Visual Cortex and Deep Networks

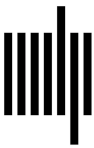
Learning Invariant Representations

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

Appendix

A.1 Invariant Representations and Bounds on Learning Rates: Sample Complexity

In this section we derive a mathematical proof of how an invariant representation can effectively reduce the sample complexity of the recognition task. In particular, we show that far fewer examples are needed to train a classifier for recognition (sample complexity) if we use an invariant representation. We start with the basic problem of learning how two sets of random variables x, y are probabilistically related. If their relation is described by the distribution $P(x, y)$, the problem is usually solved by minimizing the square loss

$$E(f) = \int (y - f(x))^2 P(x, y) dx\, dy.$$

However, the distribution $P(x, y)$ is unknown, and what can be minimized is only the empirical error of a set or random sampled points (x_i, y_i), $i = 1, \ldots, m$,

$$E_m(f) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2. \tag{A.1}$$

Let f_m be the minimizer in eq. (A.1). We would like to have that the probability that $E(f_m)$ and $E_m(f_m)$ differ significantly is very small,

$$Prob|E(f_m) - E_m(f_m) > \epsilon| < 1 - \delta, \quad \epsilon, \delta > 0. \tag{A.2}$$

This problem is related (see, e.g., [122]) to the calculation of the *covering number* of the function space (hypothesis space), $\mathcal{F}$, where the minimizer lives. More precisely, if the samples (x_i, y_i) are contained in a ball B of radius r, the covering number, $\mathcal{N}(\epsilon, B, \mathcal{F})$, is defined as the minimum number of $\epsilon-$ balls needed to cover B. It can be proved [122] that

$$\delta = 2\mathcal{N}(\mathcal{F}, \frac{\epsilon}{8}) e^{-\frac{m\epsilon^2}{8}}. \tag{A.3}$$

Inverting eq. (A.3), we have a bound on the number of samples, m, in terms of the covering number of the functional space of the possible minimizers.

More concretely, suppose the samples are taken from a sample space $\mathcal{X}$ of images, where the distribution $P(x, y)$ is defined. If, for example, we express the images in a basis, they can be viewed as vectors in $\mathbb{R}^d$ (e.g., discrete wavelet basis). The sample complexity of a learning rule depends on the covering number of the ball, $B \in \mathbb{R}^d$, that contains all the image distribution:

$$\mathcal{N}(\epsilon, B, \mathcal{X}) \sim \left(\frac{r}{\epsilon}\right)^d.$$

In the case of linear learning rules, the sample complexity is proportional to the *logarithm* of the covering number.

As a basic example, consider $\mathcal{X}$ to be d-dimensional and a hypothesis space of linear functions

$$f(I) = \langle w, I \rangle, \quad \forall I \in \mathcal{X}, w \in \mathcal{X},$$

so that the data representation is simply the identity. Then the ϵ-covering number of the set of linear functions with $\|w\| \leq 1$ is given by

$$N_\epsilon \sim \epsilon^{-d}.$$

If the input data lie in a subspace of dimension $s \leq d$, then the covering number of the space of linear functions becomes $N_\epsilon \sim \epsilon^{-s}$. In the next section we further comment on this example and provide an argument to illustrate the potential benefits of invariant representations.

A.1.1 Translation Group

Consider the simple example of a set of images of $p \times p$ pixels, and each containing an object within a (square) window of $k \times k$ pixels and surrounded by a uniform background. Imagine the object positions to be possibly anywhere in the image. Without loss of generality, fix the window centered in the middle of the image. Then it is easy to see that as soon as objects are translated so that they do not overlap, we get an orthogonal subspace. Then we see that there are $r^2 = (p/k)^2$ possible subspaces of dimension k^2, that is, the set of translated images can be seen as a distribution of vectors supported within a ball in $d = p^2$ dimensions. Following the discussion in the previous section, the best algorithm based on a linear hypothesis space will incur a sample complexity proportional to d. Assume now we have access to an oracle that can register each image so that each object occupies the centered position. In this case, the distribution of images is effectively supported within a ball in $s = k^2$

dimensions, and the sample complexity is proportional to s rather than d. In other words, a linear learning algorithm would need

$$r^2 = d/s$$

fewer examples to achieve the same accuracy. The idea is that invariant representations can act as an invariance oracle and have the same impact on the sample complexity.

A.1.2 Scale and Translation: 1-D Affine Group

In the following we consider the case of a non-Abelian locally compact discrete group of transformations G and its associated wavelet transform (e.g., $g_{j,k} \in G,\ j,k \in \mathbb{Z}$ such that $g_{j,k}f(x) = 2^{j/2}f(2^j x - k)$, $f \in L^2(\mathbb{R})$). Let $I \in L^2(\mathbb{R})$ be an image, and $\bar{G}$ be a finite subset of G. We have the following.

Lemma A.1 Let $I \in \mathcal{X}^\Omega = L^2(\Omega)$ be an image belonging to the set of square integrable functions with support in $\Omega \subseteq \mathbb{R}$ and $W : L^2(\mathbb{R}) \to \mathbb{R}^{|G|}$, the discrete wavelet transform operator, where $|\cdot|$ indicates the cardinality. Let $\mathcal{N}^{\bar{G}I}(B, \epsilon)$ (B is a ball of radius r) be the covering number of the space of functions obtained by applying the group transformations $g \in \bar{G}$ to I. Let $m_{\bar{G}I}$ be its associated sample complexity. We have

$$m_{\text{inv}} = m_{\text{image}} \frac{|G_I|}{|\bar{G}G_I|} \tag{A.4}$$

with

$$|G_I| = |\text{supp}(W(I))|,$$

where "supp" indicates the support, and W is the discrete wavelet transform operator.

Proof: Note first that eq. (A.4) is well defined, since the function I has compact support in Ω, that is, $|G_I| < \infty$, $|\bar{G}| < \infty$ by hypothesis, and $|\bar{G}G_\Omega| < |\bar{G}||G_I|$.

Given the definitions in the theorem, in the case of a linear classifier, we have

$$\mathcal{N}^I(B, \epsilon) = \left(\frac{r}{\epsilon}\right)^{|G_I|}.$$

Consider all the $\bar{G}$-group shifted versions of I. Using the covariance property of the wavelet coefficients, we have that the number of nonzero coefficients is $|\bar{G}G_I|$, and therefore the associated covering number is

$$\mathcal{N}^{\bar{G}I}(B, \epsilon) = \left(\frac{r}{\epsilon}\right)^{|\bar{G}G_I|}.$$

Thus, with the sample complexity $m \propto \ln(\mathcal{N})$, we have

$$m_{\text{image}} \propto |\bar{G}G_I|.$$

The sample complexity gain is in this case

$$\frac{m_{\text{inv}}}{m_{\text{image}}} = \frac{|G_I|}{|\bar{G}G_I|}.$$

A.2 One-Layer Architecture: Invariance and Selectivity

A.2.1 Equivalence between Orbits and Probability Distributions

The following is taken from Anselmi et al. [1].

Let $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{P}(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$. Recall that for any compact group, the Haar measure is finite, so if appropriately normalized, it corresponds to a probability measure.

In the following we assume G to be Abelian and compact and the corresponding Haar measure to be normalized. The first step in our reasoning is the following definition.

Definition A.2 Representation via Orbit Probability For all $I \in \mathcal{X}$, define the random variable

$$Z_I : (G, dg) \to \mathcal{X}, \quad Z_I(g) = gI, \quad \forall g \in G,$$

with law

$$\rho_I(A) = (Z_x)\rho(A) = \int_{Z_I^{-1}(A)} dg,$$

for all measurable sets $A \subset \mathcal{X}$. Let

$$P : \mathcal{X} \to \mathcal{P}(\mathcal{X}), \quad P(I) = \rho_I, \quad \forall I \in \mathcal{X}.$$

The map P associates to each point a corresponding probability distribution. From definition A.2 we see that we are essentially viewing an orbit as a distribution of points, and mapping each point in one such distribution. Then we have the following result.

Theorem A.3 For all $I, I' \in \mathcal{X}$,

$$I \sim I' \quad \Leftrightarrow \quad P(I) = P(I'). \tag{A.5}$$

Proof: We first prove that $I \sim I' \Rightarrow \rho_I = \rho_{I'}$. Recalling that if $\mathcal{C}_c(\mathcal{I})$ is the set of continuous functions on $\mathcal{I}$ with compact support, ρ_I can be alternatively defined as the unique probability distribution such that

$$\int f(z)d\rho_I(z) = \int f(Z_I(g))dg, \quad \forall f \in \mathcal{C}_c(\mathcal{I}). \tag{A.6}$$

Therefore $\rho_I = \rho_{I'}$ if and only if for any $f \in \mathcal{C}_c(\mathcal{I})$, we have $\int_{\mathcal{G}} f(Z_I(g))dg = \int_{\mathcal{G}} f(Z_{I'}(g))dg$, which follows immediately from a change of variable and invariance of the Haar measure:

$$\int_{\mathcal{G}} f(Z_I(g))dg = \int_{\mathcal{G}} f(gI)dg = \int_{\mathcal{G}} f(gI')dg = \int_{\mathcal{G}} f(g\tilde{g}I)dg = \int_{\mathcal{G}} f(\hat{g}I)d\hat{g}.$$

To prove that $\rho_I = \rho_{I'} \Rightarrow I \sim I'$, note that $\rho_I(A) - \rho_{I'}(A) = 0$ for all measurable sets $A \subseteq \mathcal{X}$ implies in particular that the support of the probability distributions of I has non-null intersection on a set of non-zero measure. Since the support of the distributions $\rho_I, \rho_{I'}$ is exactly the orbits associated to I, I', respectively, then the orbits coincide, that is, $I \sim I'$.

This result shows that an invariant representation can be defined considering the probability distribution naturally associated to each orbit. However, its computational realization would require dealing with high-dimensional distributions. Indeed, we next show that the representation can be further developed to consider only probability distributions on the real line.

A.2.2 Cramér-Wold Theorem and Random Projections for Probability Distributions

Simple operations for neurons are (high-dimensional) inner products, $\langle \cdot, \cdot \rangle$, between inputs and stored templates, which are neural images. In this section we show how it is possible to give a good estimate of the (high-dimensional) probability distribution described in the previous section through a collection of one-dimensional probability distributions.

Classical results such as the Cramér-Wold theorem [123], ensure 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 (one-dimensional) results of projections $\langle I, t^k \rangle$, where t^k, $k = 1, \ldots, K$ are a set of randomly chosen images called templates. A probability function in d variables (the image dimensionality) induces a unique set of 1-D projections, which is discriminative; empirically, a small number of projections is usually sufficient to discriminate among a finite number of different probability distributions. Theorem A.5 says (informally) that an approximately invariant and unique signature of an image I can be obtained from the estimates of K 1-D probability distributions

$P_{\langle I,t^k\rangle}$ for $k = 1, \ldots, K$. The number K of projections needed to discriminate n orbits, induced by n images, up to precision ϵ (and with confidence $1 - \delta^2$) is $K \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Thus the discriminability question can be answered positively (up to ϵ) in terms of empirical estimates of the one-dimensional distributions $P_{\langle I,t^k\rangle}$ of projections of the image onto a finite number of templates t^k, $k = 1, \ldots, K$, under the action of the group.

A.2.3 Finite Random Projections Almost Discriminate among Different Probability Distributions

Given the discussion in Chapter 1, a *signature* may be associated to I by constructing a histogram approximation of P_I, but this would require dealing with high-dimensional histograms. The following classic theorem gives a way around this problem.

For a *template* $t \in \mathbb{S}(\mathbb{R}^d)$, where $\mathbb{S}(\mathbb{R}^d)$ is unit sphere in $\mathbb{R}^d$, let $I \mapsto \langle I, t\rangle$ be the associated projection. Moreover, let $P_{\langle I,t\rangle}$ be the distribution associated to the random variable $g \mapsto \langle gI, t\rangle$ (or equivalently, $g \mapsto \langle I, g^{-1}t\rangle$, if g is unitary). Let $\mathcal{E} = [t \in \mathbb{S}(\mathbb{R}^d), \text{ s.t. } P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}]$.

Theorem A.4 (Cramér-Wold [123]) For any pair P, Q of probability distributions on $\mathbb{R}^d$, we have that $P = Q$ if and only if $\mathcal{E} = \mathbb{S}(\mathbb{R}^d)$.

In words, two probability distributions are equal if and only if their projections on any of the unit sphere directions are equal. This result can be equivalently stated as saying that the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 1 if and only if $P = Q$, and the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 0 if and only if $P \neq Q$ (see [123, theorem 3.4]). The theorem suggests a way to define a metric on distributions (orbits) in terms of

$$d(P_I, P_{I'}) = \int d_0(P_{\langle I,t\rangle}, P_{\langle I',t\rangle}) d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

where d_0 is any metric on one-dimensional probability distributions, and $d\lambda(t)$ is a distribution measure on the projections. Indeed, it is easy to check that d is a metric. In particular, note that in view of the Cramér Wold theorem, $d(P, Q) = 0$ if and only if $P = Q$. As mentioned, each one-dimensional distribution $P_{\langle I,t\rangle}$ can be approximated by a suitable histogram $\mu^t(I) = (\mu^t_h(I))_{h=1,\ldots,H} \in R^H$, so that in the limit in which the histogram approximation is accurate,

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I')) d\lambda(t), \quad \forall I, I' \in \mathcal{X}, \tag{A.7}$$

where d_μ is a metric on histograms induced by d_0.

A natural question is whether there are situations in which a finite number of projections suffice to discriminate any two probability distributions, that is, $P_I \neq P'_I \Leftrightarrow d(P_I, P_{I'}) \neq 0$. Empirical results show that this is often the case with a small number of templates (see [125] and HMAX experiments). The problem of mathematically characterizing the situations in which a finite number of (one-dimensional) projections are sufficient is challenging. Here we provide a partial answer to this question.

