M-THEORY

Jesika Haria, EECS 2014

M-theory is a theoretical framework for the next phase of machine learning beyond supervised learning: the unsupervised learning of representations that reduce the sample complexity of the final supervised learning stage. It is being developed since 2011 by Tomaso Poggio[1], a professor at the Massachusetts Institute of Technology.

TABLE OF CONTENTS

Introduction

Supervised vs Unsupervised learning

Supervised machine learning algorithms as practiced in the industry today are sensitive to noise under example scarcity[2]. Assuming zero-mean white noise, these algorithms rely on large numbers of inputs in order to improve predictions, and hence are computationally expensive to run. In general, accuracy of estimates increases logarithmically as n goes to infinity, where n is the number of input samples.

M-theory was based on the observation that representations that are invariant to translation, scale and other transformations can considerably reduce the sample complexity of learning. What this enables is recognition of new object classes from a very few examples, or as n goes to 1. Empirical estimates of one-dimensional projections of the distribution induced by a group of affine transformations are proven to represent a unique and invariant signature associated with a particular image. M-theory demonstrates how projections yielding invariant signatures for future images can be learned automatically and updated continuously, during unsupervised visual experience.

Inspiration

Recognition in the visual cortex is primarily based on hierarchical models, given their practical success in modeling the cortex as well as computer vision systems. Hubel and Weil's original proposal for visual area V1 can be used to construct translation-invariant detectors, a key insight underlying networks for visual recognition, such as HMAX and convolutional networks.

Hierarchical model of recognition in the visual cortex

A popular hierarchical model that works well is called HMAX. In HMAX, a layer is a 3-D array of units that collectively represent the activity of some set of features (F) at each location in a 2-D grid of points in retinal space (X, Y). Each cell is computed using the cells in the layers immediately below as inputs. S1 are Gabor Filter Layers, where there is only one feature at each gird point, the pixel intensity. Next up are C1 or Local Invariance Layers, which compute a location maximum over (X, Y) and scale. Each feature in a S2 or Intermediate Feature Layers looks for a particular local combination of different Gabor filter responses, and finally is fed into the C2 or Global Invariance Layer, where the maximum response over all (X, Y) and all scales is found[3].

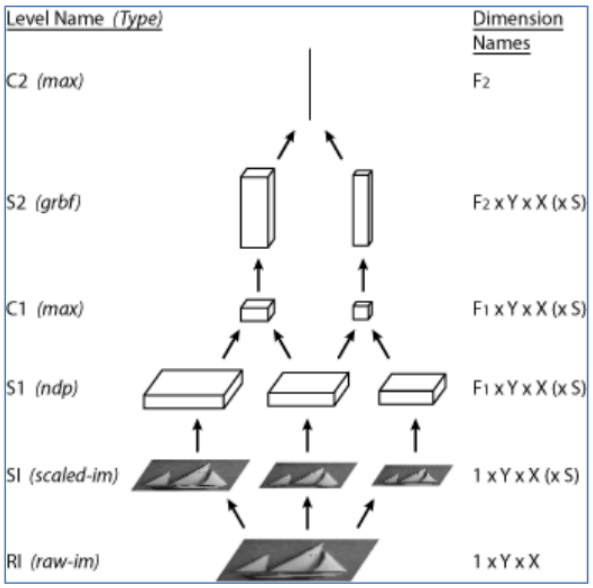


Figure 1: HMAX Model Example

Relation to deep learning architectures

In a similar vein to hierarchical modeling of the cortical functions for sensory recognition, there has been a recent rise in the success and popularity of deep learning networks, which are essentially convolutional networks[4]. In convolutional networks, individual neurons are tiled together to overlap in regions of the visual field, in multiple layers [5].

Core Theory

Image representation in the ventral stream

The initial assumption in forming a model of ventral stream architecture is that the goal of the ventral stream is to compute a representation of objects that is invariant to transformations. New images can be encoded in an invariant way by processes involving storing a 'movie' of a transformation. The conjecture that most of the complexity in classification problems arise from difference in viewpoint or illumination that swamp the intrinsic characteristics of the object, is widely supported by experimental observations. For instance, in the task in Figure 2, the solid line refers to a classification example for a rectified task, whereas the dotted line is for an unrectified task.

