio, View-invariance and mirror-symmetric tuning in a model of the macaque face-processing system, in: Computational and Systems Neuroscience, Salt Lake City, USA, 2013, p. I-54.

[18] N. Pinto, D. Doukhan, J. DiCarlo, D. Cox, A high-throughput screening approach to discovering good forms of biologically inspired visual representation, PLoS Comput. Biol. 5 (11) (2009) e1000579, http://dx.plos.org/10.1371/journal.pcbi.1000579.

[19] B. Olshausen, et al., Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (6583) (1996) 607–609.

[20] S. Mallat, Group invariant scattering, Comm. Pure Appl. Math. 65 (10) (2012) 1331–1398, http://dx.doi.org/10.1002/cpa.21413.

[21] E. Candes, J. Romberg, Sparsity and incoherence in compressive sampling, Invers. Probl. 23 (2007) 969–985.

[22] J.Z. Leibo, Q. Liao, F. Anselmi, T. Poggio, The invariance hypothesis implies domain-specific regions in visual cortex, bioRxiv http://dx.doi.org/10.1101/004473.

[23] S. Soatto, Steps towards a theory of visual information: active perception, signal-to-symbol conversion and the interplay between sensing and control, 2011, pp. 1–151, arXiv:1110.2053.

[24] S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, T. Poggio, Mathematics of the neural response, Found. Comput. Math. 10 (1) (2010) 67–91, http://www.springerlink.com/index/k650503137550j23.pdf.

Please cite this article in press as: F. Anselmi et al., Unsupervised learning of invariant representations, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.06.048

[25] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, T. Poggio, A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex, CBCL paper #259/AI memo #2005-036, http://dspace.mit.edu/handle/1721.1/36407.

[26] S.S. Chikkerur, T. Serre, C. Tan, T. Poggio, What and where: a Bayesian inference theory of attention, Vis. Res. 50 (22) (2010) 2233–2247, http://dx.doi.org/10.1016/j.visres.2010.05.013.

[27] D. George, J. Hawkins, A hierarchical bayesian model of invariant pattern recognition in the visual cortex, in: Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 3, IJCNN, 2005, pp. 1812–1817.

[28] S. Geman, Invariance and selectivity in the ventral visual pathway, J. Physiol. 100 (4) (2006) 212–224, http://linkinghub.elsevier.com/retrieve/pii/S0928425707000034.

[29] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (4) (1943) 115–133, http://link.springer.com/article/10.1007/BF02478259.

[30] P. Földiák, Learning invariance from transformation sequences, Neural Comput. 3 (2) (1991) 194–200, http://www.mitpressjournals.org/doi/abs/10.1162/neco.1991.3.2.194.

[31] N. Kanwisher, Functional specificity in the human brain: a window into the functional architecture of the mind, Proc. Natl. Acad. Sci. USA 25 (107) (2010) 11163.

[32] D. Tsao, W. Freiwald, Faces and objects in macaque cerebral cortex, Nature 6 (9) (2003) 989–995, http://www.nature.com/neuro/journal/v6/n9/abs/nn1111.html.

[33] D. Marr, T. Poggio, From understanding computation to understanding neural circuitry, AIM-357, http://scholar.google.com/scholar?q=marr+poggio+levels&hl=en&btnG=Search&as_sdt=40000001&as_sdtp=on#2.

[34] J. Mutch, D. Lowe, Multiclass object recognition with sparse, localized features, in: Comput. Vis. Pattern Recognit. 2006, vol. 1, 2006, pp. 11–18, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1640736.

[35] T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization, Proc. Natl. Acad. Sci. USA 104 (15) (2007) 6424–6429, http://cat.inist.fr/?aModele=afficheN&cpsidt=18713198.

[36] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: The Handbook of Brain Theory and Neural Networks, 1995, pp. 255–258, URL, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9297&rep=rep1&type=pdf.

[37] Y. LeCun, F. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, CVPR 2004, IEEE, 2004, p. II-97.

Please cite this article in press as: F. Anselmi et al., Unsupervised learning of invariant representations, Theoret. Comput. Sci. (2015), http://dx.doi.org/10.1016/j.tcs.2015.06.048