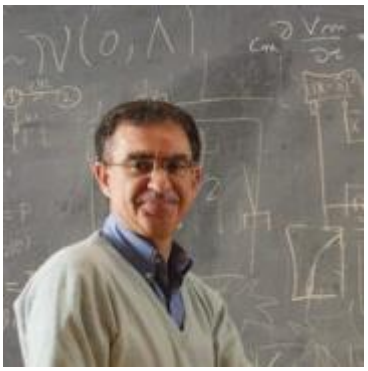
Deep Learning: mathematics and neuroscience

By Tomaso Poggio April 26, 2016



Science and Engineering of Intelligence

The problems of Intelligence are, together, the greatest problem in science and technology today. Making significant progress towards their solution will require the interaction of several disciplines involving neuroscience and cognitive science in addition to computer science, robotics and machine learning.

In this perspective I will discuss the implications of the recent empirical success in many applications, such as image categorization, face identification, localization, action recognition, depth estimation, speech recognition of a machine learning technique dubbed Deep Learning and based on multilayer neural networks. Hierarchical neural networks have become a core tool in machine learning. Here we discuss some of the technical advances that have led to this recent progress.

The original idea came from basic neuroscience work in visual cortex by Hubel and Wiesel *(1)* who recorded in the 60's from simple and complex cells in the primary visual cortex and implicitly suggested a hierarchy of simple, complex and hyper-complex cells which was later taken to mean a series of layers of units supporting more and more complex features and greater and greater invariance. The first quantitative model was the Neocognitron by Fukushima *(2)*: its architecture was identical to today's multilayer neural networks, comprising convolution, max pooling and nonlinear units; the training was, however different, and did not rely on the supervised stochastic gradient descent (SGD) technique introduced by Hinton *(see (3))* and used in today's deep learning networks. Several other quantitative models followed, some geared towards performance such as LeNet *(4)* and some towards the original goal — modeling the ventral stream pathway, such as HMAX *(5,6)*.

While the roots of deep learning architectures are in the science of the brain, the advances in performance during the last five years are exclusively an engineering feat, the cumulative effect of much increased computer power, the availability of manually labeled large data sets and a small number of incremental technical improvements. It is telling that several of the algorithmic tricks that were touted as breakthroughs just a couple of years ago, are now regarded as unnecessary. Some notable examples are "pretraining", "dropout", "max pooling" and "inception modules". I think that

some of the other ideas are more fundamental, as I explain below, and likely to be more durable, though their exact form is bound to change somewhat. The main ones are “data augmentation” (once called virtual examples *(7)*), “batch normalization” and “residual learning” *(8)*.

Deep Learning: scientific breakthroughs?

It seems that Deep Learning is more than a very good engineering implementation of existing knowhow: it may be the beginning of a breakthrough with the potential of opening new fields for science. There are several reasons for this statement. I mention the two that are most surprising to me.

The first has to do with neuroscience. Deep networks trained with Imagenet seem to mimic not only the recognition performance but also the tuning properties of neurons in cortical areas of the visual cortex of monkeys *(9)*. There are major caveats here. The variance explained is only around 50%, the comparison is between actual neurons and optimal linear combinations of units in a layer of the model, non-uniformities in cortical areas (such as cortical layers as well as face patches etc) are neglected. For the sake of the argument let me take here an optimistic view. The optimistic claim is then as follows: in order to solve the problem of the ventral stream — its organization, its function, its properties — it is sufficient to spell out a computational task, such as high-variation object categorization task, and then optimize performance of a parametrized architecture on millions of data points, that is images and their labels. The constraints imposed on the architecture are rather weak: multilayer networks with simple nonlinear units and generic shift invariant connectivity of the Neocognitron and HMAX type. Since the constrained optimization problem seems to yield properties similar to what neuroscience shows, one is led to conclude that the optimization problem with such weak constraints has a unique solution and, furthermore, that evolution has found the same solution. As mentioned in *(9)* this would imply that Marr’s (and mine) computational level of analysis *(10)* effectively determines the lower levels of algorithms and implementation, something I find nice, difficult to believe and important to verify and characterize.

The second reason has to do with the existing mathematical theory of machine learning and the need to extend it. In particular, trained multilayer networks (deep networks) seem capable of generalizing to other tasks and other databases quite different from the one on which they were originally trained (see for a review *(3)),* in a way that is much better than classical one-layer kernel machines (shallow networks), which typically suffer from overfitting any specific training set. Networks trained to perform classification on Imagenet achieve very good performance on different image

datasets with little additional training. The same networks can be used to perform detection or segmentation, again with relatively minor training of an additional output layer. All of this is consistent with the intriguing ability of deep learning networks to avoid overfitting and to show a testing error similar to the training error, at least when trained with very large datasets. Clearly this situation poses an interesting set of challenges and opportunities to the mathematics of learning and to the field of function approximation.