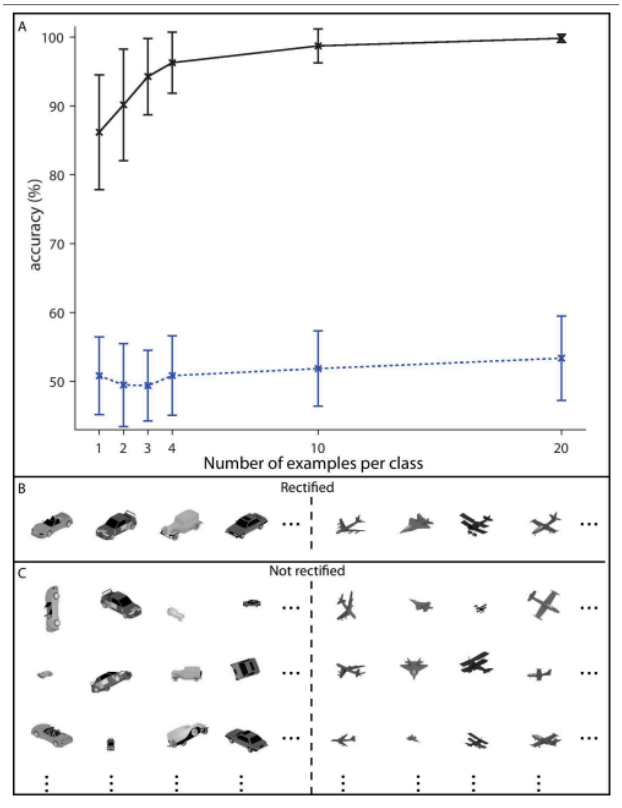


Figure 2 Complexity of recitifed vs unrectified task

Affine Transformations

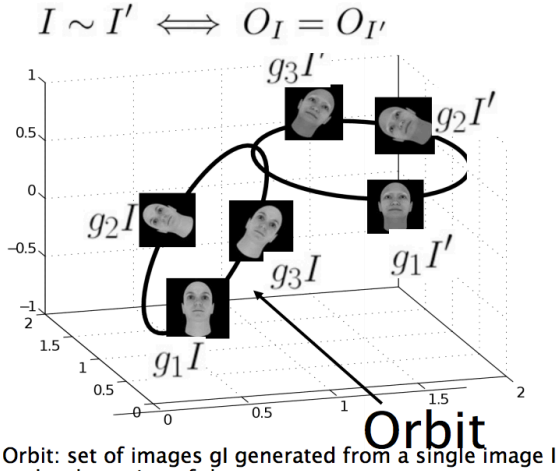
Affine transformations are defined as geometric transformations of the image I transformations $T \circ I$ such that:

$$T \circ I = I(x', y')$$

For example, $x' = Ax + t_x$. Rotations, translations and scaling can be thought of as affine.

Group Invariance Theorem

The signature assigned to a particular image I is then associated with vector that is unique and invariant. Consider finite compact groups, G. Define $gI(x) = I(g^{-1}x)$ as the action of a unitary transformation of the group's actions on an image. Now, one might define an orbit O_I, which represents the set of images gI that are generated from the image I.



Orbit: set of images gI generated from a single image I under the action of the group.

Figure 3 Images from the same orbit

The claim is that two sets of images are equivalent if they belong in the same orbit, which formalizes the notion of an orbit being invariant and unique. Mathematically, this could be written as $I \sim I' if \ \exists g \in G \mid I' = gI$. This implies that if two orbits have a point of intersection, then they are identical everywhere. Conversely, two images are different if no part of their orbit matches.

Neurons' way of computing Invariance

Considering gI as a realization of a random variable, if two orbits coincide, then their associated distributions under the group G are identical. Hence,

$$I \sim I' \Leftrightarrow O_I = O_{I'} \Leftrightarrow P_I = P_{I'}$$

Neuroscience suggests that a natural function for a neuron is to compute a high-dimensional dot product between an 'image patch' and another image patch (a template), stored in terms of their synaptic weights that capture memory.

Using the Cramer-Wold Theorem, the P_I of the image can be almost uniquely characterized by K one-dimensional probability distributions $P_{<I,t^k>}$ induced by the (one-dimensional) set of projections $< I, t^k >$ where $t^k, k = 1, \dots, K$ are a set of randomly chosen 'template' images.

