The challenge of the deep learning paradigm to the scientific method: hierarchy between models and identification of fundamental laws from algorithmic information theory

Enrico Prati
Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, Piazza Leonardo da Vinci 32, I-20133 Milano, Italia

Planets do not precisely follow the Newton's law because it is a first order approximation of general relativity. In what sense do we intend that planets move according to general relativity? The laws of physics are extrapolated by complex sets of experiments and they are considered fundamental the more they allow to describe concisely acceptable ideal approximations of real systems. The subsequent mathematical description, combined with initial conditions, enables to predict time evolution (or time-dependent probabilities in quantum physics). Currently, general relativity is the best model to fit the orbit of planets once external perturbation are estimated and subtracted. Recently the deep learning paradigm is challenging scientific method thanks to its Bayesian capability to account for complicated systems when complexity sets in, based on the exploitation of huge amount of data. Deep learning ignores Newton's and Einstein's laws but it is significantly much more efficient for instance to design the best trajectory to send a spacecraft from the Earth to Mars in the shortest time. I define a criterion based on algorithmic information theory to assess a hierarchy between scientific models to define the fundamental one, and to compare how different methods such as scientific method, deep learning, and non-scientific methods perform in the description of a set of experiments.

I. CREATION AND ELIMINATION OF NEW CATEGORIES IN SCIENCE

The description of physical world is based on the assumption that it is governed by a limited set of mathematical laws determining how physical quantities of one or more objects change as a function of the others. Their existence is inferred by indirectly probing the microscopic world by special experimental conditions, like particles measured by large arrays of detectors. Also at a macroscopic level most of the natural phenomena can be well approximated by mathematical laws, even in the case of chaotic or random (like Brownian) motion. The physical laws have been postulated after systematically carrying experiments in disparate fields (gases, light, dynamics of rigid bodies) and they have been progressively combined together during the centuries. This process has been possible by identifying pairs of quantities, related to apparently different fields by what Einstein called *conversion constants* which depend on the choice of the unit system (differing from *universal constants*, such as e^2/hc, which are irrespective of unit system). Therefore heat has been reduced as a kind of energy, the square of both electric and magnetic field have been identified with energy density, frequency of light and mass of particles have been notoriously identified with energy by Max Plank and Albert Einstein respectively. Electromagnetic and weak forces have been next unified into a single mathematical electroweak model. Grand Unified Theories intend to describe electroweak and strong forces as the low energy manifestation of the same interaction. Many efforts have been done in the years to describe such Grand Unified Theory and gravity by a single theory.[1] On the other side, generalizing the description and extending the understanding may require to introduce new categories. Even energy conservation may be viewed as a special case suitable when measuring quantities at macroscopic timescale, in which virtual particles can live long enough because they are on the mass shell. Following the view of Griffiths [2] all the particles are virtual, but some of them live long enough to be measured because they are tuned on the mass shell so we call them real as they 'antropically' obey to energy conservation law. It is well known that energy conservation is violated even in low energy physics experiments, such as in elastic and inelastic cotunneling in quantum dots, in processes lasting less than the time window set by the Heisemberg's indeterminacy [3].

At least two key ingredients made the evolution of theories in Physics possible: the general description of apparently different manifestations of some effects by using a larger set of abstract concepts, the new categories; and the identification of some of such categories by postulating they are the same. The process is therefore made by the continuous creation of new categories and the elimination of those redundant, once the mathematical maturity of the model makes such identification natural.

Let's focus on the first ingredient in the following by an example, suitable to be developed also in the next section. If we track the position of the planets in the night sky, we need a very complex model to describe it, corresponding to the epicycles model of Ptolemy. Using the reference frame of the Earth which orbits around the Sun, it is the most straightforward choice but the description of the motion becomes much simpler if we use a reference frame centered in the Sun and whose axis are bound to stars. Kepler then discovered that each orbit is elliptic. After a while other regularities were also observed, like the time to cover an area of such an ellipse of each planet (II Kepler's law), and even more surprisingly time and the axis of the ellipses of different planets are linked together (III Kepler's law) like if there is a common rule governing the motion of all of them. These discoveries, allowed by the change of reference frame to describe the orbits, paved the way to deriving an extremely simple equation not based just on the position of the planet, but instead by assuming that weight and mass are two distinct entities, linked by the concept of acceleration (the property of changing the speed). By extending the number of categories (position in space, time, weight) and by generalizing the concept of weight into force as mass times acceleration (the II Newton's law), it became possible to develop the differential calculus and to obtain the orbits as a beautiful consequence of a mathematically compact equation, populated by random values of the mass of the planets and apparently random distances.