Deep Learning: open scientific questions and some answers

The ability of Deep Learning network to predict properties of visual cortex seems a major breakthrough. *What if we now know how to develop an important subset of the basic building blocks for brain-like intelligence?* Of course training with millions of labeled examples is biologically implausible ("*not the way children learn to distinguish a dog from a cat"*). However, labels are of course arbitrary. What is clearly important for an organism is to be able to group together (with the implicit label "same identity") images of the same object (or for classification, of the same object class). I call this *implicit labeling*: explicit labels are not provided but there is contextual or other information that allows implicit labeling. Several plausible ways are available to biological organisms, especially during development, to implicitly label in this ways images and other sensory stimuli such as sounds and words. For images, time continuity is a powerful and primitive cue for implicit labeling. In fact, time continuity was proposed by i-theory to learn invariance to transformations during development *(11,12)* by associating together images of different transformations of the same template. The same strategy is used by Wang and Gupta *(13)* by tracking patches of images in videos. Many other strategies can also be used. Here are some back of the envelope calculations of how many unlabeled images a baby robot could get during its first year of life. Suppose one saccade per second per 8 hours per 360 days: this gives in the order of ~10M images. Even if only a small fraction, like 10%, could be implicitly labeled this would provide a sufficiently rich training set as suggested by the empirical results on Imagenet. Clearly a single DLN architecture cannot deal with challenging cognitive tasks like humans perform all the time. However, it may be possible to build a mind with a set of different modules several of which are Deep Learning networks.

Let me turn now to the set of mathematical questions posed by the successful engineering of Deep Learning. The headline question is *when and why are deep networks better than shallow?* In fact, there are two basic sets of questions about Deep Neural Networks. The first is about the power of the architecture -- which classes of functions can it approximate well? What is the role of convolution and

pooling? The second set of questions is about learning the unknown coefficients from the data: do multiple solutions exist? How "many"? Why is SGD so unreasonably efficient, at least in appearance? Are good local minima easier to find in deep rather than in shallow networks? This last point may be very important in practice.

Answers to some of these questions are just beginning to emerge. There is now a reasonably complete theory of convolutional and pooling layers that extends the current neural network design to other groups beyond the translation group and to non-group transformations (yielding partial invariance under appropriate conditions). Some of the less obvious and interesting mathematical results are (12, 14):

- Pooling of a sufficient number of nonlinear units is sufficient for invariance and uniqueness (see Theorems in section 3 of *(12)*), provided there is no clutter within the pooling region. The combination of just convolution and subsampling is thus sufficient for selective invariance. The form of the nonlinearity used in pooling is flexible: rectifiers and threshold functions are among the simplest choices.

Figure 1. a) A shallow universal network with one node containing *n* units or channels b) A binary tree hierarchical network. Some of the best deep learning networks *(8)* are quite similar to the binary tree architecture.

- The size of the pooling region does not affect uniqueness apart from increasing susceptibility to clutter. Reducing clutter effects on selective invariance by restricting the pooling region is probably a main role for "attention".
- Max pooling is a form of pooling that provides invariant, but not strictly selective information. There is no need for max pooling apart from computational considerations.

The results about pooling do not explain why and when multiple layers are better than on a single hidden layer. Using the language of machine learning and function approximation the following statements *(15)* can be made.

- Both shallow (a) and deep (b) networks are *universal*, that is they can approximate arbitrarily well any continuous function of variables in a compact domain.

- It is intuitive that a hierarchical network matching the structure of compositional function as in figure 1 should be ``better'' at approximating it than a generic shallow network but universality of shallow networks makes the statement less than obvious. To answer the question one has to compare the number of parameters needed for an error between shallow and deep binary networks of the form of Figure 1.

- A shallow universal network in 8 variables with one node containing several units (see Figure 1) can approximate a generic continuous multivariate function. Figure 1b shows a binary tree hierarchical network in 8 variables, which reflect the structure of compositional functions of the form . It is important to emphasize that state-of-art NNs with their small kernel size and many layers *(8)* are quite similar to the binary tree architecture of Figure 1b, which is itself similar to hierarchical models of visual cortex, which shows a doubling of receptive field sizes from one area to the next. Of course, the theorem also applies to trees that are not strictly binary trees.
- Each of the nodes consists of ReLU units and computes the ridge function with each units correspond to a so called channel *(3)*. Though not needed for the theorem, the figure assumes shift invariance (equivalent to weight sharing and convolution): thus the nodes in each layer compute the same function (for in the first layer).