We start by observing that the metric (A.7) can be approximated by uniformly sampling K templates and considering

$$\hat{d}_K(P_I, P_{I'}) = \frac{1}{K} \sum_{k=1}^{K} d_\mu(\mu^k(I), \mu^k(I')), \tag{A.8}$$

where $\mu^k = \mu^{t^k}$. The following result shows that a finite number K of templates is sufficient to obtain an approximation within a given precision ϵ. Toward this end, let

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\| \mu^k(I) - \mu^k(I') \right\|_{\mathbb{R}^H}, \tag{A.9}$$

where $\|\cdot\|_{\mathbb{R}^H}$ is the Euclidean norm in $\mathbb{R}^H$. The following theorem holds.

Theorem A.5 Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Then

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon, \tag{A.10}$$

with probability $1 - \delta^2$, for all $I, I' \in \mathcal{X}_n$.

Proof: The proof follows from an application of the Hoeffding inequality (see box A.1) and a union bound.

Fix $I, I' \in \mathcal{X}_n$. Define the real random variable $Z : \mathbb{S}(\mathbb{R}^d) \to \mathbb{R}$,

$$Z(t^k) = \left\| \mu^k(I) - \mu^k(I') \right\|_{\mathbb{R}^H}, \ \ k = 1, \ldots, K.$$

From the definitions it follows that $\|Z\| \leq c$, and $\mathbb{E}(Z) = d(P_I, P_{I'})$. Then the Hoeffding inequality implies

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| = |\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}(Z) - Z(t^k)| \geq \epsilon,$$

with probability at most $e^{-c\epsilon^2 k}$. A union bound implies a result holding uniformly on $\mathcal{X}_n$; the probability becomes at most $n^2 e^{-c\epsilon^2 K}$. The desired result is obtained, noting that this probability is less than δ^2 as soon as $n^2 e^{-c\epsilon^2 K} < \delta^2$, that is, $K \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$.

Box A.1
Hoeffding Inequality

Suppose we have a set of variables X_i. It is intuitive that when we average some of them, we should normally get a result close to the expected value. Hoeffding quantifies normally and close. More formally, let $X_1, \cdots, X_N$ be a set of independent, independently distributed random variables such that $0 \leq X_i \leq 1$. We have

$$Pr\Big(|\frac{1}{N}\sum_{i=1}^{N} X_i - E(X)| > \epsilon\Big) \leq \delta = 2e^{-2N\epsilon^2}.$$

This result shows that the discriminability question can be answered in terms of empirical estimates of the one-dimensional distributions of projections of the image and transformations induced by the group on a number of templates t^k, $k = 1, \ldots, K$.

Theorem A.5 can be compared to a version of the Cramér-Wold theorem for discrete probability distributions. Heppes [125, theorem 1] shows that for a probability distribution consisting of k atoms in $\mathbb{R}^d$, at most $k+1$ directions ($d_1 = d_2 = \ldots = d_{k+1} = 1$) are enough to characterize the distribution and thus a finite (albeit large) number of one-dimensional projections.

The signature $\Sigma(I) = (\mu_1^1(I), \ldots, \mu_H^K(I))$ is obviously invariant (and unique), since it is associated to an image and all its transformations (an orbit). Each component of the signature is also invariant; it corresponds to a group average. Indeed, each measurement can be defined as

$$\mu_h^k(I) = \frac{1}{|G|}\sum_{g \in G} \eta_h\Big(\Big\langle gI, t^k \Big\rangle\Big), \tag{A.11}$$

for G finite group, or equivalently,

$$\mu_h^k(I) = \int_G dg\ \eta_h\Big(\Big\langle gI, t^k \Big\rangle\Big) = \int_G dg\ \eta_h\Big(\Big\langle I, g^{-1}t^k \Big\rangle\Big), \tag{A.12}$$

when G is a (locally) compact group. Here, the nonlinearity η_h is chosen to define a histogram approximation. Then it is clear that from the properties of the Haar measure we have

$$\mu_h^k(\bar{g}I) = \mu_h^k(I), \quad \forall \bar{g} \in G, I \in \mathcal{X}. \tag{A.13}$$

This is a key property of the signature and of our entire model.

A.2.4 Number of Templates Depends on Pooling Size

Let us consider how a finite number of projections suffice to discriminate the distributions generated by the subset $|G|$ of a finite, discrete group. We assume

that the set of images (of dimensionality d, where d is the number of pixels) is also finite of cardinality n. Though not necessary for the main results of this section, we may also assume that the images themselves can assume a finite set of discrete (vector) values (b bits per pixel) so that the associated distribution $P_G(I)$ consists of $b^d \times |G|$ atoms.

Instead of "bits per pixel" a different representation can be used, such as bits per wavelet coefficient.

We think here of the group acting on the images, while the templates $t^k, k = 1, \cdots, K$ are fixed and not transformed. Later, once we have results connecting distributions on images induced by the group with one-dimensional distributions of projections of the images on the templates, we can think of the transformations acting on the templates instead of the images.

Extension of Johnson-Lindenstrauss (J-L)lemma to discrete distributions. Recall the Johnson-Lindenstrauss lemma.

Proposition A.6 For any set V of n points in $\mathbb{R}^d$, there exists a map $P : \mathbb{R}^d \to \mathbb{R}^K$ such that for all $u, v \in V$,

$$(1 - \epsilon) \parallel u - v \parallel \leq \parallel Pu - Pv \parallel \leq (1 + \epsilon) \parallel u - v \parallel,$$

where the map P is a random projection on $\mathbb{R}^K$ and

$$K \geq C(\epsilon)\ln(n), \quad C(\epsilon) = 4C_0\epsilon^{-2}.$$

Suppose u and v are two images in $\mathbb{R}^d$, and consider $P_G(u)$ and $P_G(v)$, generated by the action of G on u and v, respectively. Each distribution has G atoms (with possible but unlikely multiplicity) supported in $\mathbb{R}^d$. The key observation is that $P_G(u)$ and $P_G(v)$ are equivalent to the orbits of u and v, respectively, and thus contain the same sample points (e.g., images) irrespective of the order. If there is a point in $P_G(u)$ that is ϵ-different from any other point in $P_G(v)$, the two distributions are different. The J-L projections provide the required accuracy if $K \geq C_0\epsilon^{-2}\ln(n|G|)$. This is because in the J-L proof the Hoeffding inequality (implies the result in proposition (A.6)) for a specific u and v with probability at least $1 - e^{-c\epsilon^2 k}$. If we require this to hold uniformly for all $|G|$ elements of an orbit and for all $n(n-1)$ possible pairs of orbits, the number of required templates must satisfy $K \geq C_0\epsilon^{-2}\ln(n(n-1)|G|)$. The last step is to show that this implies relations between the multidimensional distributions $P_G(u), P_G(v)$ and the one-dimensional distributions $P_G(\langle u, t_k\rangle), P_G(\langle v, t_k\rangle)$. The correspondence rests on the discreteness of the distributions and the fact that the J-L lemma result implies that if u and v are very close (that is, $||u - v|| \leq \eta$), their projections $P(u)$ and $P(v)$ are very close in every norm, in particular componentwise (that is, $\max_k |P(u)_k - P(v)_k| \leq \eta$).

In particular, if each of the $|G|$ $g_i u$ and $g_j v$ are close to each other, it follows that the K distributions of the projections are also close, for instance, in the "sup" norm (and vice versa). If any of the distributions is not close, it means that at least one of the $g_i u$ is not close to any of the $g_j v$, and this in turn implies that the $P_G(u)$ and $P_G(v)$ are not close. The distributions can be computed from the components of Pu in terms of their moments or directly, since the number of atoms in each one-dimensional distribution is finite and equal to $|G|$.

Since if two orbits are different none of the $|G|$ elements can be the same, the conditions on the bounds may be relaxed. Somewhat counterintuitively, the comparison of two orbits should have higher confidence than the comparison of two points.

To give some perspective, consider an old-fashioned setup with the discrete distribution $P_G(u)$ and its moment-generating function $M_u(t) = \sum_{j=1}^{|G|} e^{\langle t, g_j u \rangle} p(g_j u)$. Choose $|G|$ appropriate templates t_i. Call m the $|G|$ dimensional vector of the moments and a the vector of probabilities. Then

$$m = Ma,$$

where M is the $|G| \times |G|$ matrix of $e^{\langle t_i, x_j \rangle}$. The equation can be solved for the probabilities a provided M^{-1} exists. This suggests that from the empirical moment-generating function it is possible to recover exactly the group orbits with a number of templates K in the order of $|G|$. We can summarize this reasoning in the following:

Theorem A.7 Let $\mathcal{X}_n$, $|\mathcal{X}_n| = n$, a finite subset of vectors from the vector space $\mathcal{X} = \mathbb{R}^d$, and let G be a finite discrete group. The probability distributions associated to any pair of orbits are ϵ-different iff any pair of the orbit points is ϵ-different. A similar result holds if we consider instead of the orbits their projections onto $K \propto \ln(n(n-1)|G|)$ random vectors and the associated probability distributions.

Proof: $\Leftarrow$ Let us start by considering a pair of vectors of $\mathcal{X}_n$, I, I' and their orbits (the generalization to n orbits is easy). Suppose the two orbits coincide up to an ϵ factor. This means that for the ordered $|G|$-couples of vectors ($v_i = g_i I \in O_I$, $u_i = g_i I' \in O_{I'}$) we have

$$\sup_j |(v_i)_j - (u_i)_j| \leq \epsilon, \ \forall i = 1, \ldots, |G|, \ j = 1, \ldots, d. \tag{A.14}$$

The condition in (A.14) guarantees that

$$KS(P_I, P_{I'}) \leq \frac{d|G|\epsilon}{d|G|} = \epsilon,$$

where KS is the Kolmogorov-Smirnov distance between the cumulative distribution functions of the orbits of I and I', calculated as

$$KS(P_I, P_{I'}) = \sup_i |\mathrm{cdf}_i(P_I) - \mathrm{cdf}_i(P_{I'})|,$$

where with cdf_i indicates the i^{th} component of the cdf histogram plotting the distribution of the vector values. In fact, the maximum difference between the cdfs will occur in the case when all vectors have the same entries; in that case, we will have an accumulation error from all the differences of $d\epsilon|G|$, which simplifies with the cdf normalization factor $1/d|G|$.

$\Rightarrow$ Conversely, if the probability distributions associated to two orbits differ by an ϵ factor in the orbit points, any couple of ordered points in the orbits can maximally differ by the same factor.

This reasoning can be repeated if we consider projections of the orbits onto random vectors and the associated probability distributions. By the J-L lemma we can distinguish $2|G|$ vectors (number of elements of the two orbits) with $K \propto \ln(2|G|)$ projections. If we want this to hold uniformly over $\mathcal{X}_n$, we have to consider all possible orbit couples, and therefore the number of required projections will be $K \propto \ln(n(n-1)|G|)$. We want to prove that this is enough for the $n(n-1)$ associated probability distributions to be distinguished. Consider the projections $\langle g_i I, t\rangle$, $I \in \mathcal{X}_n$, and construct the associated probability distributions. The same reasoning can be applied: instead of the orbit elements u_i, v_i, we have their projections $\langle u_i, t\rangle, \langle v_i, t\rangle$ and the associated probability distributions are $P_{\langle I,t\rangle}, P_{\langle I',t\rangle}$.

A.2.5 Partially Observable Groups

A.2.5.1 Invariance for Partially Observable Groups Implies Localization

In this section we analyze how the invariance property described in eq. (A.13) of the signature defined in eqs. (A.11) and (A.12) changes when we do not have access to the whole set of the transformations (orbits) but to part of it. In particular, we show how invariance is linked to localization of the dot product $\langle I, gt\rangle$ in the group transformation.

Since the group is only partially observable, we introduce the notion of *partial invariance* for images and transformations G_0 that are within the observation window. Partial invariance is defined in terms of invariance of

$$\mu_h^k(I) = \frac{1}{V_0} \int_{G_0} dg\ \eta_h\Big(\Big\langle gI, t^k\Big\rangle\Big). \tag{A.15}$$

We recall that when gI and t^k do not share any common support on the plane, or I and t are uncorrelated, then $\langle gI, t^k\rangle = 0$. The following theorem, where

G_0 corresponds to the pooling range states, is a sufficient condition for partial invariance in the case of a locally compact group (see also figure A.1).

