Learning in Hierarchical Architectures: from Neuroscience to Derived Kernels

tomaso poggio, McGovern Institute BCS, CSAIL MIT

Learning is the gateway to understanding the brain and to making intelligent machines.

Problem of learning:
a focus for
  - math
  - computer algorithms
  - neuroscience
Message of today

Neuroscience *may begin* to provide new ideas and approaches to machine learning, AI and computer vision.
Message of today

Neuroscience *may begin* to provide new ideas and approaches to machine learning, AI and computer vision.

A case in point: (un)supervised learning
Learning in Hierarchical Architectures: from Neuroscience to Derived Kernels

tomaso poggio, McGovern Institute BCS, CSAIL MIT

Learning is the gateway to understanding the brain and to making intelligent machines.

Problem of learning:
a focus for
  - math
  - computer algorithms
  - neuroscience
1. Today’s supervised learning algorithms: sample complexity problem and shallow architectures
2. Visual Cortex: hierarchical architecture, from neuroscience to a class of models
3. Physiology, psychophysics, computer vision
4. Models suggest new architectures for learning
5. Extensions and limitations of models
Supervised learning

*Given* a set of \( l \) examples (data) \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \)

**Question**: find function \( f \) such that

is a *good predictor* of \( y \) for a *future* input \( x \) (*fitting the data is not enough!*):

\[
f(x) = \hat{y}
\]
Supervised learning

There is an unknown probability distribution on the product space \( Z = X \times Y \), written \( \mu(z) = \mu(x, y) \). We assume that \( X \) is a compact domain in Euclidean space and \( Y \) a closed subset of \( IR \).

The training set \( S = \{(x_1, y_1), \ldots, (x_n, y_n)\} = \{z_1, \ldots, z_n\} \) consists of \( n \) samples drawn i.i.d. from \( \mu \).

\( \mathcal{H} \) is the hypothesis space, a space of functions \( f : X \rightarrow Y \).

A learning algorithm is a map \( L : Z^n \rightarrow \mathcal{H} \) that looks at \( S \) and selects from \( \mathcal{H} \) a function \( f_S : x \rightarrow y \) such that \( f_S(x) \approx y \) in a predictive way.
Definitions

Given a function $f$, a loss function $V$, and a probability distribution $\mu$ over $Z$, the **expected or true error** of $f$ is:

$$l[f] = E_Z V[f, z] = \int_Z V(f, z) d\mu(z) \quad (1)$$

which is the **expected loss** on a new example drawn at random from $\mu$.

The **empirical error** of $f$ is:

$$l_S[f] = \frac{1}{n} \sum V(f, z_i) \quad (2)$$

A very natural requirement for $f_S$ is distribution independent **generalization**

$$\forall \mu, \lim_{n \to \infty} |l_S[f_S] - l[f_S]| = 0 \text{ in probability} \quad (3)$$

In other words, the training error for the solution must converge to the expected error and thus be a “proxy” for it. Otherwise the solution would not be “predictive”.
Classical learning theory and Kernel Machines
(Regularization in RKHS)

\[
\min_{f \in H} \left[ \frac{1}{n} \sum_{i=1}^{n} V(f(x_i) - y_i) + \lambda \left\| f \right\|_K^2 \right]
\]

implies

\[
f(x) = \sum_{i}^{n} \alpha_i K(x, x_i)
\]

Equation includes splines, Radial Basis Functions and SVMs (depending on choice of \( V \)).

For a review, see Poggio and Smale, The Mathematics of Learning, Notices of the AMS, 2003; see also Schoelkopf and Smola, 2002; Bousquet, O., S. Boucheron and G. Lugosi.
Classical learning theory and Kernel Machines (Regularization in RKHS)

\[
\min_{f \in H} \left[ \frac{1}{n} \sum_{i=1}^{n} V(f(x_i) - y_i) + \lambda \|f\|^2_K \right]
\]

implies

\[
f(x) = \sum_{i}^{n} \alpha_i K(x, x_i)
\]

Kernel machines correspond to shallow networks
Present learning algorithms: “high” sample complexity and shallow architectures

How do the learning machines described by classical learning theory -- such as kernel machines -- compare with brains?

