M-theory:
unsupervised learning of hierarchical invariant representations

tomaso poggio
CBMM
McGovern Institute,
BCS,
LCSL, CSAIL
MIT

Thursday, December 5, 13
1. Motivation: models of cortex (and deep convolutional networks)

2. Core theory
   - the basic invariance module
   - the hierarchy

3. Computational performance

4. Biological predictions

5. Theorems and remarks
   - $n \to 1$
   - invariance and sample complexity
   - connections with scattering transform
   - invariances and beyond perception
   - ...

Thursday, December 5, 13
Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)
Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)
Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)
Motivation: feedforward models of recognition in Visual Cortex
(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)


Thursday, December 5, 13
Motivation: feedforward models of recognition in Visual Cortex

(Hubel and Wiesel + Fukushima and many others)

*Modified from (Gross, 1998)


software available online with CNS (for GPUs)
Hierarchical, Hubel and Wiesel (HMAX-type) models work well, as model of cortex and as computer vision systems but...why? and how can we improve them?

Similar convolutional networks called deep learning networks (LeCun, Hinton,...) are unreasonably successful in vision and speech (ImageNet+Timit)... why?
Hierarchical, Hubel and Wiesel (HMAX-type) models work well, as model of cortex and as computer vision systems but...why? and how can we improve them?

Similar convolutional networks called deep learning networks (LeCun, Hinton,...) are unreasonably successful in vision and speech (ImageNet+Timit)...

why?
Collaborators (MIT-IIT, LCSL) in recent work

F. Anselmi, J. Mutch, J. Leibo, L. Rosasco, A. Tacchetti, Q. Liao

+ +

Evangelopoulos, Zhang, Voinea

Plan

1. Motivation: models of cortex (and deep convolutional networks)

2. Core theory
   - the basic invariance module
   - the hierarchy

3. Computational performance

4. Biological predictions

5. Theorems and remarks
   - $n \to 1$
   - invariance and sample complexity
   - connections with scattering transform
   - invariances and beyond perception
   - ...

Thursday, December 5, 13
The main computational goal of the *feedforward* ventral stream hierarchy is to compute a representation for each incoming image which is invariant to transformations previously experienced in the visual environment.

Remarks:

- A *theorem* (T&R) shows that invariant representations may reduce by orders of magnitude the sample complexity of a classifier at the top of the hierarchy
- Empirical evidence (T&R) also supports the claim
- Hypothesis suggests unsupervised learning of transformations
The main computational goal of the feedforward ventral stream hierarchy is to compute a representation for each incoming image which is invariant to transformations previously experienced in the visual environment.

Remarks:

- A *theorem* (T&R) shows that invariant representations may reduce by orders of magnitude the sample complexity of a classifier at the top of the hierarchy.

- Empirical evidence (T&R) also supports the claim.

- Hypothesis suggests unsupervised learning of transformations.
Theorem *(translation case)* Consider a space of images of dimensions $d \times d$ pixels which may appear in any position within a window of size $rd \times rd$ pixels. The usual image representation yields a sample complexity (of a linear classifier) of order $m = O(r^2 d^2)$; the oracle representation (invariant) yields (because of much smaller covering numbers) a much better sample complexity of order

$$m_{oracle} = O(d^2) = \frac{m_{image}}{r^2}$$
Use of invariant representation ---> signature vectors for memory access at several levels of the hierarchy

\[ \sum = \text{signature vector} \cdot \]
Remarks:

- Images can be represented by a set of functionals on the image, e.g. a set of measurements.

- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an “image patch” and another image patch (called template) which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$).

- Projections via dot products are natural for neurons: here simple cells.

Neuroscience definition of dot product!
Neuroscience constraints on image representations

Remarks:

- Images can be represented by a set of functionals on the image, e.g. a set of measurements.

- Neuroscience suggests that natural functionals for a neuron to compute is a high-dimensional dot product between an “image patch” and another image patch (called template) which is stored in terms of synaptic weights (synapses per neuron $\sim 10^2 - 10^5$).

- Projections via dot products are natural for neurons: here simple cells.

$\langle x, t \rangle$  
**Neuroscience definition of dot product!**
Signatures: the Johnson-Lindenstrauss theorem (features do not matter much!)

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 $u, v \in V$

$$(1 - \epsilon) \| u - v \|^2 \leq \| Pu - Pv \|^2 \leq (1 + \epsilon) \| u - v \|^2$$

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

$$kC(\epsilon) \geq \ln(n), \quad C(\epsilon) = \frac{1}{2}\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)$$

JL suggests that good image representations for classification and discrimination of $n$ objects can be provided by $k$ dot products with random templates!
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, ) undergoes all in plane rotations

An histogram of the values of the dot products of with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao
Computing an invariant signature with the HW module
(dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations

An histogram of the values of the dot products of with the image (e.g. a face) is computed.
Histogram gives a unique and invariant image signature
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $gt$

An histogram of the values of the dot products of $t$ with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $gt$

An histogram of the values of the dot products of $gt$ with the image (e.g. a face) is computed.