Theorem A.8 (Localization and Invariance) Let $I, t \in H$, be a Hilbert space, $\eta_h : \mathbb{R} \to \mathbb{R}^+$ a set of bijective (positive) functions, and G a locally compact group. Let $G_0 \subseteq G$, and suppose supp $(\langle gI, t^k \rangle) \subseteq G_0$. Then for any given $\bar{g} \in G$, $t^k, I \in \mathcal{X}$, the following conditions hold:

$$\begin{aligned} &\left\langle gI, t^k \right\rangle = 0, \ \forall g \in G/(G_0 \cap \bar{g}G_0) \\ &\text{or equivalently,} \qquad\qquad \Rightarrow \ \mu_h^k(I) = \mu_h^k(\bar{g}I) \\ &\left\langle gI, t^k \right\rangle \neq 0, \ \forall g \in G_0 \cap \bar{g}G_0. \end{aligned} \tag{A.16}$$

Proof: To prove the implication, note that if $\left\langle gI, t^k \right\rangle = 0, \ \forall g \in G/(G_0 \cap \bar{g}G_0)$, with $G_0 \Delta \bar{g}G_0 \subseteq G/(G_0 \cap \bar{g}G_0)$ (Δ is the symbol for symmetric difference $A \Delta B = (A \cup B)/(A \cap B)$ A, B sets), we have

$$\begin{aligned} 0 &= \int_{G/(G_0 \cap \bar{g}G_0)} dg \ \eta_h\Big(\left\langle gI, t^k \right\rangle\Big) \\ &= \int_{G_0 \Delta \bar{g}G_0} dg \ \eta_h\Big(\left\langle gI, t^k \right\rangle\Big) \\ &\geq \Big| \int_{G_0} dg \ \Big(\eta_h\Big(\left\langle gI, t^k \right\rangle\Big) - \eta_h\Big(\left\langle g\bar{g}I, t^k \right\rangle\Big)\Big)\Big|. \end{aligned} \tag{A.17}$$

The second equality is true, since, being η_h positive, the fact that the integral is zero implies $\left\langle gI, t^k \right\rangle = 0 \ \forall g \in G/(G_0 \cap \bar{g}G_0)$ (and therefore in particular $\forall g \in G_0 \Delta \bar{g}G_0$). The right hand side of the inequality being positive, we have

$$\Big| \int_{G_0} dg \ \Big(\eta_h\Big(\left\langle gI, t^k \right\rangle\Big) - \eta_h\Big(\left\langle g\bar{g}I, t^k \right\rangle\Big)\Big)\Big| = 0, \tag{A.18}$$

that is, $\mu_h^k(I) = \mu_h^k(\bar{g}I)$ (see figure A.1 for a graphical representation).

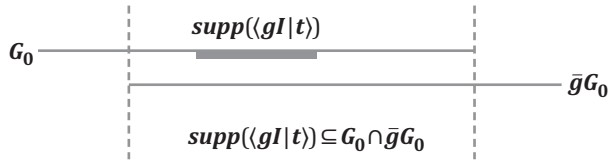


Figure A.1

A sufficient condition for invariance for locally compact groups. If the support of $\langle gI, t \rangle$ is sufficiently localized, it will be completely contained in the pooling interval even if the image is group shifted or, equivalently (as shown), if the pooling interval is group shifted by the same amount.

Eq. (A.16) describes a *localization* condition on the inner product of the transformed image and the template. The preceding result naturally raises the question of whether the localization condition is also necessary for invariance. Clearly, this would be the case if eq. (A.17) could be turned into an equality, that is,

$$\begin{aligned} &\int_{G_0 \Delta \bar{g} G_0} dg\ \eta_h\big(\big\langle gI, t^k \big\rangle\big) \\ &= |\int_{G_0} dg\ \Big(\eta_h\big(\big\langle gI, t^k \big\rangle\big) - \eta_h\big(\big\langle g\bar{g}I, t^k \big\rangle\big)\Big)| \\ &= |\mu_h^k(I) - \mu_h^k(\bar{g}I)|. \end{aligned} \tag{A.19}$$

Indeed, in this case, if $\mu_h^k(I) - \mu_h^k(\bar{g}I) = 0$, and we further assume the natural condition $\langle gI, t^k \rangle \neq 0$ if and only if $g \in G_0$, then the localization condition (1.8) would be necessary, since η_h is a positive bijective function.

Invariance, Localization, and Wavelets

The conditions equivalent to optimal translation and scale invariance, which correspond to maximum localization in space and frequency, cannot be simultaneously satisfied because of the classic *uncertainty principle*: if a function $t(x)$ is essentially zero outside an interval of length Δx and its Fourier transform $\hat{I}(\omega)$ is essentially zero outside an interval of length $\Delta\omega$, then

$$\Delta x \cdot \Delta\omega \geq 1. \tag{A.20}$$

In other words, a function and its Fourier transform cannot both be highly concentrated. Interestingly, for our setup the uncertainty principle also applies to sequences (see [25]).

It is well known that the equality sign in the uncertainty principle above is achieved by Gabor functions (see [24]) of the form

$$\psi_{x_0,\omega_0}(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x},\ \sigma_x \in \mathbb{R}^+,\ \omega_0 \in \mathbb{R}. \tag{A.21}$$

The uncertainty principle leads to the concept of optimal localization instead of exact localization. In a similar way, it is natural to relax our definition of strict invariance (e.g., $\mu_h^k(I) = \mu_h^k(g'I)$) and to introduce ϵ-invariance as $|\mu_h^k(I) - \mu_h^k(g'I)| \leq \epsilon$. If we suppose, for instance, the localization condition

$$\langle T_x I, t\rangle = e^{-\frac{x^2}{\sigma_x^2}}, \quad \langle D_s I, t\rangle = e^{-\frac{s^2}{\sigma_s^2}},\ \sigma_x, \sigma_s \in \mathbb{R}, \tag{A.22}$$

we have

$$|\mu_h^k(T_{\bar{x}}I) - \mu_h^k(I)| = \frac{1}{2}\sqrt{\sigma_x}\Big(\text{erf}\big([-b,b]\Delta[-b+\bar{x},b+\bar{x}]\big)\Big),$$
$$|\mu_h^k(D_{\bar{s}}I) - \mu_h^k(I)| = \frac{1}{2}\sqrt{\sigma_s}\Big(\text{erf}\big([-1/S,S]\Delta[\bar{s}/S,S\bar{s}]\big)\Big),$$

where "erf" is the error function. The differences, with an opportune choice of the localization ranges σ_s, σ_x, can be made as small as desired.

We end with a conjecture: the optimal ϵ-invariance is satisfied by templates with noncompact support that decay exponentially, such as a Gaussian or a Gabor wavelet. We can then speak of *optimal invariance* meaning optimal ϵ-invariance. This reasoning leads to the following theorem:

Theorem A.9 Assume invariants are computed from pooling within a pooling window with a set of linear filters. Then the optimal templates (e.g., filters) for maximum simultaneous invariance to translation and scale are Gabor functions

$$t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}. \tag{A.23}$$

Summarizing, we showed why Gabor templates are implied if we want optimal invariance in the case of partially observable groups of dilations and translations.

- The Gabor function $\psi_{x_0,\omega_0}(x)$ corresponds to a *Heisenberg box*, which has an x-spread $\sigma_x^2 = \int x^2|g(x)|dx$ and an ω spread $\sigma_\omega^2 = \int \omega^2|\hat{g}(\omega)|d\omega$ with area $\sigma_x\sigma_\omega$. Gabor wavelets arise under the action on $\psi(x)$ of the translation and scaling groups as follows. The function $\psi(x)$, as defined, is zero-mean and normalized,

  $$\int \psi(x)dx = 0 \tag{A.24}$$

  and

  $$\|\psi(x)\| = 1. \tag{A.25}$$

  A family of Gabor wavelets is obtained by translating and scaling ψ:

  $$\psi_{u,s}(x) = \frac{1}{s^{\frac{1}{2}}}\psi(\frac{x-u}{s}). \tag{A.26}$$

  Under certain conditions (in particular, the Heisenberg boxes associated with each wavelet must together cover the space-frequency plane) the Gabor wavelet family becomes a Gabor wavelet frame.

- Optimal self-localization of the templates (which follows from localization), when valid simultaneously for space and scale, is also equivalent to Gabor wavelets. If they are a frame, full information can be preserved in an optimal quasi-invariant way.

A.2.6 Approximate Invariance to Nongroup Transformations

In the previous section we analyzed the relation between localization and invariance in the case of group transformations. By relaxing the requirement of exact invariance and exact localization we show how the same strategy for computing invariants can still be applied even in the case of nongroup transformations if certain localization properties of $\langle TI, t\rangle$ hold, where T is a smooth transformation.

We first notice that the localization condition of theorems A.8 and A.9, when relaxed to approximate localization, takes the form (e.g., for $1-D$ translations group) $\left\langle I, T_x t^k \right\rangle < \delta \quad \forall x$ s.t. $|x| > a$, where δ is small in the order of $1/\sqrt{d}$ (n is the dimension of the space) and $\left\langle gI, t^k \right\rangle \approx 1 \quad \forall x$ s.t. $|x| < a$.

We call this property *sparsity of I in the dictionary t^k under G.* This condition can be satisfied by templates that are similar to images in the set and are sufficiently rich to be incoherent for small transformations. Note that from the preceding reasoning the sparsity of I in t^k under G is expected to improve with increasing d and with noiselike encoding of I and t^k by the architecture.

Another important property of sparsity of I in t^k (in addition to allowing local approximate invariance to arbitrary transformations) is *clutter tolerance* in the sense that if n_1, n_2 are additive uncorrelated spatial noisy clutter $\left\langle I + n_1, gt^k + n_2 \right\rangle \approx \langle I, gt\rangle$.

Interestingly, the *sparsity condition under the group* is related to associative memories, for instance, of the holographic type (see [121, 127]). If the sparsity condition holds only for $I = t^k$ and for very small set of $g \in G$ (if it has the form $\left\langle I, gt^k \right\rangle = \delta(g)\delta_{I,t^k}$) (this implies strict memory-based recognition (see noninterpolating look-up table in [52]) with inability to generalize beyond stored templates or views.

While the first regime — exact (or ϵ-) invariance for generic images, yielding universal Gabor templates — applies to the first layer of the hierarchy, this second regime (sparsity) — approximate invariance for a class of images, yielding class-specific templates — is important for dealing with nongroup transformations at the top levels of the hierarchy, where receptive fields may be as large as the visual field.

Several interesting transformations do not have the group structure, for instance, the change of expression of a face or the change of pose of a body. We

show here that approximate invariance to transformations that are not groups can be obtained if the approximate localization condition holds, and if the transformation can be locally approximated by a linear transformation, for instance, a combination of translations, rotations, and nonhomogeneous scalings, which corresponds to a locally compact group admitting a Haar measure.

As an example, consider the transformation induced on the image plane by rotation in depth of a face. It can be decomposed into piecewise linear approximations around a small number of key templates, each one corresponding to a specific 3-*D* rotation of a template face. Each key template corresponds to a complex cell containing as simple cells a number of observed transformations of the key template within a small range of rotations. Each key template corresponds to a different signature that is invariant only for rotations around its center. Notice that the form of the linear approximation or the number of key templates needed does not affect the algorithm or its implementation. The templates learned are used in the standard dot-product-and-pooling module. The choice of the key templates, each one corresponding to a complex cell and thus to a signature component, is not critical as long as there are enough of them. For one-parameter groups, the key templates correspond to the knots of a piecewise linear spline approximation. Optimal placement of the centers, if desired, is a separate problem that we leave aside for now.

A.3 Multilayer Architecture: Invariance, Covariance, and Selectivity

A.3.1 Recursive Definition of Simple and Complex Responses

So far we have studied the invariance, uniqueness, and stability properties of signatures, both in the case when a whole group of transformations is observable (see eq. (1.2) in chapter 1), and in the case in which it is only partially observable (see eqs. (A.11) and (A.12)). We now discuss how the preceding ideas can be iterated to define a multilayer architecture.

Consider first the case when G is finite. Given a subset $G_0 \subset G$, we can associate a window gG_0 to each $g \in G$. Then we can use eq. (1.5) to define for each window a signature $\Sigma(I)(g)$ given by the measurements

$$\mu_h^k(I)(g) = \frac{1}{|G_0|} \sum_{\bar{g} \in gG_0} \eta_h\Big(\big\langle I, \bar{g}t^k \big\rangle\Big).$$

Note that the averaging in the integral is done on transformed templates, not on transformed images. We keep this form as the definition of signature. For fixed h, k a set of measurements corresponding to different windows can be seen as a $|G|$-dimensional vector. A signature $\Sigma(I)$ for the whole image

is obtained as *a signature of signatures*, that is, a collection of signatures $(\Sigma(I)(g_1), \ldots, \Sigma(I)(g_{|G|})$ associated to each window.

Since we assume that the output of each module is made zero-mean and normalized before further processing at the next layer, *conservation of information from one layer to the next requires saving the mean and the norm* at the output of each module at each level of the hierarchy. We conjecture that the neural image at the first layer is uniquely represented by the final signature at the top of the hierarchy and the means and norms at each layer (see [1]).

This discussion can be extended to continuous (locally compact) groups considering,

$$\mu_h^k(I)(g) = \frac{1}{V_0} \int_{gG_0} d\bar{g} \eta_h\Big(\Big\langle I, \bar{g}t^k \Big\rangle\Big), \quad V_0 = \int_{G_0} d\bar{g},$$

where, for fixed h,k, $\mu_h^k(I) : G \to \mathbb{R}$ can now be seen as a function on the group. In particular, if we denote by $K_0 : G \to \mathbb{R}$ the indicator function on G_0, we can write

$$\mu_h^k(I)(g) = \frac{1}{V_0} \int_G d\bar{g} K_0(\bar{g}^{-1}g) \eta_h\Big(\Big\langle I, \bar{g}t^k \Big\rangle\Big).$$

The signature for an image can again be seen as a collection of signatures corresponding to different windows, but in this case it is a function $\Sigma(I) : G \to \mathbb{R}^{HK}$, where $\Sigma(I)(g) \in \mathbb{R}^{HK}$, is a signature corresponding to the window G_0 centered at $g \in G$.

This construction can be iterated to define a hierarchy of signatures. Consider a sequence $G_1 \subset G_2, \ldots, \subset G_L = G$. For $q : G \to \mathbb{R}^p$, $p \in \mathbb{N}$ with an abuse of notation, we let $gq(\bar{g}) = q(g^{-1}\bar{g})$. Then we can consider the following construction.