One of the most obvious differences is the apparent ability of people and animals to learn from very few examples (“poverty of stimulus” problem).

A comparison with real brains offers another, related, challenge to learning theory. Classical “learning algorithms” correspond to one-layer architectures. The cortex suggests a hierarchical architecture.

Are hierarchical architectures with more layers the answer to the sample complexity issue?

The Mathematics of Learning: Dealing with Data Tomaso Poggio and Steve Smale
Towards a hierarchical learning theory

How then do the learning machines described in the theory compare with brains?

- One of the most obvious differences is the ability of people and animals to learn from very few examples.

- Are hierarchical architectures with more layers justifiable in terms of learning theory?

- Why hierarchies in the brain?
1. Today’s supervised learning algorithms: sample complexity problem and shallow architectures
2. Visual Cortex: hierarchical architecture, from neuroscience to a class of models
3. Physiology, psychophysics, computer vision
4. Models suggest new architectures for learning
5. Extensions and limitations of models
Hypothesis: the hierarchical architecture of the ventral stream in monkey visual cortex has a key role in object recognition...of course subcortical pathways may also be important (thalamus, in particular pulvinar...).
The Ventral Stream

visual recognition is a difficult learning problem
(e.g., “is there an animal in the image?”)
The Ventral Stream

• Human Brain
  – $10^{10}$-$10^{11}$ neurons (~1 million flies 😊)
  – $10^{14}$-$10^{15}$ synapses

• Ventral stream in rhesus monkey
  – $10^9$ neurons
  – $5 \times 10^6$ neurons in AIT (Anterior InferoTemporal) cortex
The ventral stream in monkey visual cortex has a key role in solving this problem. Of course, subcortical pathways may also be important (thalamus, in particular, pulvinar...).
The ventral stream hierarchy: V1, V2, V4, IT

A gradual increase in the receptive field size, in the complexity of the preferred stimulus, in tolerance to position and scale changes

Kobatake & Tanaka, 1994
The ventral stream

Feedforward connections only?
The ventral stream

Feedforward connections as well as backprojections:

How far we can push the simplest type of feedforward hierarchical models?
The ventral stream

(Thorpe and Fabre-Thorpe, 2001)
Model of Visual Recognition (millions of units) based on neuroscience of cortex

*Modified from (Gross, 1998)

[software available online]

Model of Visual Recognition (millions of units) based on neuroscience of cortex

- It is in the family of “Hubel-Wiesel” models (Hubel & Wiesel, 1959; Fukushima, 1980; Oram & Perrett, 1993, Wallis & Rolls, 1997; Riesenhuber & Poggio, 1999; Thorpe, 2002; Ullman et al., 2002; Mel, 1997; Wersing and Koerner, 2003; LeCun et al 1998; Amit & Mascaro 2003; Deco & Rolls 2006…)

- As a biological model of object recognition in the ventral stream – from V1 to PFC -- it is perhaps the most quantitative and faithful to known neuroscience

- Feedforward only: an approximation of the first 100 msec of visual perception. *Potential key limitation*

- Hierarchy of disjunctions of conjunctions (~Geman)

Two operations, one circuit?

- Tuning operation (Gaussian-like, AND-like)
  \[ y = \exp(-|x - w|^2) \]
  \[ y = \frac{w \cdot x}{|x|} \]
- Simple units

- Max-like operation (OR-like)
  \[ y = \max\{x_1, x_2, \ldots\} \]
- Complex units
Instead of Gaussian, normalized dot product

A plausible biophysical implementation of a Gaussian-like tuning (Kouh, Poggio, 2008):

\[
\text{normalized dot product} \quad \frac{W \cdot x}{|x|}
\]
Two operations, one circuit?