Histogram gives a unique and invariant image signature

poggio, anselmi, rosasco, tacchetti, leibo, liao

Thursday, December 5, 13
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $gt$

An histogram of the values of the dot products of $gt$ with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature

$poggio, anselmi, rosasco, tacchetti, leibo, liao$
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $gt$

An histogram of the values of the dot products of $gt$ with the image (e.g. a face) is computed.

Histogram gives a unique and invariant image signature

$Hist\langle \text{face}, gt \rangle$
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g., a car, ) undergoes all in plane rotations

Random template could be used instead of car

An histogram of the values of the dot products of with the image (e.g., a face) is computed.

Histogram gives a unique and invariant image signature
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in-plane rotations

Random template could be used instead of car

An histogram of the values of the dot products of $t$ with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $g_t$

Random template could be used instead of car

An histogram of the values of the dot products of with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature.

poggio, anselmi, rosasco, tacchetti, leibo, liao
Computing an invariant signature with the HW module (dot products and histograms of an image in a window).

A template (e.g. a car, $t$) undergoes all in plane rotations $gt$.

Random template could be used instead of car.

An histogram of the values of the dot products of $gt$ with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature.
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $g_t$.

Random template could be used instead of car.

An histogram of the values of the dot products of $g_t$ with the image (e.g. a face) is computed. Histogram gives a unique and invariant image signature.
Computing an invariant signature with the HW module (dot products and histograms of an image in a window)

A template (e.g. a car, $t$) undergoes all in plane rotations $gt$

Random template could be used instead of car

An histogram of the values of the dot products of $gt$ with the image (e.g. a face) is computed.

Histogram gives a unique and invariant image signature

$Hist\langle \cdot, gt \rangle$
This is it

• The basic HW module works for all transformations (no need to know anything about it, just collect unlabeled videos)

• Recipe:
  - memorize a set of images/objects called templates
  - for each template memorize observed transformations
  - to generate a representation/signature invariant to those transformation for each template
    - compute dot products of its transformations with image
    - compute histogram of the resulting values

• The same rule works on many types of transformations:
  - affine in 2D, image blur, image undersampling,...
  - 3D pose for faces, pose for bodies, perspective deformations, color constancy, aging, face expressions,...
Overview of a “deep” theory

• Formal proofs --> exact invariance for generic images under group transformations using the basic HW module with generic templates (it is an invariant Johnson-Lindenstrauss-like embedding)
  - optimal templates for maximum range of simultaneous invariance in position and scale are Gabor functions

• Formal proofs --> approximate invariance under smooth non-group transformations using the same basic HW module with object-class specific templates
Transformation example: affine group

The action of a group transformation on an image $I$ is defined as:

In the case of affine group:
Transformation example: affine group

The action of a group transformation $g$ on an image $I$ is defined as:

In the case of affine group:
Transformation example: affine group

The action of a group transformation $g$ on an image $I$ is defined as:

$$gI(\tilde{x}) = I(g^{-1}\tilde{x})$$

In the case of affine group:
Transformation example: affine group

The action of a group transformation \( g \) on an image \( I \) is defined as:

\[
gI(\vec{x}) = I(g^{-1}\vec{x})
\]

In the case of affine group:

\[
gI(\vec{x}) = I(A^{-1}\vec{x} - \vec{b}), \quad A \in GL(2), \vec{b} \in R^2
\]
Theorems for the compact group

The image orbit and its associated probability distribution is invariant and unique.

For a SINGLE new image invariant and unique signature consisting of 1D distributions:

This “movie” is stored during development:

: set of templates
Theorems for the compact group

\[ I \sim I' \iff O_I = O_{I'}, \iff P_I = P_{I'}, \]

The image orbit and its associated probability distribution is invariant and unique.