We call *complex operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $\mu_\ell(I) : G \to \mathbb{R}^{HK}$,

$$\mu_\ell^{h,k}(I)(g) = \frac{1}{|G_\ell|} \sum_{\bar{g} \in gG_\ell} \eta_h\Big(v_\ell^k(I)(\bar{g})\Big), \tag{A.27}$$

and *simple operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $v_\ell(I) : G \to \mathbb{R}^K$,

$$\mu_\ell^k(I)(g) = \Big\langle \mu_{\ell-1}(I), gt_\ell^k \Big\rangle, \tag{A.28}$$

with t_ℓ^k the kth template at layer ℓ and $\mu_0(I) = I$. Several comments are in order:

- Besides the first layer, the inner product defining the simple cell operator is that in $L^2(G) = \{q : G \to \mathbb{R}^{HK}, \mid \int dg |q(g)|^2 < \infty\}$.

- The index ℓ corresponds to different layers, corresponding to different subsets G_ℓ.
- At each layer a (finite) set of templates $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ ($\mathcal{T}_0 \subset \mathcal{X}$) is assumed to be available. For simplicity, we assume that $|\mathcal{T}_\ell| = K$ for all $\ell = 1, \ldots, L$. The templates at layer ℓ can be thought of as *compactly supported functions*, with support much smaller than the corresponding set G_ℓ. Typically templates can be seen as image patches in the space of complex operator responses, that is, $t_\ell = \mu_{\ell-1}(\bar{t})$ for some $\bar{t} \in \mathcal{X}$.
- Similarly we assume (for simplicity in dealing with indexes) that the number of nonlinearities η_h, considered at every layer, is the same.

The extension to continuous‘ (locally compact) groups is straightforward. We express it in the following definition.

Definition A.10 Simple and Complex Responses For $\ell = 1, \ldots, L$, let $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ (and $\mathcal{T}_0 \subset \mathcal{X}$) be a sequence of template sets. The complex operator at layer ℓ maps an image $I \in \mathcal{X}$ to a function $\mu_\ell(I) : G \to \mathbb{R}^{HK}$. In components

$$\mu_\ell^{h,k}(I)(g) = \frac{1}{V_\ell} \int d\bar{g} K_\ell(\bar{g}^{-1} g) \eta_h \left(\nu_\ell^k(I)(\bar{g}) \right), \; g \in G, \tag{A.29}$$

where K_ℓ is the indicator function on G_ℓ, $V_\ell = \int_{G_\ell} d\bar{g}$ and

$$\nu_\ell^k(I)(g) = \left\langle \mu_{\ell-1}(I), g t_\ell^k \right\rangle, \quad g \in G, \tag{A.30}$$

($\mu_0(I) = I$) is the simple operator at layer ℓ that maps an image $I \in \mathcal{X}$ to a function $\nu_\ell(I) : G \to \mathbb{R}^K$.

Remark A.11 Note that eq. (A.29) can be written as

$$\mu_\ell^{h,k}(I) = K_\ell * \eta_h(\nu_\ell^k(I)), \tag{A.31}$$

where $*$ is the group convolution.

A.3.2 Inheriting Transformations: Covariance

We call the map Σ covariant iff

$$\Sigma(gI) = g^{-1} \Sigma(I), \quad \forall g \in G, I \in \mathcal{X},$$

when g acting on Σ is properly defined. In the following we show the covariance property for the $\mu_1^{h,k}$ response (figure A.2). Inductive reasoning then can be applied for higher-order responses. *We assume that the architecture*

Figure A.2
Covariance: the response for an image I at position g is equal to the response of the group-shifted image at the shifted position.

is isotropic in the relevant covariance dimension (this implies that all the modules in each layer should be identical with identical templates) and that there is a continuum of modules in each layer. The following proposition proves the covariance property of the signature defined in eq. (A.29).

Proposition A.12 Let G be a locally compact group and $\bar{g} \in G$. Let $\mu_1^{h,k}$ be as defined in (A.29). Then $\mu_1^{h,k}(\tilde{g}I)(g) = \mu_1^{h,k}(I)(\tilde{g}^{-1}g)$.

Proof: Using eq. (A.29), we have

$$\begin{aligned}
\mu_1^{h,k}(\tilde{g}I)(g) &= \frac{1}{V_1}\int_G d\bar{g} K_1(\bar{g}^{-1}g)\eta_h\left(\left\langle \tilde{g}I, \bar{g}t^k\right\rangle\right) \\
&= \frac{1}{V_1}\int_G d\bar{g} K_1(\bar{g}^{-1}g)\eta_h\left(\left\langle I, \tilde{g}^{-1}\bar{g}t^k\right\rangle\right) \\
&= \frac{1}{V_1}\int_G d\hat{g} K_1(\hat{g}^{-1}\tilde{g}^{-1}g)\eta_h\left(\left\langle I, \hat{g}t^k\right\rangle\right) \\
&= \mu_1^{h,k}(I)(\tilde{g}^{-1}g),
\end{aligned}$$

where in the third line we used the change of variable $\hat{g} = \tilde{g}^{-1}\bar{g}$ and the invariance of the Haar measure.

Remark A.13 The covariance property described in proposition A.12 can be stated equivalently as $\mu_1^{h,k}(I)(g) = \mu_1^{h,k}(\bar{g}I)(\bar{g}g)$. This expression has a intuitive meaning, as shown in figure A.2.

Remark A.14 With respect to the range of invariance, the following property holds for multilayer architectures in which the output of a layer is defined as covariant if it transforms in the same way as the input. For a given transformation of an image or part of it, the signature from complex cells at a certain layer is either invariant or covariant with respect to the group of transformations. If it is covariant, there will be a higher layer in the network at which it is invariant (see theorem A.16), assuming that the image is contained in the

visual field. This property gives a *stratification* of ranges of invariance in the ventral stream: invariances should appear in a sequential order, meaning that smaller transformations will be invariant before larger ones, in earlier layers of the hierarchy (see [128]).

Remark A.15 *Covariance and wavelets*. The result provend in Theorem A.9 shows how optimal invariance of complex cells for scale and translation transformations leads to Gabor wavelet templates, and thus simple cells operate a Gabor wavelet transform of the image. However, the necessity of the wavelet transform (not necessarily Gabor) can be also proved from the very general requirements of linearity, covariance, and energy conservation of the representation,[129, p 41]. A related approach was taken by Stevens in [130], who showed that (1) the wavelet transform is covariant with the similitude group, and (2) the wavelet transform follows from the requirement of covariance (rather than invariance) for the simple cells.

A.3.2.1 Invariance in Hierarchical Architectures

We now find the conditions under which the functions μ_ℓ are locally invariant, that is, invariant within the restricted range of the pooling. We further prove that the range of invariance increases from layer to layer in the hierarchical architecture. The fact that for an image, in general, no more global invariance is guaranteed allows a novel definition of parts of an image.

The local invariance conditions are a simple reformulation of theorem A.8 in the context of a hierarchical architecture. In the following, for simplicity, we suppose that at each layer we only have a template t and a nonlinear function η.

Proposition A.16 (Invariance and Localization: Hierarchy) Let $I, t \in H$,be a Hilbert space, $\eta : \mathbb{R} \to \mathbb{R}^+$ a bijective (positive) function, and G a locally compact group. Let $G_\ell \subseteq G$, and suppose $\text{supp}(\langle g\mu_{\ell-1}(I), t\rangle) \subseteq G_\ell$. Then for any given $\bar{g} \in G$,

$$\begin{aligned} \mu_\ell(I) = \mu_\ell(\bar{g}I) \;\Leftrightarrow\; & \langle g\mu_{\ell-1}(I), t\rangle = 0,\ g \in G/(G_\ell \cap \bar{g}G_\ell), \\ & \langle g\mu_{\ell-1}(I), t\rangle \neq 0,\ g \in G_\ell \cap \bar{g}G_\ell. \end{aligned} \tag{A.32}$$

Proof: The proof follows the reasoning in theorem A.8 with I substituted by $\mu_{\ell-1}(I)$ using the covariance property $\mu_{\ell-1}(gI) = g\mu_{\ell-1}(I)$. Note how, similarly to eq. (A.16) , eq. (A.32) gives a localization condition.

We can now give a formal definition of *object part* as the subset of the signal I whose complex response, at layer ℓ, is invariant under transformations in the range of the pooling at that layer. This definition is consistent, since the invariance is increasing from layer to layer, therefore allowing bigger and bigger parts. Consequently, for each transformation there exists a layer $\bar{\ell}$ such that any signal subset will be a part at that layer.

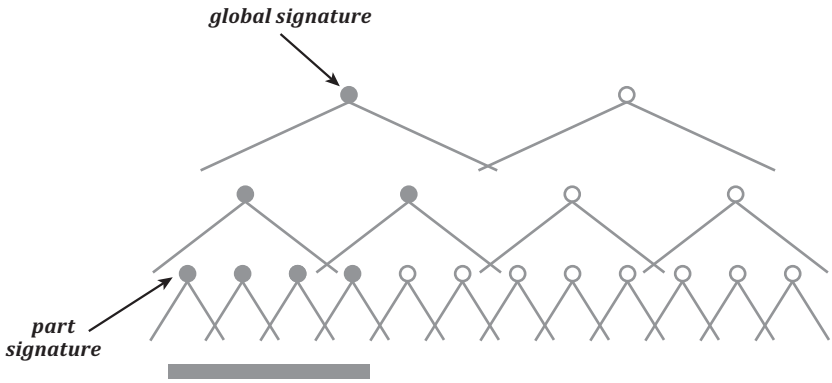


Figure A.3

An image I (rectangle) with a finite support may or may not be fully included in the receptive field of a single complex cell at layer ℓ (and the transformed image may not be included in the pooling range of the complex cell). However, there will be a higher layer such that the support of its neural response is included in the pooling range of a single complex cell.

We can now state the following theorem.

Theorem A.17 (Whole and Parts) Let $I \in \mathcal{X}$ (an image or a subset of it) and μ_ℓ be the complex response at layer ℓ. Let $G_0 \subseteq \cdots \subset G_\ell \subseteq \cdots \subseteq G_L = G$ be a set of nested subsets of the group G. Suppose η is a bijective (positive) function and that the template t and the complex response at each layer has finite support. Then $\forall \bar{g} \in G$, $\mu_\ell(I)$ is invariant for some $\ell = \bar{\ell}$,

$$\mu_m(\bar{g}I) = \mu_m(I), \;\; \exists\, \bar{\ell} \;\text{ s.t. }\; \forall m \geq \bar{\ell}.$$

Proof: The proof follows from the observation that the pooling range over the group is a bigger and bigger subset of G with a growing layer number. In other words, there exists a layer such that the image and its transformations are within the pooling range at that layer (figure A.3). This is clear, since for any $\bar{g} \in G$ the nested sequence

$$G_0 \cap \bar{g}G_0 \subseteq ... \subseteq G_\ell \cap \bar{g}G_\ell \subseteq ... \subseteq G_L \cap \bar{g}G_L = G$$

will include a set $G_{\bar{\ell}} \cap \bar{g}G_{\bar{\ell}}$ such that

$$\left\langle g\mu_{\bar{\ell}-1}(I), t \right\rangle \neq 0 \;\; g \in G_{\bar{\ell}} \cap \bar{g}G_{\bar{\ell}},$$

being $\text{supp}(\left\langle g\mu_{\bar{\ell}-1}(I), t \right\rangle) \subseteq G$.

Theorem A.17 gives a new formal definition of an object part as the part of the visual signal whose representation is invariant at a certain layer of the hierarchical architecture.

A.4 Complex Cells: Wiring and Invariance

We prove the equivalence of signatures obtained by pooling over the projections of template orbits or their principal component. Let $L^2(\mathcal{G}) = \{F : \mathcal{G} \to \mathbb{R} \mid \int |F(g)|^2 dg < \infty\}$, and

$$T_t : \mathcal{X} \to L^2(\mathcal{G}), \qquad (T_t f)(g) = \langle f, T_g t \rangle,$$

where $t \in \mathcal{X}$. It is easy to see that T_t is a linear bounded and compact operator if $\|T_g t\| < \infty$. (In fact, it is easy to see that T is Hilbert-Schmidt, $\mathrm{Tr}(T_t^* T_t) = \int dg \|T_g t\|$.) Denote by $(\sigma_i; u_i, v_i)_i$ the singular system of T_t, where $(u_i)_i$ and $(v_i)_i$ are orthonormal bases for $\mathcal{X}$ and $L^2(\mathcal{G})$, respectively.

For $\sigma : \mathbb{R} \to \mathbb{R}$ measurable, define (complex response)

$$c : \mathcal{X} \to \mathbb{R}, \qquad c(I) = \sum_i \sigma(\langle I, u_i \rangle).$$

If $\sigma(a) = |a|^2$, $a \in \mathbb{R}$, and T_t / b_t is an isometry, where b_t is a constant possibly depending on t (see [131]), then c is invariant. Indeed,

$$c(I) = \|I\|^2 = \frac{1}{b_t^2} \|T_t I\|^2_{L^2(\mathcal{G})} = \frac{1}{b_t^2} \int |\langle I, T_g t \rangle|^2 dg$$

for all $I \in \mathcal{X}$, and $\|T_t I\|^2_{L^2(\mathcal{G})} = \|T_t T_{g'} I\|^2_{L^2(\mathcal{G})}$ for all $g' \in \mathcal{G}$.

If $\mathcal{G}$ is the *affine group* and $\mathcal{X} = L^2(\mathbb{R})$, then under the admissibility condition

$$\int |\langle T_g t, t \rangle|^2 < \infty,$$

it is possible to take $b_t = \sqrt{C_t}$, with $C_t = 2\pi \int \frac{|\hat{t}(\omega)|^2}{\omega} d\omega$, where $\hat{t}$ denotes the Fourier transform of t.