A plausible biophysical implementation for both Gaussian tuning (~AND) + max (~OR): normalization circuits with divisive inhibition (Kouh, Poggio, 2008; also RP, 1999; Heeger, Carandini, Simoncelli,…)
Task-specific circuits (from IT to PFC) - **Supervised** learning: ~ classifier

- Overcomplete dictionary of “templates” or image “patches” is learned during an **unsupervised** learning stage (from ~10,000 natural images) by tuning S units.

see also (Foldiak 1991; Perrett et al 1984; Wallis & Rolls, 1997; Lewicki and Olshausen, 1999; Einhauser et al 2002; Wiskott & Sejnowski 2002; Spratling 2005)
Learning: supervised and unsupervised

We refer to the sample complexity of the preprocessing stage as the # of labeled examples required by the classifier at the top

- Preprocessing stages lead to a representation that has lower sampling complexity than the image itself

1. Today’s supervised learning algorithms: sample complexity problem and shallow architectures
2. Visual Cortex: hierarchical architecture, from neuroscience to a class of models
3. Physiology, psychophysics, computer vision
4. Models suggest new architectures for learning
5. Extensions and limitations of models
Hierarchical feedforward models of the ventral stream

Millions of units

CBCL software available on the Web
Hierarchical feedforward models of the ventral stream

Hierarchical Feedforward Models: predict/are consistent //w neural data

- V1:
  - Simple and complex cells tuning (Schiller et al 1976; Hubel & Wiesel 1965; Devalois et al 1982)
    - MAX-like operation in subset of complex cells (Lampl et al 2004)
  - V4:
    - Tuning for two-bar stimuli (Reynolds Chelazzi & Desimone 1999)
      - MAX-like operation (Gawne et al 2002)
    - Two-spot interaction (Freiwald et al 2005)
  - Tuning for boundary conformation (Pasupathy & Connor 2001, Cadieu, Kouh, Connor et al., 2007)
    - Tuning for Cartesian and non-Cartesian gratings (Gallant et al 1996)
  - IT:
    - Tuning and invariance properties (Logothetis et al 1995, paperclip objects)
      - Read out results (Hung Kreiman Poggio & DiCarlo 2005)
    - Pseudo-average effect in IT (Zoccolan Cox & DiCarlo 2005; Zoccolan Kouh Poggio & DiCarlo 2007)
      - Human:
        - Rapid categorization (Serre Oliva Poggio 2007)
        - Face processing (fMRI + psychophysics) (Riesenhuber et al 2004; Jiang et al 2006)
Hierarchical feedforward models of the ventral stream

Rapid Categorization
Mask should force visual cortex to operate in feedforward mode.

Thorpe et al 1996; Van Rullen & Koch 2003; Bacon-Mace et al 2005
Hierarchical feedforward models of the ventral stream

Feedforward Models:
“predict” rapid categorization (82% model vs. 80% humans)
Hierarchical feedforward models of the ventral stream

- Image-by-image correlation:
  - Heads: $\rho=0.71$
  - Close-body: $\rho=0.84$
  - Medium-body: $\rho=0.71$
  - Far-body: $\rho=0.60$
Hierarchical feedforward models of the ventral stream

Feedforward Models: perform well compared to engineered computer vision systems (in 2006)
“Mutations” of the architecture perform well!
“Mutations” of the architecture perform well!

1. Generate thousands of model variants
2. Unsupervised training of each variant
   - GPU-accelerated (~1000x speed-up)
3. Supervised testing of each variant

"Mutations" of the architecture perform well!

Cox, Pinto, Doukhan, Corda & DiCarlo (2008); Pinto, DiCarlo & Cox (in prep)
1. Today’s supervised learning algorithm: sample complexity problem and shallow architectures
2. Visual Cortex: from neuroscience to a class of models
3. Physiology, psychophysics, computer vision
4. Models suggest new architectures for learning
5. Extensions (video+attention) and limitations of feedforward hierarchical models
Hierarchical feedforward models of visual cortex may be wrong ...but present a challenge for “classical” learning theory:

an unusual, hierarchical architecture with unsupervised and supervised learning working well...