For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates
Theorems for the compact group

\[ I \sim I' \iff O_I = O_{I'}, \iff P_I = P_{I'} \]

The image orbit and its associated probability distribution is invariant and unique

\[ P_{(I, t^k)} \]

For a SINGLE new image, invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates
Theorems for the compact group

\[ I \sim I' \iff O_I = O_{I'}, \iff P_I = P_{I'} \]

The image orbit and its associated probability distribution is invariant and unique

\[ P_I \quad \leftrightarrow \quad P_{\langle I, t^k \rangle} \]

For a SINGLE new image invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates
Theorems for the compact group

\[ I \sim I' \iff O_I = O_{I'}, \iff P_I = P_{I'} \]

The image orbit and its associated probability distribution is invariant and unique

\[ P_I \leftarrow P_{\langle I, t^k \rangle} \]

For a SINGLE new image, invariant and unique signature consisting of 1D distributions

This “movie” is stored during development

: set of templates
Theorems for the compact group

\[ I \sim I' \iff O_I = O_{I'}, \iff P_I = P_{I'} \]

\[ P_I \quad \longleftrightarrow \quad P_{\left< I, t^k \right>} \]

\[ \left< gI, t^k \right> = \left< I, g^{-1}t^k \right> \]

The image orbit and its associated probability distribution is invariant and unique.

For a SINGLE new image invariant and unique signature consisting of 1D distributions.

This “movie” is stored during development.

: set of templates
Theorems for the compact group

\[ I \sim I' \iff O_I = O_{I'}, \iff P_I = P_{I'} \]

The image orbit and its associated probability distribution is invariant and unique

\[ P_I \quad \text{and} \quad P_{\langle I, t^k \rangle} \]

For a SINGLE new image invariant and unique signature consisting of 1D distributions

\[ \langle gI, t^k \rangle = \langle I, g^{-1} t^k \rangle \]

This “movie” is stored during development

\[ t^k, k = 1, \ldots, K : \text{set of templates} \]
Theorem: Consider $n$ images $I_j$ in $X_n$. Let

$$K \geq \frac{c}{\varepsilon^2} \log \frac{n}{\delta}$$

where $c$ is a universal constant. Then

$$| d(P_I - P_{I'}) - \hat{d}_K (P_I - P_{I'}) | \leq \varepsilon$$

with probability $1 - \delta^2$, for all $I, I' \in X_n$. 


Our basic machine: a HW module
(dot products and histograms for an image in a receptive field window)

- The signature provided by complex cells at each “position” is associated with histograms of the simple cells activities that is

\[ \mu_n^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + n\Delta) \]

- Related quantities such as moments of the distributions are also invariant, for instance as computed by the energy model of complex cells or the max, related to the sup norm ---> we have a full theory of pooling

- Neural implementation of histograms requires complex cells -- usual neurons with different thresholds

- Histograms provide uniqueness independently of pooling range
Preview: from a HW module to a hierarchy via covariance
Preview: from a HW module to a hierarchy via covariance

HW module

complex cell node gives output of the HW module
Covariance theorem (informal): for isotropic networks the activity at a layer of “complex” cells for shifted an image at position $g$ is equal to the activity induced by the group shifted image at the shifted position.
Covariance theorem (informal): for isotropic networks the activity at a layer of “complex” cells for shifted an image at position $g$ is equal to the activity induced by the group shifted image at the shifted position.

Remarks:

- Covariance allows to consider a higher level HW module, looking at the neural image at the lower layer and apply again the invariance/covariance arguments
Toy example: 1D translation

= 

Thursday, December 5, 13
M-Theory

So far: compact groups in $R^2$

M-theory extend result to

- partially observable groups
- non-group transformations
- hierarchies of magic HW modules (multilayer)
Invariance for POGs implies a localization property we call

**sparsity of the image** $I$ wrt the $t$ dictionary under the set of transformations $G$

Example: consider the case of a 1D parameter translation group: invariance of $\mu_n^k(I)$ with pooling region $[-b,b]$ is ensured if

$$\langle I, g_r t^k \rangle = 0, \quad \text{for } |r| > b - a$$
Invariance, sparsity, wavelets

Thus sparsity implies, and is implied by, invariance. Sparsity can be satisfied in two different regimes:

• exact sparsity for *generic* images holds for affine group.

• approximate sparsity of a subclass of $I$ w.r.t. dictionary of transformed templates $g_t^k$ holds locally for any smooth transformation.
Invariance, sparsity, wavelets

**Theorem:** Sparsity is *necessary and sufficient* condition for translation and scale invariance. Sparsity for translation (respectively scale) invariance is equivalent to the support of the template being small in space or frequency.