A.5 Mirror-Symmetric Templates Lead to Odd-Even Covariance Eigenfunctions

As explained in section 4.5, in the brain face patch system, specifically in the anterior patch AL, neurons are found to be tuned to even receptive fields. A possible explanation in terms of extracted PCs of a set of rotated-in-depth faces is as follows. Let $R : \mathbb{R}^K \to \mathbb{R}^K$ denote the reflection operator. The eigenfunctions of R are odd and even functions with eigenvalue $-1, 1$, respectively. We want to prove that the covariance matrix associated with the set of templates and their mirror symmetric has odd or even eigenfunctions.

Consider a set that only contains one template t and its reflection $R\tau$ (a unitary transformation). Let Z denote the operator that takes a neural frame T and returns the set consisting of t and its reflection.

$$Z = (\mathbb{1}, R),$$
$$Z(t) = \mathbb{T} = \begin{pmatrix} T \\ Rt \end{pmatrix},$$

We want to prove that R commutes with $Z^\intercal Z$, namely, that they have common eigenvectors. We have

$$Z^\intercal Z(Rt) = \begin{pmatrix} Rt + t \\ t + Rt \end{pmatrix} = R(Z^\intercal Z(t)). \tag{A.33}$$

Therefore $R(Z^\intercal Z) = (Z^\intercal Z)R$ (they commute), and $Z^\intercal Z$ and R must have the same eigenfunctions. Since the eigenfunctions of R are even and odd, the principal components of $\mathbb{T}^\intercal \mathbb{T}$ must also be even and odd.

A.6 Gabor-like Shapes from Translation Group

A.6.1 Spectral Properties of Template Transformations Covariance Operator: Cortical Equation

This section focuses on characterizing the spectral properties associated with the covariance of the template transformations. It proposes a cortical equation whose solution provides the eigenfunctions of the covariance of the transformed templates. We consider a layer of 2-D apertures and the covariance of the template transformations associated with each aperture resulting from transformations of images seen through one of these apertures. This leads to an explicit solution for the first layer in the case of translations.

For any fixed t we want to solve the spectral problem associated to the set

$$\mathbb{T}_t = (g_0 t, g_1 t, \ldots, g_{|G|} t)^T,$$

that is, we want to find the eigenvalues λ_i and eigenfunctions ψ_i such that

$$\mathbb{T}_t^* \mathbb{T}_t \psi_i = \lambda_i \psi_i. \tag{A.34}$$

To state the problem precisely we need some definitions. We start with the 1-D problem. We show how to derive an analytical expression of the visual cortex cells tuning based on the following:

- *Observables*. Images, transformed by a locally compact group, seen through an aperture;

- *Hebbian learning*. Hebbian-like synapses exist in the visual cortex.

We fix few objects:

- $\mathcal{X}$ space of signals: $L^2(\mathbb{R}, dx)$;
- $\mathcal{T} \subseteq \mathcal{X}$ is the template set.

We solve the eigenproblem associated to the continuous version of eq. (A.34). In this case the basic observable is given by the operator $T : \mathcal{X} \to \mathcal{X}$,

$$(TI)(y) \equiv [t * M_a I](y) = \int dx t(y-x)a(x)I(x), \;\; t \in \mathcal{T},\; a, I \in \mathcal{X}, \tag{A.35}$$

where

$$(M_a I)(x) = a(x)I(x),\; a \in \mathcal{X}.$$

Eq. (A.35) is the mathematical expression of the observable T, that is, a translating template t seen through the function a, which will be called the aperture.

Remark A.18 T is linear (from the properties of the convolution operator) and bounded (from $\|T\| = \|\mathfrak{F}(T)\| = \|t\|\,\|a\|$).

Remark A.19 M_a is a self-adjoint operator. The adjoint operator $T^* : \mathcal{X} \to \mathcal{X}$ is given by

$$\begin{aligned}\left\langle TI, I' \right\rangle &= \int dy\, \bar{I}'(y) \int dx\, t(y-x)a(x)I(x) \\ &= \int dx\, I(x)a(x) \int dy\, t(y-x)\bar{I}'(y) = \left\langle I, T^* I' \right\rangle,\end{aligned}$$

which implies $T^* I = M_a(t^- * I)$, $t^-(x) = t(-x)$. Note that $\|t\| = \left\|t^-\right\| \Rightarrow \|T\| = \|T^*\|$, that is, $\|T^*\|$ is bounded.

Assuming Hebbian learning, we have that the tuning of the cortical cells is given by the solution of the spectral problem of the covariance operator associated to T, $T^*T : \mathcal{X} \to \mathcal{X}$,

$$\begin{aligned}[T^*TI](y) &= M_a[t^- * (t * (M_a I))](y) = M_a[(t^- * t) * (M_a I)](y) \\ &= M_a(t^{\circledast} * (M_a I)) = a(y) \int dx\, a(x) t^{\circledast}(y-x) I(x),\end{aligned}$$

where $t^{\circledast}$ is the autocorrelation of t. This expression can be written as

$$[T^*TI](y) = \int dx\, K(x,y) I(x), \quad K(x,y) = a(x)a(y)t^{\circledast}(y-x).$$

Since the kernel K is Hilbert-Schmidt,

$$Tr(K) = \int dx\, K(x,x) = \int dx\, a^2(x) t^{\circledast}(0) < \infty,$$

we have that

- the eigenfunctions corresponding to distinct eigenvalues are orthogonal;
- the eigenvalues are real and positive;
- there is at least one eigenvalue and one eigenfunction (when K is almost everywhere nonzero), and in general a countable set of eigenfunctions.

Now, we aim to find $\psi_n \in \mathcal{X}$ and $\lambda_n \in \mathbb{R}$ such that

$$a(y) \int dx\, a(x) t^{\circledast}(y-x)\psi_n(x) = \lambda_n \psi_n(y) \tag{A.36}$$

and study their properties. In particular, we find approximate solutions and show that they are Gabor-like shapes. An exact solution in some particular cases can be found in section A.6.2.

A.6.1.1 Square Aperture, Circulants, and Fourier

We start from the simplest discrete case in which we assume periodic boundary conditions on each aperture (one in a layer of receptive fields), resulting in a circulant structure of the collection of transformed templates. Define T as the circulant matrix where each column represents a template t shifted relative to the previous column. This corresponds to assuming that the visual world translates and is seen through a square aperture with periodic boundary conditions. Let us assume in this example that the image is one-dimensional. Thus the image seen through an aperture,

$a(x)$ s.t. $a(x) = 1$ for $0 \leq x \leq A$; $a(x) = 0$ otherwise,

is $t(x-y)a(x)$ when the image is shifted by y. We are led to the following problem. Find the eigenvectors of the symmetric matrix $T^T T$, where T is a circulant matrix (see [132]). If we consider the continuous version of the problem, that is, the eigenvalue problem

$$\int_0^A dx \psi_n(x) t^{\circledast}(y-x)dx = \lambda_n \psi_n(y),$$

with $t^{\circledast}(x)$ being the autocorrelation function associated with t, the solution is $\psi_n(x) = e^{-i2\pi \frac{n}{A}x}$, which is the Fourier basis between 0 and A.

A.6.2 Cortical Equation: Derivation and Solution

Suppose $a(x) = e^{-\alpha x^2}$. Eq (A.36) can be written as

$$\lambda_n \psi_n(y) - e^{-\alpha y^2} \int dx\, e^{-\alpha x^2} t^{\circledast}(y-x)\psi_n(x) = 0,$$

or equivalently,

$$\lambda_n \xi_n(y) - \int dx\, e^{-2\alpha x^2} t^{\circledast}(y-x)\xi_n(x) = 0, \;\; \xi_n(x) = e^{+\alpha x^2}\psi_n(x).$$

Decomposing $t^{\circledast}(x)$ in the Fourier basis,

$$\lambda_n \xi_n(y) - \frac{1}{\sqrt{2\pi}} \int dx\, e^{-2\alpha x^2} \int d\omega\, t^{\circledast}(\omega) e^{i\omega(y-x)} \xi_n(x) = 0.$$

Deriving twice in the y variable and rearranging the expression,

$$\lambda_n \xi_n^{''}(y) + \frac{1}{\sqrt{2\pi}} \int d\omega\, \omega^2 t^{\circledast}(\omega) e^{i\omega y} \int dx\, e^{-2\alpha x^2} \psi_n(x) e^{-i\omega x} = 0. \qquad \text{(A.37)}$$

This expression is equivalent to the original eigenproblem in eq. (A.36) adding the conditions (e.g., $\xi_n(0) = c$, $\xi_n^{'}(0) = c^{'}$). In fact, given the derivative of a function, the original function is uniquely determined, having its value on a point. Therefore the function $\xi_n^{'}(y)$ uniquely determines $\xi_n(y)$ if we have, for instance, $\xi_n(0) = c$, and the function $\xi_n^{''}(y)$ uniquely determines $\xi_n^{'}(y)$ if we have, for instance, $\xi_n^{'}(0) = c^{'}$.

Using the Fourier operator $\mathfrak{F}$, we can rewrite (A.37) as

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\mathfrak{F}^{-1}\Big(\omega^2 t^{\circledast}(\omega)\mathfrak{F}\big(e^{-2\alpha x^2}\psi_n(x)\big)\Big) = 0.$$

By the convolution theorem,

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\mathfrak{F}^{-1}\big(\omega^2 t^{\circledast}(\omega)\big) * \big(e^{-2\alpha x^2}\xi_n(x)\big) = 0.$$

Expanding $\omega^2 t^{\circledast}(\omega)$ in a Taylor series and recalling that $\mathfrak{F}^{-1}(\omega^m) = i^m \delta^m(x)$, we are led to the differential equation

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\big(c_0\delta + ic_1\delta^{'} + \cdots\big) * \big(e^{-2\alpha x^2}\xi_n(x)\big) = 0, \qquad \text{(A.38)}$$

where c_i are the coefficients of the Taylor series expansion of $\omega^2 t^{\circledast}(\omega)$.

A.6.2.1 $1/\omega$ Spectrum

In the case of the average natural images spectrum

$$t(\omega) = \frac{1}{\omega},$$

the differential eq. (A.38) assumes the particularly simple form

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\, e^{-2\alpha y^2}\psi_n(y) = 0.$$

In the harmonic approximation

$$\xi_n^{''}(y) + \frac{\sqrt{2\pi}}{\lambda_n}(1 - 2\alpha y^2)\xi_n(y) = 0, \qquad \text{(A.39)}$$

which is of the form of a Weber differential equation (or 1-D Schrödinger equation for the harmonic oscillator). The explicit real solution is given by

$$\xi_n(y) = D\Big(-\frac{1}{2} + \frac{\pi^{\frac{1}{4}}}{2^{\frac{5}{4}}\sqrt{\alpha\lambda_n}}, \frac{2^{\frac{7}{8}}\alpha^{\frac{1}{4}}\pi^{\frac{1}{8}}}{\lambda_n{}^{\frac{1}{4}}}\Big),$$

where $D(\eta, y)$ is the parabolic cylinder function. It can be proved [134, p.139] that the solutions have two different behaviors, exponentially increasing or exponentially decreasing, and that we have exponentially decreasing solutions if and only if the following quantization condition holds:

$$-\frac{1}{2} + \frac{\pi^{\frac{1}{4}}}{2^{\frac{5}{4}}\sqrt{\alpha\lambda_n}} = n,\ n = 0, 1, \ldots.$$

Therefore, recalling that $\alpha = 1/\sigma_\alpha^2$, we obtain the spectrum quantization condition

$$\lambda_n = \sqrt{\frac{\pi}{2}}\frac{\sigma_\alpha^2}{(2n+1)^2},\ n = 0, 1, \ldots. \tag{A.40}$$

Further, using the identity (true if $n \in \mathbb{N}$)

$$D(n, y) = 2^{-\frac{n}{2}} e^{-\frac{y^2}{4}} H_n\Big(\frac{y}{\sqrt{2}}\Big),$$

where $H_n(y)$ are Hermite polynomials, we have

$$\xi_n(y) = 2^{-\frac{n}{2}} e^{-\frac{2n+1}{\sigma_\alpha^2}y^2} H_n\Big(\frac{\sqrt{2(2n+1)}}{\sigma_\alpha}y\Big),$$

or (see figure A.4)

$$\psi_n(y) = 2^{-\frac{n}{2}} e^{-\frac{2(n+1)}{\sigma_\alpha^2}y^2} H_n\Big(\frac{\sqrt{2(2n+1)}}{\sigma_\alpha}y\Big). \tag{A.41}$$

Remark A.20 The solution in eq. (A.41) is also an approximate solution for any template spectrum such that

$$\omega t(\omega) = const + O(\omega).$$

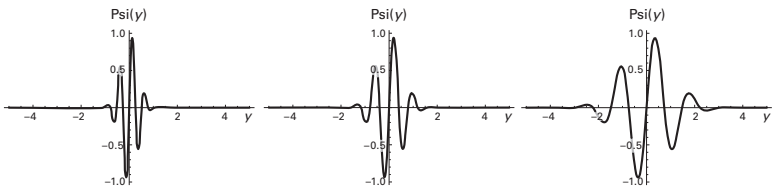


Figure A.4

Wavelet-like behavior of the solution for $\sigma_\alpha/\sqrt{\lambda} = C$. $C = 10,\ \lambda = 5, 10, 20$.