Today, the common understanding is that forces are all apparent manifestation of gauge theories and of the choice of the reference frame, while mass is another form of energy, so they look not fundamental at all. Therefore, it is surprising that, starting from macroscopically measurable quantities such as space, time and weight, corresponding to our coarse grained categories - some of which already identified by Aristotle in Greece in III century BC and by Kanada in ancient India in IV-II century BC, by using proxy 'artificial' categories such as mass and acceleration, we become able to fully predict position of planets in time forever in the future and in the past. The mathematical rule approximately holds because the Newton's equation of gravity is a first order approximation of the Einstein's general relativity mathematical model, a more consistent (see the issue of the instantaneous action of the gravity in the Newton's theory) and coordinate-independent theory, where forces are apparent. General relativity is believed not being the ultimate theory of gravitation, but the classical low energy manifestation of a more fundamental theory of quantum gravity, where space and time emerge. The point is that theories evolved by enlarging their mathematical basis thanks to more complex mathematical tools and by adding new categories to those in use, and progressively bridging some of them consistently under the same category. The resulting compact mathematical description does not imply that it is necessarily simple to apply and sometimes it requires heavy use of computational power to achieve an approximate solution.

II. EXPLOITING PHYSICS WITH DEEP LEARNING

Let's now turn to deep learning. Deep learning [4] has been developed by emulating how images are processed in the visual cortex, which allows animals to recognize objects, faces and shapes irrespectively from their rotation, position, background, illumination, contrasts. Deep learning is based on layered neural networks: the simplest example is the perceptron based on two layers. One layer is constituted by a single neuron (node) and the second layer by a matrix of neurons connected to the first one via synaptic connections with different transmission capability (synaptic weights).

One can conceive two kind of neural networks: feed-forward (nodes of input layer activate those of the next one, recursively ahead up to the output layer: no connections are looping back some outputs to neurons of a previous layer) which is suitable for classification and image recognition; and recurrent neural networks (after feeding the input layer, some of the next nodes are connected to nodes of the previously solicited layers) which are useful to identify time dependent patterns as the information circulates during time in the network so there is room for time-correlation to be revealed and memorized, like for speech recognition [5]. *Deep* indicates that the network, like in the visual cortex, has multiple layers. Such two kind of neural network can be trained (updated) by three different learning methods: *supervised* (when many classified examples are available), *unsupervised* (when the network is asked to build categories on its own, so we can assign a meaning at a later stage once it is trained), and *reinforcement learning.*

The latter has been used by Hutter as a key ingredient to define universal artificial intelligence [6, 7] and it is especially useful when there is a competition between short-term and long-term rewards, so the network is asked to develop the best strategy to achieve the compromise of winning in average on the long term despite apparently neutral or counterintuitive decisions in the short term. This fact is particularly impressive because the optimization is achieved on the medium and long term despite some short term losses during the process.

Both feed forward [8] and recurrent [9] neural network own the general property of being universal approximators thanks to their respective theorems, granting that the networks always allow to discriminate two arbitrarily close input patterns. Deep learning challenges human brain because, if left to run free, it is able to develop its own set of representations to describe structured data and solve the problems. This has been observed in deep learning in several fields: in economy and finance, new unanticipated economic laws and trading patterns [10] have been discovered after unpacking the available data set on novel representations created by the artificial intelligence. In language translation, an artificial intelligence developed an intermediate language [11], which is not human and it exploits combinations of features which have nothing to do with natural languages. The most intuitive example is provided by face recognition: the layers of the neural network initially splits the image in small features (first layer) such as vertical, horizontal, diagonal lines and contrasts, circles, and it recombines them in the next layers progressively in more complex patterns, so each face corresponds at the end to a suitable combination of such basic patterns.[12] Such high level patterns, suitable for unpacking the data, can be extrapolated in any other field where either a process or a law creates order and correlation among them, including scientific data. To explain why deep learning challenges scientific method, we may consider a representative example.

Artificial intelligence has been explored to design space missions launch time and trajectory, to minimize travel time required to artificial satellites and spacecrafts to reach other planets.[13] To train the neural network one would need time- and money-consuming launches, so for this purpose simulation data are suitable to speed up the process, but in principle one does not need to know a mathematical law describing gravitation, as just bare data from experiments are needed. As a result, the neural network unpacks all the possible trajectories by using representations of features built by empirical laws not easily understandable by a human brain. Such features can be obtained by spreading into the next neural network layers the input vector representing for instance the state in the phase space of the planets and of the spacecraft from the initial conditions to the final destination in arrays. Such approach challenges the scientific method which looks for a compact set of mathematical equations to provide understanding and predictions (even if this may be extremely difficult to solve in practice, like Newton's law for more than two bodies). Deep learning allows to extrapolate general features based on the comparison of several trajectories and it is therefore capable to rapidly indicate the best direction and speed of the launch given the position of the Earth and that of the planet to be explored.