...so... we need theories -- not just models!
Derived Kernels

joint work with J. Bouvrie (MIT), A. Caponnetto (CityU), L. Rosasco (MIT and Genoa), S. Smale (TTI-C)

January 24, 2009
Plan

1. Derived Kernel and Neural Response
2. Theoretical Analysis
The goal

We consider hierarchical architectures to preprocess an image in order to reduce sample complexity of a classifier based on labeled examples and “looking” at the output of this unsupervised preprocessing architecture. The architecture is synthesized from a large number of unlabeled, unsupervised examples. Ideally we would like to show that more layers are “better” and we would like to derive conditions under which these statements can be characterized.
The visual cortex model

The ingredients needed to define the derived kernel consist of:

- a finite *architecture* of nested patches
- a set of *transformations* from a patch to the next larger one,
- a suitable family of *function spaces* defined on each patch,
- a set of *templates* which connect the mathematical model to a real world setting.
An Architecture of Patches

We first consider an architecture composed of three layers of patches: $v_1, v_2$ and $S_q$ in $\mathbb{R}^2$, with $v_1 \subset v_2 \subset S_q$, 
Images as Functions

We consider a function space on $S_q$, denoted by

$$\text{Im}(S_q) = \{ f : S_q \rightarrow [0, 1] \},$$

as well as the function spaces $\text{Im}(v_1), \text{Im}(v_2)$ defined on subpatches $v_1, v_2$, respectively.

Functions can be interpreted as grey scale images when working with a vision problem for example.
Transformations

Next, we assume a set $H_{v_1}$ of *transformations* that are maps from the smallest patch to the next larger patch

$$h : v_1 \rightarrow v_2.$$ 

Similarly $H_{v_2}$ with

$h : v_2 \rightarrow Sq.$

The sets of transformations are assumed to be finite. Examples of transformations are translations, scalings and rotations... In the vision interpretation, a translation $h$ can be thought of as moving the image over the “receptive field”. $f \circ h$ “restricts $f$ to a smaller patch associated with $h$. 


Templates

Template sets are finite, $T_{v_1} \subset \text{Im}(v_1)$ and $T_{v_2} \subset \text{Im}(v_2)$.

- they are image patches sampled from some set of unlabeled images.
- link the mathematical development to real world problems.
For a general $n$ layer architecture $v_1 \subset v_2 \subset \cdots \subset v_n = Sq$, let $K_n = K_{v_n}$ and $H_n = H_{v_n}$, $T_n = T_{v_n}$.

Definition

Given a non-negative valued, normalized, reproducing kernel $\hat{K}_1$, the $m$-layer derived kernel $\hat{K}_m$, $m = 2, \ldots, n$, is obtained by normalizing

$$K_m(f, g) = \langle N_m(f), N_m(g) \rangle_{L^2(T_{m-1})}$$

where

$$N_m(f)(t) = \max_{h \in H} \hat{K}_{m-1}(f \circ h, t), \quad t \in T_{m-1}$$

with $H = H_{m-1}$.
Recursive Definition

For a general $n$ layer architecture $v_1 \subset v_2 \subset \cdots \subset v_n = S q$, let $N_n = N_{v_n}$ and $H_n = H_{v_n}$, $T_n = T_{v_n}$. 
Recursive Definition

For a general $n$ layer architecture $v_1 \subset v_2 \subset \cdots \subset v_n = Sq$, let $N_n = N_{v_n}$ and $H_n = H_{v_n}$, $T_n = T_{v_n}$.