**Proposition:** Maximum simultaneous invariance to translation and scale is achieved by Gabor templates:

\[
t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}
\]
M-Theory

M-theory extends result to

• non compact groups

• non-group transformations

• hierarchies of magic HW modules (multilayer)
Non-group transformations: approximate invariance in class-specific regime

\[ \mu^k_n(I) \] is locally invariant if:

- \( I \) is sparse in the dictionary of \( t^k \)
- \( I \) transforms in the same way (belong to the same class) as \( t^k \)
- the transformation is sufficiently smooth
Class specific pose invariance for faces
M-Theory

M-theory extend result to

- non compact groups
- non-group transformations
- hierarchies of magic HW modules (multilayer)
Hierarchies of magic HW modules:
key property is covariance
Local and global invariance: whole-parts theorem

For any signal (image) there is a layer in the hierarchy such that the response is invariant w.r.t. the signal transformation.
Why multilayer architectures

- Compositionality: signatures for wholes and for parts of different size at different locations
- Minimizing clutter effects
- Invariance for certain non-global affine transformations
- Retina to V1 map
Invariance and uniqueness

Figure 3: Two distinct stimuli (left) are presented at various location in the visual field. The Euclidean distance between C2 response vectors in HMAX is reported (right). It can be seen how the response are invariant to global translation and discriminative. The C2 units represent the top of a hierarchical, convolutional architecture.
Invariance for parts and stability for wholes

Figure 4: (a) shows the reference image on the left and a local deformation of it (the eyes are closer to each other); (b) shows that a C1 signature from complex cells whose receptive fields covers the left eye is invariant to the deformation; in (c) C2 cells whose receptive fields contain the whole face are (Lipschitz) stable with respect to the deformation. In all cases just the euclidean norm of the response is shown on the y axis.
Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
   - the basic invariance module
   - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
   - $n \rightarrow 1$
   - invariance and sample complexity
   - connections with scattering transform
   - invariances and beyond perception
   - ...
Implementations/specific models: computational performance

- Deep convolutional networks (such as Lenet) as an architecture are a special case of Mtheory (with just translation invariance and max/sigmoid pooling)
- HMAX as an architecture is a special case of Mtheory (with translation + scale invariance and max pooling) and used to work well
HMAX models perform well at computational level

HMAX models -- a special case of M-theory -- perform well compared to engineered computer vision systems (in 2006) on several databases

Bileschi, Wolf, Serre, Poggio, 2007; Mutch Lowe
HMAX models perform well at computational level

HMAX models -- a special case of M-theory -- perform well compared to engineered computer vision systems (in 2006) on several databases

Bileschi, Wolf, Serre, Poggio, 2007; Mutch Lowe
HMAX models perform well at computational level

HMAX models -- a special case of M-theory -- perform well compared to engineered computer vision systems (in 2006) on several databases

Bileschi, Wolf, Serre, Poggio, 2007; Mutch Lowe
Models: computational performance

- Deep convolutional networks (such as Lenet) as an architecture are special case of Mtheory (with just translation invariance and max/sigmoid pooling)
- HMAX as an architecture is a special case of Mtheory (with translation + scale invariance and max pooling) and used to work well
- Encouraging initial results in speech and music classification (Evangelopoulos, Zhang, Voinea)
- Example in face identification (Liao, Leibo) --->
Computational performance: example faces

Labeled Faces in the Wild

Contains 13,233 images of 5,749 people
Accuracy (%)

- Our model: 87.6%
- APEM: 81.7%
- Li et al. (2013): 81.7%
- Rahimzadeh et al. (2013): 79.08%
- Sanderson et al. (2009): 72.95%

LFW - no outside data used & no alignment

Q. Liao, J. Leibo
Plan

1. Motivation: models of cortex (and deep convolutional networks)
2. Core theory
   - the basic invariance module
   - the hierarchy
3. Computational performance
4. Biological predictions
5. Theorems and remarks
   - $n \to 1$
   - invariance and sample complexity
   - connections with scattering transform
   - invariances and beyond perception
   - ...
Theory of unsupervised invariance learning in hierarchical architectures

• neurally plausible: HW module of simple-complex cells
• says what simple-complex cells compute
• provides a theory of pooling: energy model, average, max...
• leads to a new characterization of complex cells
• provides a computational explanation of why Gabor tuning
• may explain tuning and functions of V1, V2, V4 and in face patches!
• suggests generic, Gabor-like tuning in early areas and specific selective tuning higher up

poggio, anselmi, rosasco, tacchetti, leibo, liao
Musing on technology: a second phase in Machine Learning?

• The first phase -- from ~1980s -- led to a rather complete theory of supervised learning and to practical systems (MobilEye, Orcam,...) that need lots of examples for training: $n \to \infty$

• The second phase may be about unsupervised learning of (invariant) representations that make supervised learning possible with very few examples: $n \to 1$
A theory of feedforward vision

• The basic equation of physics can be derived from a small number of symmetry properties: invariance wrt space+time, conservation of energy, invariance to measurement units....

• Is the architecture and tuning properties of visual (and auditory...) cortex predictable from basic symmetries of geometric transformations of images?