- Let us first define W to be the set of all with the unit cube, which have a continuous gradient with Sobolev norm; we further define to be the set of shallow networks of the form. The following proposition *(15)* then holds: *Let be integers, and be a univariate function which is not a polynomial. Then the closure of contains the space of all polynomials in d variables with coordinate wise degree $< N$.*

- The proposition above yields the following observation which we state as a theorem to emphasize its role in providing an answer to the question why and when are deep networks better than shallow networks:

Theorem: *For approximation error, a lower bound in the number of parameters for the shallow network is whereas for the deep binary network matched to the compositional structure of the function the bound is; The VC dimension of the shallow network is, whereas the VC dimension of the deep network is.*

Thus both number of parameters and VC dimension are *exponential in* for the shallow network and *polynomial in* for the deep network. Notice that compositionally is

related to sparsity of the approximating polynomial; sparsity however does not imply compositionality in the way defined here.

The above results are just the beginning of a study of multilayer function approximation techniques and related machine learning results that are likely to be interesting and useful. For instance, it may turn out that the emphasis on (Tikhonov) regularization for learning should be restricted to small data sets and non-compositional functions. Notice that in classical kernel methods there are as many parameters as data points and this implies that regularization is needed (otherwise the estimation of the coefficients may be ill-conditioned because of noise).

Importance of Science of Intelligence

The next phase in the developing tale of deep learning has to be a scientific phase, before another engineering cycle. I sketched some of the scientific questions in both neuroscience and mathematics. There may be even be a deep connection between the two fields, as mentioned in (16): "*A comparison with real brains offers another, and probably related, challenge to learning theory. The "learning algorithms" we have described in this paper correspond to one-layer architectures. Are hierarchical architectures with more layers justifiable in terms of learning theory? It seems that (classical)...learning theory... does not offer any general argument in favor of hierarchical learning machines for regression or classification. This is somewhat of a puzzle since the organization of cortex — for instance visual cortex -- is strongly hierarchical. At the same time, hierarchical learning systems show superior performance in several engineering applications...."*

Beyond the case of Deep Learning, I believe that a concentrated effort on the basic Science of Intelligence and not only on the Engineering of Intelligence is high priority for our society. We need to do it for basic curiosity driven research, for the engineering of tomorrow and for understanding how to build robots and artificial intelligences that are ethical.

Acknowledgment

The author is supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF – 1231216.

References

1. D.H. Hubel, T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex The Journal of Physiology 160, 1962.

2. K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4):193-202, 1980.
3. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature 521, 436–444, 2015
4. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541{551, 1989
5. M. Riesenhuber, T. Poggio. Models of object recognition. Nature Neuroscience, 3(11), 2000.
6. T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, T. Poggio, A Quantitative Theory of Immediate Visual Recognition. In: Progress in Brain Research, 2007
7. P. Niyogi, T. Poggio, F. Girosi. Incorporating Prior Information in Machine Learning by Creating Virtual Examples. In: IEEE Proceedings on Intelligent Signal Processing, Vol. 86, No 11, 2196-2209, 1998
8. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition," arXiv preprint arXiv:1512.03385v1 [cs.CV] 10 Dec 2015 , 201
9. D.L.K. Yamins , J.D. Dicarlo Using goal-driven deep learning models to understand sensory cortex, Nature Neuroscience, 19,3, 356-365, 2016
10. T. Poggio, The Levels of Understanding framework, revised. Perception, volume 41, 1017 - 1023, December 2012
11. T. Poggio, sections with J. Mutch, J.Z. Leibo, L. Rosasco, The Computational Magic of the Ventral Stream: Towards a Theory, Nature Precedings, 2011
12. F. Anselmi, J.Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio, Unsupervised learning of invariant representations. Theoretical Computer Science, 2015.
13. X. Wang, A. Gupta, Unsupervised Learning of Visual Representations using Videos, CVPR, 2015
14. F. Anselmi,T. Poggio. Visual Cortex and Deep Networks, MIT press, Cambridge, MA, 2016
15. H. Mhaskar, Q. Liao, T. Poggio. Learning Real and Boolean Functions: When Is Deep Better Than Shallow, 2016
16. T. Poggio, S. Smale, The Mathematics of Learning: Dealing with Data, Notices of the American Mathematical Society (AMS), Vol. 50, No. 5, 537-544, 2003