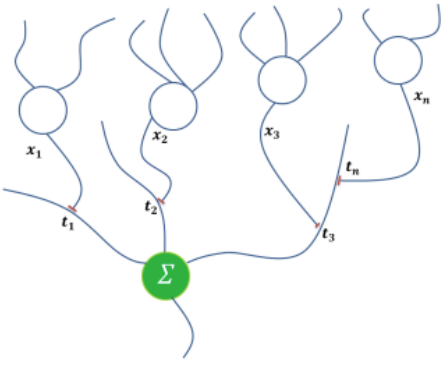


Figure 4 Neuron performing a-dimensional inner products

However, this requires the observation of the image and all its transformations gI. The goal of this technique is to be able to compute an invariant signature for a new object without any more examples. Using the fact that $< gI, t^k > = < I, g^{-1} t^k >$, it is inferred that the same one-dimensional distribution is obtained on from all projections of a particular image onto a fixed template as is obtained from the projections of that particular image onto all transformations of that fixed template. In other words, implicit knowledge of transformations received from transformations of the template enable the system to automatically be invariant to those transformations for previously unseen inputs.

Example

Plane Rotation Example

Here, we can see an example of a face rotation in a plane. Once the template face has undergone the gt^k transformation, it generates CDF distributions that are very similar. Conducting the Inter-person K-S test[6] allows us to distinguish between people, while for

K-S tests conducted on varying representations of the same person, invariance can be found.

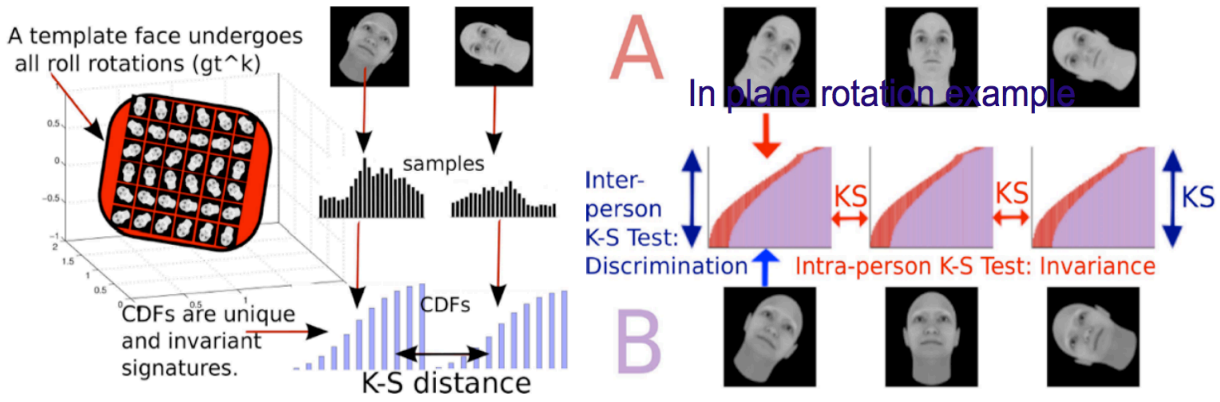


Figure 5 Plane Rotation Example of Face

Extensions

Given the assumptions of the theory in terms of restrictions on the group, a number of extensions are possible, a few of which are briefly listed as follows.

Non-compact groups

For a transformation observed via a 'receptive field' as in non-compact groups, there is only partial invariance. Localization implies invariance, and one could achieve it either via a wavelet-like regime or via a sparse regime.

Hierarchies of magic modules (multi-layer)

For multi-layer architectures, the main considerations are compositionality, factorization of invariant ranges, minimizing clutter effect memory access and optimizing local connections.

Non-group transformations

Rotation of the image in another plane is not a group transformation, and new information is being integrated into learning. The linking conjecture predicts the Gabor-like tuning of simple cells in V1, while the complex cells are more viewpoint tolerant.

REFERENCES

[1] *Unsupervised Learning of Invariant Representations in Hierarchical Architectures*, F. Anselmi et al, Submitted to Porceedings of the National Academy of Sciences of the United States of America

[2] *General Limitations on Machine Learning,* A. Hoffman, Technische Universit□at Berlin, http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=2CCFDE7CD88949AC38E818E5AD13C9A2?doi=10.1.1.48.2327&rep=rep1&type=pdf

[3] *HMAX Models Architecture*, J. Mutch, March 30, 2010 http://www.mit.edu/~9.520/spring10/slides/class15-visualneuroscience/class15-hmax.pdf

[4] *Scaling Learning Algorithms towards AI*, Y. Bengio and Y. LeCun, To appear in "Large-Scale Kernel Machines", L. Bottou, O. Chapelle, D. DeCoste, J. Weston (eds) MIT Press, 2007http://yann.lecun.com/exdb/publis/pdf/bengio-lecun-07.pdf

[5] Convolutional Neural Networks (LeNet), Dec 13, 2013, http://deeplearning.net/tutorial/lenet.html

[6] Kolmogorov–Smirnov test, http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test