**Definition**

Let us assume that a non-negative valued, normalized, feature map $N_1$ is given. Thus the derived kernel can be defined as $K_1(f, g) = \langle N_m(f), N_m(g) \rangle_{L^2(T_{m-1})}$. The recursion is

$$N_m(f)(t) = \max_{h \in H} \hat{K}_{m-1}(f \circ h, t), \quad t \in T_{m-1}$$

with $t \in T_{m-1}$, $H = H_{m-1}$, $K_m(f, g) = \langle N_m(f), N_m(g) \rangle_{L^2(T_{m-1})}$ and $\hat{K}_{m-1}$ being the normalized kernel.
Plan

1. Derived Kernel and Neural Response
2. Theoretical Analysis
Initial results on the *invariance* of the (normalized) neural response to some set of transformations $\mathcal{R}$.

**Theorem**

If the initial kernel satisfies $\hat{K}_1(f, f \circ r) = 1$ for all $r \in \mathcal{R}$, $f \in \text{Im}(v_1)$, then

$$\hat{N}_m(f) = \hat{N}_m(f \circ r),$$

for all $r \in \mathcal{R}$, $f \in \text{Im}(v_m)$ and $m \leq n$. 
Example: Discrimination Results for 1-D strings

Consider an exhaustive architecture: \( v_m = \{1, \ldots, m\} \),
\( T_m = \text{Im}(v_m) = S^m \), for \( m = 1, \ldots, n \) and transformations are all possible translations.

**Theorem**

If \( f, g \) are \( n \)-strings and \( \hat{K}_n(f, g) = 1 \) then one of the following statements is true:

- \( f, g \) are the same string
- one is the reversal of the other
- \( f, g \) are the “checkerboard” pattern:
  \( f = ababa \cdots, g = babab \cdots \), with \( f \) and \( g \) odd length strings.
Hierarchy can reduce sample complexity: empirical support

Sample Complexity: 9-class NN-Classification Example

Boutrie, Rosasco, Poggio, Smale, 2009
Summary

- Compact mathematical description of a feedforward model of the visual cortex
- “Derived” kernel recursively defined
- Initial results on invariance/discrimination properties

- Open problem: Discrimination/Approximation properties
- Open problem: Number of layers and sample complexity (poverty-of-stimulus)
- Open problem: Efficient learning of the templates
- Conjecture: small effective dimensionality at each layer
1. Today’s supervised learning algorithm: sample complexity problem and shallow architectures
2. Visual Cortex: from neuroscience to a class of models
3. Physiology, psychophysics, computer vision
4. Models suggest new architectures for learning
5. Extensions (video+attention) and limitations of feedforward hierarchical models
Extension to motion: model of the dorsal stream

Automatic recognition of continuous rodent behavior (over hours): automatic phenotyping

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human agreement</td>
<td>72%</td>
</tr>
<tr>
<td>proposed system</td>
<td>71%</td>
</tr>
<tr>
<td>commercial system</td>
<td>56%</td>
</tr>
<tr>
<td>chance</td>
<td>12%</td>
</tr>
</tbody>
</table>
From neuroscience to models: extension to attention

![Graph showing FEF and V4 activity with time from stimulus onset (ms) and potential target and non-target stimuli](image-url)
From neuroscience to models: extension to attention

An integrated Bayesian model with B. Desimone + E. Miller


see also Rao 2005; Lee & Mumford 2003
From neuroscience to models: extension to attention

An integrated Bayesian model with B. Desimone + E. Miller


see also Rao 2005; Lee & Mumford 2003
From neuroscience to models: extension to attention

An integrated Bayesian model with B. Desimone + E. Miller

Limitations of present feedforward hierarchical models

• Vision is more than categorization or identification: it is image understanding/inference/parsing
• Our visual system can “answer” almost any kind of question about an image or video (a Turing test for vision…)

• Three options: 1) top-down (attentional) control of task-dependent routines 2) probabilistic inference in the ventral stream (~Mumford, Geman) 3) hierarchical probabilistic inference by the brain (Tenenbaum)
Collaborators in recent work