A.6.2.2 Extension to the Two-Dimensional Problem

Suppose the template t is moving only in the x direction. The spectral problem in 2-D is

$$\lambda_n \xi_n(x,y) - \int dx\, e^{-2\alpha(\eta^2+\gamma^2)} t^{\circledast}(x-\eta,y)\xi_n(\eta,\gamma) = 0,$$
$$\xi_n(x,y) = e^{+\alpha(x^2+y^2)}\psi_n(x,y).$$

Applying the operator ∇^2 and proceeding as in the 1-D case, we obtain

$$\lambda_n \nabla^2 \xi_n(x,y) + 2\pi \mathfrak{F}^{-1}\left((\omega_\eta^2+\omega_\gamma^2)t^{\circledast}(\omega_\eta,\omega_\gamma)\right) * \left(e^{-2\alpha(\eta^2+\gamma^2)}\xi_n(\eta,\gamma)\right)_{|_{y=0}} = 0.$$

If we suppose the spectrum is of the form

$$t(\omega_x,\omega_y) = \frac{1}{\sqrt{\omega_x^2+\omega_y^2}},$$

we are left with the differential equation

$$\nabla^2\xi(x,y) + \frac{2\pi}{\lambda_n} e^{-2\alpha x^2}\xi_n(x,0) = 0. \tag{A.42}$$

In the hypothesis $\xi_n(x,y) = \xi_{1,n}(x)\xi_{2,n}(y)$ we can immediately see, plugging the solution into eq. (A.42), that $\xi_n(x,y)$ is y-independent. Therefore the solution in the 2-D case is

$$\psi_n(x,y) = 2^{-\frac{n}{2}} e^{-\frac{y^2+2(n+1)x^2}{\sigma_\alpha^2}} H_n\left(\frac{\sqrt{2(2n-1)}}{\sigma_\alpha}x\right), \tag{A.43}$$

which gives

$$\frac{\sigma_y}{\sigma_x} = \sqrt{2(n+1)}.$$

A.7 Gabor-like Wavelets

A.7.1 Derivation from an Invariance Argument for 1-D Affine Group

The equality in eq. (A.19) in general is not true. However, this is clearly the case if we consider the group of transformations to be dilations and translations (see figure A.5a). We discuss this case in some detail.

Assume that $G_0 = [0,a]$. Let

$$S = \{\langle T_x I, t\rangle \ : \ x \in [0,a]\}, \ \ S_c = \{\langle T_x I, t\rangle \ : \ x \in [c,a+c]\} \tag{A.44}$$

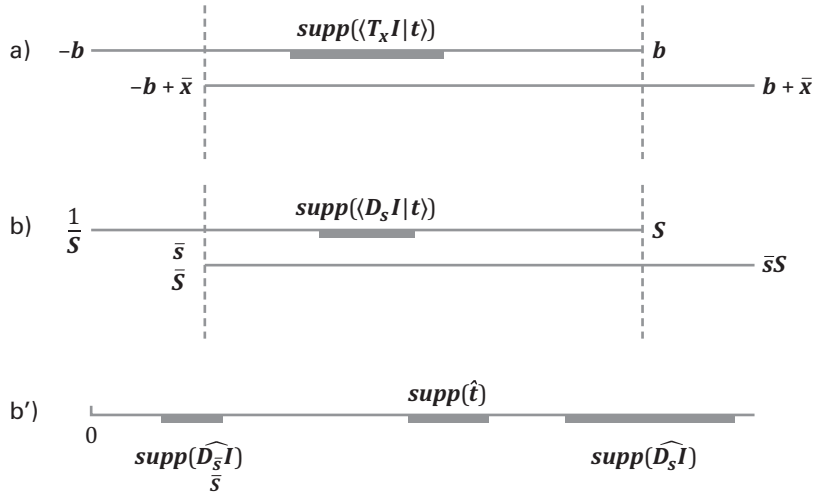


Figure A.5

If the support of the dot product between the image and the template is contained in the intersection between the pooling range and the group-translated (*a*) or dilated (*b*) pooling range the signature is invariant. In frequency condition (*b*) becomes (*b*′): When the Fourier supports of the dilated image and the template do not intersect, their dot product is zero.

for a given c, where T_x is a unitary representation of the translation operator. We can view S, S_c as sets of simple responses to a given template through two receptive fields. Let $S_0 = \{\langle T_x I, t \rangle \ : \ x \in [0, a + c]\}$, so that $S, S_c \subset S_0$ for all c. We assume S_0, S, S_c to be closed intervals for all c. Then recall that a bijective function (in this case η_h) is strictly monotonic on any closed interval, so the difference of integrals in eq. (A.19) is zero if and only if $S = S_c$. Since we are interested in considering all the values of c up to some maximum C, we can consider the condition

$$\langle T_x I, t \rangle = \langle T_x T_a I, t \rangle \, , \ \forall \, x \in [0, c], \ c \in [0, C]. \tag{A.45}$$

This condition can be satisfied in two cases: (1) both dot products are zero, which is the localization condition, or (2) $T_a I = I$ (or equivalently, $T_a t = t$), where the image or template is periodic. A similar reasoning applies to the case of scale transformations.

In the next section we show how localization conditions for scale and translation transformations imply a specific form of the templates.

A.7.2 Wavelet Templates from Best Invariance and Heisenberg Principle

In this section we identify G_0 with subsets of the affine group. In particular, we study separately the case of scale and translations (in 1-D for simplicity). It is helpful to assume that all images I and templates t are strictly contained in the range of translation or scale pooling, P, since image components outside it are not measured. We consider images I *restricted to* P. For translation this means that the support of I is contained in P. For scaling, since $g_s I = I(sx)$ and $\widehat{I(sx)} = (1/s)\hat{I}(\omega/s)$ (where $\hat{\cdot}$ indicate the Fourier transform), assuming a scale pooling range of $[s_m, s_M]$, implies a range $[\omega^I_m, \omega^I_M]$, $[\omega^t_m, \omega^t_M]$ (m and M indicate maximum and minimum) of spatial frequencies for the maximum support of I and t. Because of theorem A.8 *invariance to translation requires spatial localization of images and templates*, and *invariance to scale requires bandpass properties of images and templates*. Thus images and templates are assumed to be localized from the outset in either space or frequency. The following corollaries show that a stricter localization condition is needed for invariance and that this condition determines the form of the template. Note that in our framework images and templates are bandpass because of being zero-mean. In addition, neural images, which are input to the hierarchical architecture, are spatially bandpass because of retinal processing.

We now state the result of theorem A.8 for one-dimensional signals under the translation group and separately under the dilation group.

Let $I, t \in L^2(\mathbb{R})$, $(\mathbb{R}, +)$ be the one-dimensional locally compact group of translations and $T_x : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ be a unitary representation of the translation operator. For instance, $G_0 = [-b, b]$, $b > 0$, and suppose $\text{supp}(t) \subseteq \text{supp}(I) \subseteq [-b, b]$. Further suppose $\text{supp}(\langle T_x I, t\rangle) \subseteq [-b, b]$. Then we have the following.

Corollary A.21 *Localization in the spatial domain is necessary and sufficient for translation invariance.*

For any fixed $t, I \in \mathcal{X}$, we have

$$\mu^k_h(I) = \mu^k_h(T_x I),\ \forall x \in [0, \bar{x}] \ \Leftrightarrow\ \langle T_x I, t\rangle \neq 0,\ \forall x \in [-b + \bar{x}, b] \tag{A.46}$$

with $\bar{x} > 0$.

Similarly, let $G = (\mathbb{R}^+, \cdot)$ be the one-dimensional locally compact group of dilations, and denote by $D_s : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ a unitary representation of the dilation operator. Let $G_0 = [1/S, S]$, $S > 1$ and suppose $\text{supp}(\langle D_s I, t\rangle) \subseteq [1/S, S]$. Then eq. (A.8) gives the following.

Corollary A.22 *Localization in the spatial frequency domain is necessary and sufficient for scale invariance.*

For any fixed $t, I \in \mathcal{X}$, we have

$$\mu_h^k(I) = \mu_h^k(D_s I), s \in [1, \bar{s}] \;\; \Leftrightarrow \;\; \langle D_s I, t \rangle \neq 0, \; \forall s \in [\frac{\bar{s}}{S}, S] \tag{A.47}$$

with $S > 1$.

Localization conditions of the support of the dot product for translation and scale are depicted in figure A.5.

As shown by lemma A.23, eqs. (A.46) and (A.47) give interesting conditions on the supports of t and its Fourier transform $\hat{t}$. For translation the corollary is equivalent to zero overlap of the compact supports of I and t. In particular, using theorem A.5, for $I = t$ the maximal invariance implies the following localization conditions on t,

$$\langle gt, t \rangle = 0 \;\; g \notin G_L \subseteq G, \tag{A.48}$$

which we call self-localization. For 1-D translations it has the simple form $\langle T_x t, t \rangle = 0 \; |x| > a, \; a > 0$.

For scaling we consider the support of the Fourier transforms of I and t. The Parseval theorem allows rewriting the dot product $\langle D_s I, t \rangle$, which is in $L^2(\mathbb{R}^2)$, as $\langle \widehat{D_s I}, \hat{t} \rangle$ in the Fourier domain.

In the following we suppose that the supports of $\hat{t}$ and $\hat{I}$ are, respectively, $[\omega_m^t, \omega_M^t]$ and $[\omega_m^I, \omega_M^I]$, where $\omega_m^{t,I}$ could be very close to zero (images and templates are supposed to be zero-mean) but usually are larger than zero. Note that the effect of scaling I with (typically $s = 2^j$ with $j \leq 0$) is to change the support to $\text{supp}(\widehat{D_s I}) = s(\text{supp}(\hat{I}))$. This change of the support of $\hat{I}$ in the dot product $\langle \widehat{D_s I}, \hat{t} \rangle$ gives nontrivial conditions on the intersection with the support of $\hat{t}$ and therefore on the localization with respect to the scale invariance. We have the following lemma.

Lemma A.23 Invariance to translation in the range $[0, \bar{x}], \; \bar{x} > 0$ is equivalent to the localization condition of t in space

$$\text{supp}(t) \subseteq [-b - \bar{x}, b] - \text{supp}(I), \; I \in \mathcal{X}. \tag{A.49}$$

Separately, invariance to dilation in the range $[1, \bar{s}], \; \bar{s} > 1$ is equivalent to the localization condition of $\hat{t}$ in frequency ω

$$\begin{aligned} &\text{supp}(\hat{t}) \subseteq [-\omega_t - \Delta_t^*, -\omega_t + \Delta_t^*] \cup [\omega_t - \Delta_t^*, \omega_t + \Delta_t^*] \\ &\Delta_t^* = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S}, \omega_t = \frac{\omega_M^t - \omega_m^t}{2}. \end{aligned} \tag{A.50}$$

Proof: To prove that $\text{supp}(t) \subseteq [-b + \bar{x}, b] - \text{supp}(I)$, note that eq. (A.46) implies that $\text{supp}(\langle T_x I, t\rangle) \subseteq [-b + \bar{x}, b]$ (see figure A.5*a*). With $\text{supp}(\langle T_x I, t\rangle) = \text{supp}(I * t) \subseteq \text{supp}(I) + \text{supp}(t)$, we have $\text{supp}(t) \subseteq [-b - \bar{x}, b] - \text{supp}(I)$. To prove the condition in eq. (A.50),note that eq. (A.47) is equivalent in the Fourier domain to

$$\langle D_s I, t\rangle = \langle \widehat{D_s I}, \hat{t}\rangle = \frac{1}{s}\int d\omega\, \hat{I}\big(\frac{\omega}{s}\big)\hat{t}(\omega) \neq 0 \;\; \forall s \in [\frac{\bar{s}}{S}, S]. \tag{A.51}$$

This is depicted in figure A.5b' for S large enough. In this case we can suppose the support of $\widehat{D_{\bar{s}/S} I}$ to be on an interval to the left of $\text{supp}(\hat{t})$, and $\widehat{D_S I}$ to the right. The condition $\text{supp}(\langle \widehat{D_s I}, \hat{t}\rangle) \subseteq [\bar{s}/S, S]$ is in this case equivalent to

$$\omega_M^I \frac{\bar{s}}{S} < \omega_m^t, \;\; \omega_M^t < \omega_m^I S, \tag{A.52}$$

which gives

$$\Delta_t^* = \text{Max}(\Delta_t) \equiv \text{Max}\Big(\frac{\omega_M^t - \omega_m^t}{2}\Big) = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S}. \tag{A.53}$$

and therefore eq. (A.50).

Note that for some $s \in [\bar{s}/S, S]$ the condition that the Fourier supports are disjoint is only sufficient and not necessary for the dot product to be zero, since cancellations can occur. However, to have $\langle \widehat{D_s I}, \hat{t}\rangle = 0$ on a continuous interval of scales (except some pathological examples of the function I) implies disjointness of the supports, since $\hat{I}(\omega/s) \neq \hat{I}(\omega/s')$, $s \neq s'$, unless I has constant spectrum in the interval $[\bar{s}/S, S]$. Similar reasoning is valid for the translation case.

The preceding results lead to a statement connecting invariance with localization of the templates.

Theorem A.24 Maximum translation invariance implies a template with minimum support in the space domain (x). Maximum scale invariance implies a template with minimum support in the Fourier domain (ω).

Proof: We first illustrate the statement of the theorem with a simple example. In the case of translations, suppose, for instance, $\text{supp}(I) = [-b, b']$, $\text{supp}(t) = [-a, a]$, $a \leq b' \leq b$. Then eq. (A.49) reads

$$[-a, a] \subseteq [-b + \bar{x} + b', b - b'],$$

which gives the condition $-a \geq -b + b' + \bar{x}$, that is, $\bar{x}^{\max} = b - b' - a$. Thus for any fixed b, b', the smaller the template support $2a$ in space, the greater is translation invariance. Similarly, in the case of dilations, increasing the range

of invariance $[1, \bar{s}]$, $\bar{s} > 1$, implies a decrease in the support of $\hat{t}$, as shown by eq. (A.53). In fact, noting that $|\text{supp}(\hat{t})| = 2\Delta_t$, we have

$$\frac{d|supp(\hat{t})|}{d\bar{s}} = -\frac{2\omega_M^I}{S} < 0,$$

that is, the measure $|\cdot|$ of the support of $\hat{t}$ is a decreasing function with respect to the measure of the invariance range $[1, \bar{s}]$.