Therefore, there are two alternative *methods* to identify the best trajectories in space to land on Mars: solving the Newton's or the Einstein's equation on one side, and to exploit deep learning on the other. Neither the Newton's law nor a deep learning neural network rely on representations to describe the external world which can be proven to be fundamental (mass and acceleration for the former, first neural network layers features for the other). Here it arises the question whether it is possible or not to identify a criterion to return which one of the best *model* provided by the two methods can be considered the fundamental one, sufficiently general to be adopted to address the same queston concerning any other pair of alternative methods (deep learning versus science, astrology versus science) or models within the same method (two different traditional mathematical theories like Newton's and Einsteins's gravitation).

III. ADOPTING THE ALGORITHMIC INFORMATION THEORY TO COMPARE COMPETING METHODS AND AMONG SCIENTIFIC MODELS

Finding both a special and a universal law corresponds to summarize a set of data by the most compact mathematical rules. Raw data are naturally equipped of a method to quantify the degree of their correlation (because they are generated according to a rule) by the concept of order or information, according to the algorithmic theory of information developed first by Kolmogorov [14]. Such quantification corresponds to the negative of entropy or randomness, which in turn own a very powerful definition based on compression of the data by an algorithm. Implicitly, we associate the idea of a theory of being more fundamental with respect to another one when it is simpler, or, in other words, when the algorithm to reproduce the data is smaller, which corresponds to low entropy. Algorithmic complexity straighforwardly enable to compare fundamentalness of models competing to describe a set of data in a continuum of possible values. So, for the case of Ptolemy epycicles, Newton's gravitation and general relativity, if one considers some planets they can all nicely describe the motion of planets seen by Earth, but if we need to account for the motion of all the planets and we want to include the precession of the perielion of Mercury in the model, at a different degree both Ptolemy's model and the Newton's law fail as we should add some additional mathematical effective laws to cover the full Solar system. This action enlarges the algorithm to implement the model, which loses if compared to compactness of general relativity.

The most fundamental model correspond to the most compact in algorithmic sense when arbitrarily extending the set of experimental data to account for.

Because of how science is defined, namely a theory has to be falsifiable, one can not demonstrate that a theory provides an ultimate model, for instance to account for elementary particles. Being more or less fundamental is a relative concept which consists of comparing the models according to their entropy (or alternatively their information). The concept of being fundamental is naturally quantified by algorithmic information theory and it applies not only to unification theories of elementary particle and to quantum gravity, but instead it can be used to compare between models belonging to the scientific method and to different methods such as scientific method, deep learning method, astrology, and other non-scientific methods. According such definition it becomes easy to show that science is more fundamental than deep learning, which requires large neural networks instead of mathematical laws based on algorithms zipped in few bits, and both are far ahead if compared to non-scientific methods.

To conclude, I've considered the example of describing the motion of planets, which allows to consider radically different methods (astrology, science, deep learning) and different models within the same method (Ptolemy's, Newton's, Einstein's theory) to introduce the problem of determining what can be fundamental. By using algorithmic information theory it is possible not only to compare which method provides the most fundamental models (defining a hierarchy given here by astrology, deep learning, scientific method respectively) but also which scientific model is the most fundamental, or, in other words, the most compact when arbitrarily extending the set of experimental data. Such property is therefore relative to the method and to the model so the assumption that there is an ultimate fundamental theory is an extrapolation supported by the belief that Physics is governed, like Leonardo da Vinci anticipated and Galileo Galilei postulated, by a set of mathematical equation.

---

[1] Chamseddine, A. H. and Connes, A. (2010). Noncommutative geometry as a framework for unification of all fundamental interactions including gravity. Part I. Fortschritte der Physik, 58(6), 553-600.
[2] Griffiths, D. (2008). Introduction to elementary particles. John Wiley Sons.
[3] Prati, E. (2009). Virtual states and decoherence in single impurity atom solid state quantum device. In Journal of Physics: Conference Series (Vol. 174, No. 1, p. 012029). IOP Publishing.
[4] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.
[5] Deng, L., Hinton, G., and Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 8599-8603). IEEE.
[6] Hutter, M. (2000). A theory of universal artificial intelligence based on algorithmic complexity. arXiv preprint cs/0004001.
[7] Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17(4), 391-444.
[8] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural networks, 4(2), 251-257.
[9] Hammer, B. (2000). On the approximation capability of recurrent neural networks. Neurocomputing, 31(1), 107-123.
[10] Takeuchi, L., and Lee, Y. Y. A. (2013). Applying deep learning to enhance momentum trading strategies in stocks. In Technical Report. Stanford University.
[11] Wong, S. (2015). Google Translate AI invents its own language to translate with, New Scientist, 30 November 2016
[12] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
[13] Di Lizia, P., Radice, G., Izzo, D., and Vasile, M. (2005, September). On the solution of interplanetary trajectory design problems by global optimisation methods. In I. Garcia, LG Casado, EMT Hendrix, and B. Toth, xeditors, Proceedings of the International Workshop on Global Optimization (pp. 159-164).
[14] Chaitin, G. J. (1977). Algorithmic information theory. IBM journal of research and development, 21(4), 350-359.