Because of the assumption of maximum possible support of all I being finite, there is always localization for any choice of I and t under spatial shift. Of course, if the localization support is larger than the pooling range there is no invariance. For a complex cell with pooling range, $[-b, b]$ in space, only templates with self-localization smaller than the pooling range make sense. An extreme case of self-localization is $t(x) = \delta(x)$, corresponding to maximum localization of tuning of the simple cells.

A.8 Transformations with Lie Group Local Structure: HW Module

In this section we generalize the reasoning of chapter 1 for groups to transformations that generate manifolds with a high degree of smoothness.

Let $I \in \mathcal{X}$. Let $s : \mathcal{X} \times \mathbb{R}^N \to \mathcal{X}$, be a C^∞ transformation depending on $\Theta = (\theta_1, \cdots, \theta_N)$ parameters. For any fixed $I \in \mathcal{X}$, the set $M = (s(I, \Theta),\ \Theta \in \mathbb{R}^N)$ describes a differentiable manifold.

If we expand the transformation around, for instance, $\vec{0}$, we have

$$s(I, \Theta) = s(I, \vec{0}) + \sum_{i=1}^{N} \frac{\partial s(I, \Theta)}{\partial \theta_i}\theta_i + o(\|\Theta\|^2) = I + \sum_{i=1}^{N} \theta_i L_{\theta_i}(I) + o(\|\Theta\|^2), \tag{A.54}$$

where L_{θ_i} are the infinitesimal generators of the transformation in the i^{th} direction. The associated transformation can be expressed by exponentiation as

$$g(\Theta) = \exp(\theta_1 L_{\theta_1} + \theta_2 L_{\theta_2} + \cdots + \theta_N L_{\theta_N}).$$

Consider now the signature

$$\mu^{h,k}(I) = \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle s(I, \Theta), t^k \right\rangle). \tag{A.55}$$

If the support of the template is localized enough with respect to that of the image, we have (suppose for simplicity the generators L_{θ_i} are self-adjoint)

$$\mu^{h,k}(I) = \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle s(I,\Theta), t^k \right\rangle \approx \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle g(\Theta)I, t^k \right\rangle$$
$$= \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle I, g(\Theta)t^k \right\rangle.$$

If the representation of the group associated to the operators $g(\Theta)$ is square integrable (a continuous irreducible unitary representation $\pi(g) = g(\Theta)$ of a locally compact group G in a Hilbert space H such that for some nonzero vector ψ the function, $g \to \langle g(\Theta)\psi, \psi \rangle$, is square integrable with respect to the Haar measure on G), then $g(\Theta)t^k$ is the associated wavelet frame. The signature calculated in this way is invariant and selective (see chapter 1).

A.9 Factorization of Invariances

Initially we thought that a signature invariant to a group of transformations could be obtained by factorizing in successive layers the computation of signatures invariant to a subgroup of the transformations (e.g., the subgroup of translations of the affine group) and then adding the invariance with respect to another subgroup (e.g., rotations). In the following we show that while factorization of invariance ranges is possible in a hierarchical architecture (theorem A.17), in general the factorization in successive layers for instances of invariance to translation followed by invariance to rotation (by subgroups) is actually impossible. The mathematical reason is that subgroup averages destroy the covariance structure, factoring out the transformation.

We begin by considering invariance with respect to 2-D translations. We show how invariance can be achieved by first computing an invariant signature with respect to the x-translation and the y-translation, and discuss how a similar strategy can be used with more general groups (and subgroups). We then show that this factorization cannot be used to learn separately x and y invariance through independent templates.

Consider an architecture composed of two layers. Each layer has an associated set of templates, $t^1, \ldots, t^K \in \mathcal{X}$ and $\bar{t}^1, \ldots, \bar{t}^K \in \mathcal{X}$ (we assume for simplicity the number of templates is the same). We also assume we have just one nonlinearity per layer and that the set of translations has been suitably discretized. We denote by T_x, T_y, $x, y \in \mathbb{R}$, the action of the translations on any $I \in \mathcal{X}$.

The first layer defines an invariant signature $\mu : \mathcal{X} \to \mathbb{R}^K$ as

$$\mu_1^k(I) = \sum_{x' \in \mathbb{R}} \eta(\left\langle T_{x'} I, t^k \right\rangle), \quad k = 1, ..., K, \tag{A.56}$$

where $\langle T_{x'} I, t^k \rangle = \left\langle I, T_{x'}^{-1} t^k \right\rangle$, since the representation of the translation is unitary. This means that we can obtain the signature by looking at transforming images or transforming templates.

The signature at second layer is $\nu : \mathcal{X} \to \mathbb{R}^K$,

$$\mu_2^l(I) = \sum_{y' \in \mathbb{R}} \eta(\left\langle \mu_1(T_{y'} I), s^l \right\rangle), \quad l = 1, ..., K, \tag{A.57}$$

where the set of templates $s^1, \ldots, s^K$ can be thought of as the signatures of a set of templates with respect to the first layer, that is, $s^i = \mu_1(\bar{t}^i)$ for $\bar{t}^i \in \mathcal{X}$, $i = 1, \ldots, K$, with $s^i \in \mathbb{R}^K$.

Indeed, we can show that ν is invariant to 2-D translations $\mu_2(I) = \mu_2(T_x T_y I)$. We note that $\mu(T_x T_y I) = \mu(T_y T_x I) = \mu(T_y I)$, since μ is defined by a group integration with respect to the x-translation and the x and y translation operators commute. Then

$$\mu_2^l(T_x T_y I) = \sum_{y' \in \mathbb{R}} \eta(\left\langle \mu_1(T_{y'} T_y I), s^l \right\rangle) = \mu_2^l(T_y I), \quad l = 1, \ldots, K, \tag{A.58}$$

and finally $\mu_2^l(T_x T_y I) = \mu_2^l(T_y I) = \mu_2(I)$, since μ_2 is defined by a group integration with respect to the y-translation.

An inspection of the reasoning shows that a similar factorization holds for many transformations beyond translations. Indeed, we have the following result.

Lemma A.25 (Factorization of Invariances) Let G, R be two locally compact groups with Haar measures dg, dr, respectively. Let $\mu : \mathcal{X} \to \mathbb{R}^K$ and $\nu : \mathcal{X} \to \mathbb{R}^L$ be defined by

$$\mu_1^k(I) = \int dg \eta(\left\langle gI, t^k \right\rangle), \qquad \mu_2^l(I) = \int dr \eta(\left\langle \mu_1(rI), s^l \right\rangle); \tag{A.59}$$

then $\mu_2(grI) = \mu_2(I)$.

Remark A.26 We briefly comment on the order in which transformations need to be applied to an image to still have invariance. Clearly, factorization of invariances happens in the architecture with an order given by the group integration at each layer, so in general it might not be true that $\mu_2(rgI) = \mu_2(I)$. However, invariance clearly still holds if the groups actions commute, $rg = gr$. Moreover it also holds under the weaker requirement that for all

$(g,r) \in G \times R$ there exist $(g',r') \in G \times R$ such that $rg = g'r'$. The latter property holds, for example, if we take G,R to be 2-D translations and rotations, respectively. This is because the semidirect product of the Abelian group $\mathbb{R}^2$ (which describes translations) and the group $SO(2)$ of orthogonal 22 matrices (which describes rotations and reflections that keep the origin fixed) is isomorphic to the Euclidean group. The group of translations is a normal subgroup of the Euclidean group, and the Euclidean group is a semidirect product of the translation group and $SO(2)$.

A more interesting result would be if μ_2 of eq. (A.57) were covariant in the following sense: $\langle \mu_1(g_{y'}I), s^k \rangle) = \left\langle \mu_1(I), \mu_1(g_{y'}^{-1}\bar{t}^i) \right\rangle$, which we rewrite as

$$\left\langle \mu_1(rI), \mu_1(r\bar{t}) \right\rangle = \left\langle \mu_1(I), \mu_1(\bar{t}) \right\rangle, \tag{A.60}$$

where r is the group element corresponding to $g_{y'}$. If this were true, then one invariance could be learned via the templates t^k and another could be learned in the next layer via the templates s^i. This covariance property, however, cannot be expected for general η.

To prove the statement, we first sketch the linear situation: η is the identity function. We also assume that the set of t^k is an orthonormal basis. In this case,

$$\begin{aligned} \left\langle \mu_1(rI), \mu_1(r\bar{t}) \right\rangle &= \sum_k \int dg \left\langle grI, t^k \right\rangle \int dg' \left\langle g'r\bar{t}, t^k \right\rangle \\ &= \int dg \int dg' \sum_k \left\langle grI, t^k \right\rangle \left\langle g'r\bar{t}, t^k \right\rangle, \end{aligned}$$

and if the transformations g and r do commute, we have

$$\begin{aligned} \left\langle \mu_1(rI), \mu_1(r\bar{t}) \right\rangle &= \int dg \int dg' \left\langle grI, g'r\bar{t} \right\rangle \\ &= \int dg \int dg' \left\langle gI, g'r^{-1}r\bar{t} \right\rangle \\ &= \int dg \int dg' \left\langle gI, g'\bar{t} \right\rangle \\ &= \left\langle \mu_1(I), \mu_1(\bar{t}) \right\rangle. \end{aligned} \tag{A.61}$$

Note that if the transformations g and r do not commute, the result in eq. (A.61) does not hold. Even a weaker result, $\left\langle \mu(rI), \mu(r\bar{t}) \right\rangle = \left\langle \mu(I), \mu(r'\bar{t}) \right\rangle$ for some r', does not hold. In fact, using remark A.26, we have that for each fixed g, g', r there exists $\tilde{g}, \hat{g}, r'(g,g')$ such that

$$\left\langle grI, g'r\bar{t} \right\rangle = \left\langle \tilde{g}I, \hat{g}r'(g,g')\bar{t} \right\rangle.$$

However, r' is different for any choice of g, g'. Therefore we can neither obtain the same result as in the commutative case nor have a weaker form of it,

$\langle \mu_1(rI), \mu_1(r\bar{t}) \rangle = \langle \mu_1(I), \mu_1(r'\bar{t}) \rangle$ for some r'. Another important case when eq. (A.60) does not hold is when η is a nonlinear function. we now sketch another case of practical interest in which $\eta(x) = x^2$. We make the same simplifying assumptions. Then

$$\begin{aligned} \langle \mu_1(rI), \mu_1(r\bar{t}) \rangle &= \sum_k (\int dg \left\langle grI, t^k \right\rangle)^2 (\int dg' \left\langle g' r\bar{t}, t^k \right\rangle)^2 \\ &= \int dg \int dg' \sum_k (\left\langle grI, t^k \right\rangle)^2 (\left\langle g' r\bar{t}, t^k \right\rangle)^2. \end{aligned}$$

It can be easily proved that this expression does not satisfy eq. (A.60) even in the case of commutative groups. Thus *for the specific network* here, factorization holds only for functions η that are linear (or equivalently, only for the first moment of the distribution) and only for commutative groups.

In this section we analyzed in detail in which sense a hierarchical architecture achieves larger and larger invariance with growing layer numbers. We found that it is impossible to achieve invariance with respect to different subgroups at different layers and that only invariance per ranges is possible.

A.10 Invariant Representations and Their Cost in Number of Orbit Elements

In the model described so far, implementing an invariant (and discriminative) representation has a *memory cost* in terms of number of templates and their transformations that we need to store to calculate an invariant signature. Here we calculate this cost for two different architectures, hierarchical and shallow (figure A.6), and show how a hierarchical architecture can calculate an invariant signature with far fewer transformed templates than a shallow architecture can.

The intuition is the following. Fewer template transformations are needed to compute a certain invariance at layer ℓ in the case of an invariant hierarchical representation because the dimensionality of the space of possible images is reduced owing to the inherited invariance from the layers below.

More precisely, consider the signature

$$\mu_\ell^{h,k}(I)(g) = \int_{gG_\ell} d\bar{g}\ \eta_h(\left\langle \mu_{\ell-1}(I), \mu_{\ell-1}(\bar{g}t^k) \right\rangle), \quad k = 1, \cdots, K. \tag{A.62}$$

As proved previously, $\mu_\ell^{h,k}(I)$ is an invariant and discriminative vector. The amount of invariance (which we call $\bar{G}_\ell$), as proved in theorem A.8, depends on the extension of the support of the dot product $\langle \mu_{\ell-1}(I), \mu_{\ell-1}(\bar{g}t^k) \rangle$ with

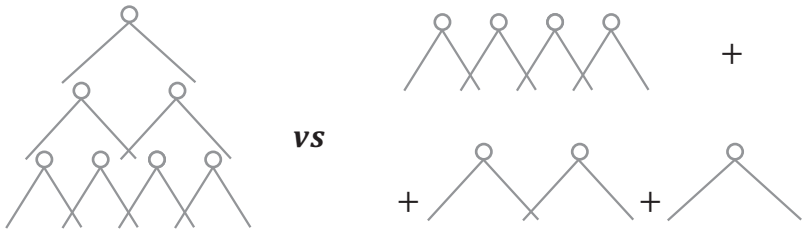


Figure A.6
Hierarchical architecture versus multiple one-layer architectures.

respect to the pooling range G_ℓ. As is clear from eq. A.62, building invariant representations has a cost in terms of the template transformations needed to compute the invariance and a cost in terms of the number of different templates (projections, K) if we want the representation to be discriminative. In the following we consider the cost of hierarchical invariant representations versus multiple one-layer representations achieving the same invariances for whole images and their parts and with the same discriminative power (see figure A.6).

Suppose, instead of a single-layer $\bar{G}$-invariant representation, we construct a sequence of increasingly invariant representations (at different layers of a hierarchy) into the tower of transformation subsets (figure A.7):

$$\bar{G}_1 \subset \bar{G}_2 \subset \cdots \subset \bar{G}_L = \bar{G} \subset G.$$

This construction not only achieves $\bar{G}$ invariance at the top layer for the whole image but also local and *independent* $\bar{G}_\ell$ invariances for (bigger and bigger) parts of the image. If we wanted to achieve the same kind of invariances with a single-layer architecture, we would need to have L one-layer architectures that implement separately the $\bar{G}_\ell$, $\ell = 1, \cdots, L$, invariances for bigger and bigger parts of the image. In the following we mathematically show that the cost (in terms of transformed templates) is far less with a hierarchical architecture than with multiple one-layer architectures. In the next section we consider the discriminative properties of the two architectures and prove that the number of templates to achieve the same discriminability is the same in the two models.

A.10.1 Discriminability Cost for Hierarchical versus Single-Layer Architecture

For simplicity we study the case of a non-Abelian locally compact discrete group of transformations G and its associated wavelet transform (e.g., $g_{j,k} \in G$,

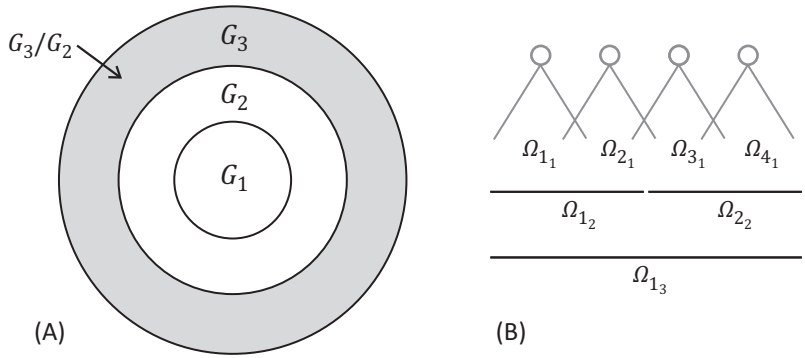


Figure A.7

(A) Tower of subsets of transformations. (B) Effective (at the image level) partitions of the image support (corresponding to different image parts) at different layers of a hierarchical architecture.

$j,k \in \mathbb{Z}$, such that $g_{j,k}f(x) = 2^{j/2}f(2^j x - k)$, $f \in L^2(\mathbb{R})$). It is well known that the number of random projections needed to distinguish among n images with ϵ precision is proportional to $\ln(n)$ and that this number does not depend on the image dimensionality by the Johnson-Lindenstrauss lemma. In section A.2 we proved that if n is the cardinality of a set of different images (none is the transform of another; G-invariant) that must be we have to distinguished with probability $1 - \delta^2$ and precision ϵ, then an estimated probability distribution obtained with

$$\epsilon(K) = \sqrt{\frac{2}{Kc} \ln\left(\frac{n}{\delta}\right)}$$

is sufficient.

Translated in the context of a hierarchical construction this means that the set of images at each module at layer ℓ that must be distinguished are all possible $\bar{G}_\ell$-invariant images, n_ℓ. This is difficult to calculate, but we can prove that the number of equivalence classes in the case of parallel architecture and hierarchical architecture is the same at each layer and part. The only difference is that in the case of a hierarchy the space of the signals at layer ℓ is already $\bar{G}_{\ell-1}$ invariant.

In general, n_ℓ depends, at each module of layer ℓ, on the extension of the support of $\langle \mu_\ell(I), \mu_\ell(gt^k) \rangle$ with respect to the pooling range G_ℓ. In particular, the invariance range is given by

$$\bar{G}_\ell = \{\hat{g} \in G \mid \text{supp}(\left\langle \mu_\ell(\hat{g}I), \mu_\ell(gt^k) \right\rangle \subseteq G_\ell\}.$$

However, note that

$$\text{supp}_{g\in G_\ell}(\langle\mu_\ell(I),\mu_\ell(gt^k)\rangle) = \text{supp}_{g\in G_\ell}(\langle I, gt^k\rangle).$$

Thus we have that $\bar{G}_\ell = G_\ell$ and therefore that the number of G_ℓ ($\bar{G}_\ell$)-invariant images is the same for the two architectures, namely, $k_\ell = \bar{k}_\ell$, with the pooling range, G_ℓ, the same. In other words, we proved that the number of templates per layer and per module to achieve discriminability among the $\bar{G}_\ell$-invariant images is the same in the case of hierarchical (k_ℓ) and parallel ($\bar{k}_\ell$) architectures.

A.10.2 Invariance Cost for Hierarchical versus Single-Layer Architecture

For simplicity, we again consider a model where the transformations are those of the discrete non-Abelian locally compact group of transformations G associated to the discrete wavelet transform in $L^2(\mathbb{R})$.

Theorem A.27 Let $I \in \mathcal{X}^\Omega$ and $\Omega_{i_\ell} \subset \Omega$, $i_\ell = 1,\cdots,M_\ell$, $\ell = 1,\cdots,L$ (see figure A.7, B). Let Ω_{i_ℓ} be such that $\bigcup_{i_\ell=1}^{M_\ell} \Omega_{i_\ell} = \Omega$, $\bigcap_{i_\ell=1}^{M_\ell} \Omega_{i_\ell} = \emptyset$, $\forall\, \ell = 1,\cdots,L$, and suppose the sets are all of the same size per fixed layer. Let $\bar{G}_1 \subset \bar{G}_2 \subset \cdots \subset \bar{G}_L = \bar{G} \subset G$ be a tower of proper subsets of G. Let $c(1)$, $c(L)$ be the costs respectively, of a multiple one-layer and of a hierarchy composed of L layers architecture for achieving the same $\bar{G}_\ell$-invariances (for image parts with support in Ω_{i_ℓ}). We have

$$r = \frac{c(L)}{c(1)} = \frac{\sum_{\ell=1}^{L}\sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell}{\sum_{\ell=1}^{L}\sum_{i_\ell=1}^{M_\ell} |G_\ell|k_\ell}, \tag{A.63}$$

where M_ℓ and k_ℓ are, respectively, the number of moduli and the number of templates at layer ℓ, and G_ℓ is the pooling range at each layer of the hierarchical architecture.

Proof: The first layer is by construction implementing a $\bar{G}_1$ invariance. Thus we need $G_1 \supseteq \bar{G}_1$ transformations per each of the k_1 templates and M_1 modules (assuming the same number of templates at each module),

$$c_1 = \sum_{i_1=1}^{M_1} |G_1|k_1,$$

where M_1 is the number of parts (modules) of the image I we consider at the first layer.

For the second layer,

$$c_2 = \sum_{i_1=1}^{M_2} |G_2/G_1|k_2.$$

The key point is that to implement the G_2 invariance we need fewer transformations, G_2/G_1, since the representation at the first layer is already G_1 invariant). This is true, since $|\text{supp}(\langle t^k, gt^k \rangle)| = 1$ and therefore $\tilde{G}_1 = G_1$.

In general, at layer ℓ,

$$c_\ell = \sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell, \ \ell = 1, \ldots, L.$$

Summing the contributions at each layer, we have for the hierarchical construction

$$c(L) \propto \sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell. \tag{A.64}$$

For the corresponding multiple single-layer construction we have instead

$$c_\ell = \sum_{i_\ell=1}^{M_\ell} |G_\ell|\tilde{k}_\ell, \ \ell = 1, \ldots, L,$$

where, as demonstrated in the previous section, $\tilde{k}_\ell = k_\ell$. Note that in this case to implement a G_ℓ invariance at layer ℓ we have to use a bigger set of transformations, G_ℓ (versus $G_\ell/G_{\ell-1}$ of the hierarchy), since no invariance is inherited from the layers below. Thus,

$$c(1) = \sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell|k_\ell, \tag{A.65}$$

which gives a ratio

$$r = \frac{c(L)}{c(1)} = \frac{\sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell}{\sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell|k_\ell}. \tag{A.66}$$

The theorem proves that far fewer stored transformations of templates are needed to achieve invariance for a group transformation G with hierarchical architecture than with shallow architecture.

A.11 Nonlinearities Are Key in Gaining Compression from Repeated Random Projections

The selectivity property of the invariant representation proposed here is guaranteed by an adequate number of projections onto random templates (or PCs) (see section A.2). In the following we prove that repeating this projection step does not provide any computational advantage unless some nonlinearities came into play.

A.11.1 Hierarchical (Linear) Johnson-Lindenstrauss Lemma

Let $\{x_1, x_2, \ldots, x_n\}$ be a set of vectors in $\mathbb{R}^d$. The J-L lemma says that for any $0 < \epsilon < 1/2$ there exists an $f : \mathbb{R}^d \to \mathbb{R}^k$ for $k = O(\epsilon^{-2}) \log(n)$ such that

$$\forall\, i,j\ (1-\epsilon) \left\|x_i - x_j\right\|_2 \leq \left\|f(x_i) - f(x_j)\right\|_2 \leq (1+\epsilon) \left\|x_i - x_j\right\|_2 .$$

The J-L lemma is equivalent to the following statement [132].

Lemma A.28 For any $0 < \epsilon < 1/2$ and positive integer d, there exists a distribution $\mathcal{D}$ over $\mathbb{R}^{k \times d}$ for $k = O(\epsilon^{-2} log(1/\delta))$ such that for any $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$,

$$Pr_{A \sim \mathcal{D}}[\,|\, \|Ax\|_2^2 - 1| > \epsilon] < \delta.$$

We want to formulate first a local version of the J-L lemma and then a recursive/hierarchical one. For the local version suppose $x \in \mathbb{R}^{d \times p}$. Let $A = \mathbb{I}_p \otimes A$ and $x_k = P_k x$ be the projection onto the kth block of d-components of the x vector, $k = 1, \ldots, p$. If we suppose $Pr_{A \sim \mathcal{D}}[\,|\, \|Ax_k\|_2^2 - 1| > \epsilon] < \delta \ \ \forall k$, then also

$$Pr_{A \sim \mathcal{D}}[\,|\, \left\|(\mathbb{I}_p \otimes A)x\right\|_2^2 - 1| > \epsilon] < \delta.$$

Therefore instead of applying the J-L lemma to the whole vector x, we can write it as $x = x_1 \oplus x_2 \oplus \ldots \oplus x_p$ and apply it to each of the p parts, and the embedding still holds. This result says that there is no difference between the applying J-L lemma locally or globally. We can then repeat the procedure on the resulting vector $y \in \mathbb{R}^{k \times p}$. However, the number of different vectors obtained after the projection will be still n, and consequently we will need again the same number of projections, so no gain.

Consider now a different situation in which the number of images depends on the vector size. In this case, we will have an effective gain in repeating the J-L projections.

A.11.2 Example of Nonlinear Extension of Johnson-Lindenstrauss Lemma

Suppose we have an architecture composed of two-layers; the first layer is composed of P parts, each of d pixels, and the second layer of one part of P pixels. Suppose also we have a one-layer architecture composed of Pd pixels. We want to compare the number of templates needed to discriminate among the different possible images that are inputs to the two architectures. The reasoning holds for any choice of a dictionary. We choose a PC dictionary as a possible example.

In order to use the J-L result we have to know the number of possible images for a certain image size s. We can say that they correspond to the number of all possible combinations of the pixels at different (finite) gray levels. Here, however, we adopt a different criterion. Suppose we have N images of size s (N can be as big as we want and in the limit goes to infinity). The criterion we adopt is the following. The number of possible images of size s is the number of images we can distinguish (up to an ϵ factor) and is given by the number of possible images we can build using a number of PCs (extracted from the N images of size s, $N \to \infty$) such that they allow a reconstruction up to an ϵ error; we call this number $f(s)$. If we further suppose that the PCs are linearly combined with coefficients that can take a finite (but arbitrarily large) number of values, say V, we have that the number of possible images with size s is given by

$$V^{f(s)}.$$

Let us now go back to the one-layer versus the two-layer comparison. If we assume that according to the J-L lemma the number of templates is given (up to a constant) by the log of the number of possible images, we have that the number of templates for the one-layer architecture is

$$\ln(V^{f(Pd)})$$

and for the two-layers architecture

$$P\ln(V^{f(d)}) + \ln(V^{f(P)}).$$

Suppose we can reuse the template projections for the different parts. Given that the log and the power are monotonically increasing functions, we are left with the comparison

$$f(d) + f(P) \lesseqgtr f(dP).$$

From a look at the plot of the number of PCs versus eigenvalues values at different sizes (see [9], figure 3), we have

$$\lambda(i,s) \approx e^{-\frac{i}{s}},$$

where the i axis is the count of the number of PCs. Reconstruction up to an ϵ factor will correspond to a constant $T(\epsilon)$ such that $\lambda < T(\epsilon)$, that is, with a simple inversion of the formula we have that the number of PCs that allows a reconstruction with an error up to ϵ is

$$f(s) = -\text{Round}\ln(T(\epsilon))s).$$

Neglecting for the moment the round approximation, we have that the comparison between the one- and two-layer architecture boils down to

$$\text{(two layers)} \quad d + P \lesseqgtr Pd \quad \text{(one layer)}.$$

For $P > 1$ (more then one part) we have that $d + P < Pd$ is always true, that is, there are fewer templates for the two-layer architecture than for the one-layer architecture. Any choice of dictionary that allows the same reconstruction error with fewer dictionary elements will give a similar result. The underlying idea is that images have structure and can be expressed with a dictionary of cardinality much lower than the number of pixels.

Remark A.29 This is an application of the J-L lemma in a nonlinear setting. In a hierarchical architecture the nonlinearity is the indicator function on the parts.

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