Understanding deep learning is also a job for physicists

Automated learning from data by means of deep neural networks is finding use in an ever-increasing number of applications, yet key theoretical questions about how it works remain unanswered. A physics-based approach may help to bridge this gap.

Lenka Zdeborová

Imagine an event for which thousands of tickets get sold out in under 12 minutes. We are not speaking of a leading show on Broadway or a concert of a rockstar, but about the Conference on Neural Information Processing Systems (NeurIPS) — the principal gathering for research in machine learning and artificial intelligence. The fields related to automated learning from data are experiencing a surge in research activity, as well as in investment. This is largely thanks to developments in a subfield called deep learning, which has led to a myriad of successes in many applications[1,2]. Research in physics is no exception to this claim, and indeed in the recent years we have seen numerous applications of machine learning to various physics problems[3,4], and even more predictions regarding which physics problems we will be able to solve with machine learning in the near future. Some even wonder whether future machine-learning systems will be able to collect suitable data and infer the laws of nature from them entirely automatically.

All this activity and progress naturally comes with many open questions — not least that deep neural networks are often described as black boxes: hard to interpret and without a solid understanding of when they provide satisfactory answers and when they do not. When applying machine learning to problems in physics (and other areas) researchers often wonder: What is the best way to take into account the corresponding domain knowledge, constraints and symmetries? How do we adapt the existing machine-learning tools to new problems, and how to interpret their results in a scientific manner? How do we reliably quantify the uncertainties and errors stemming from the fact that training and testing data may not come from the same source?

One might argue that researchers in mathematics, computer science, statistics and other related fields are working hard to answer such questions, and so for us physicists it is a matter of sitting tight waiting for tools and answers that we can subsequently put to use. In this Comment, I argue that, instead, we need to join the race of searching for these answers, because it is precisely the physicists' perspective and approach that is needed to enable progress in this endeavour.

Three ingredients to decipher deep learning

The engineering details of current deep-learning systems, such as the ones deployed by Google to translate languages[5], can be dauntingly complicated. Yet the basic principle of how learning with deep neural networks works is, in fact, pleasantly simple.

A basic example of a task in machine learning is supervised learning, where the machine learns to associate the correct outputs to input data, based on a database of examples of input–output pairs. Deep learning then uses multi-layer neural networks in which the input data are fed into the first layer, its output then fed as input into the next layer, and so on. Each layer is a multiplication of the input by a matrix of so-called weights, followed by a component-wise non-linear function. This is repeated a number of times corresponding to the number of layers.

For problems with binary output data (for example, 1 for a picture of a dog and −1 for a picture of a cat), the last layer then aims to find a hyperplane separating these output labels. This described structure is called a feed-forward fully connected neural network and is mathematically seen as a function of the input data outputting the labels and being parameterized by the matrices of weights. The weights are then adjusted using a simple gradient descent of a so-called loss function that quantifies the amount of mismatch between the current and desired outputs. Finally, the performance is evaluated against a so-called test dataset that was not seen during the training. Interestingly, the basic design principles of multi-layer neural networks have been known since the early days of research on artificial neural networks[6]. Arguably, the unprecedented engineering progress of the last two decades is largely due to better and larger training datasets and faster computing, such as highly parallelizable GPU processors, rather than due to fundamental improvements in the network architectures or the training algorithms themselves.

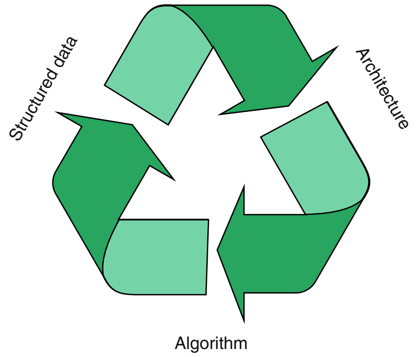

Fig. 1 | Interplay of key ingredients. Building theory of deep learning requires an understanding of the intrinsic interplay between the architecture of the neural network, the behaviour of the algorithm used for learning and the structure in the data.

In 1995, the influential statistician Leo Breiman summarized three main open problems in machine learning theory[7]: "Why don't heavily parameterized neural networks overfit the data? What is the effective number of parameters? Why doesn't back-propagation — the term used for the gradient-descent-based algorithm used to train the state-of-the-art neural networks — get stuck in poor local minima with low value of the loss function, yet bad test error?" While Breiman formulated these questions 25 years ago, they are still open today and subject to most of the ongoing works in the learning-theory community, including numerous papers and countless discussions at the NeurIPS conference.

Let us now try to clarify why the theoretical progress in understanding deep learning so difficult. When describing current deep-learning systems, the interplay of three key ingredients needs to be considered, as depicted in Fig. 1.

Architecture. As sketched out above, neural networks used in deep learning consist of multiple layers, with the number of layers known as the 'depth'. Each layer has a certain dimension, called width, and is associated with a non-linear function, called activation. Layers are of different types, for example, fully connected or convolutional. One always needs to decide how to choose the depth, width, activations or types of layers, which determines the architecture of the neural network.

Algorithm. Given the data and the architecture of the network, one needs an algorithm to set the weight matrices so that the network outputs are the correct ones for previously unseen input data. This is most often done by the minimization of a function on the training set that quantifies the mismatch between current network outputs and the desired ones. The corresponding algorithm, widely known as back-propagation, is based on simple gradient descent of this function with respect to the weights.

Structured data. In supervised learning the training data consist of pairs of input data and labels. For instance, for classification of pictures of dogs and cats, one sample consists of one picture and one label that identifies whether the picture is a dog or a cat (for example, label 1 for a dog and −1 for a cat). A neural network aims to learn the function from inputs to outputs, but crucially only on inputs that are of the same type as those in the training set. The input data are therefore not arbitrary vectors, but ones that have a particular structure, representing a picture of a cat or that of a dog in our example.

None of the above three ingredients can be excluded from consideration when building a theory. Indeed, the network architecture should have multiple layers because the empirical evidence for the superiority of such architectures is overwhelming. Concerning the algorithm, one clearly always needs to ensure computational tractability of the learning problem. In other words, it is not sufficient that there exists a set of parameters providing good performance, this set of parameters needs to be discoverable with efficient algorithms. Concerning the structure of the data, it is known that learning even simple neural networks is computationally prohibitive for the worst-case data[8], thus there must be some property of the data that makes learning tractable.

These three ingredients — architecture, algorithm and structure of data — are intrinsically inter-dependent, since network architectures are chosen so that they represent the structure in the data in a way that is learnable with a given algorithm. There is a growing body of empirical and numerical evidence that the classical learning theory is not able to explain the observed behaviour[9].

Where physics comes in

To illustrate the situation to a physics audience, one could compare the current state of deep learning theory to the physics theory of light and matter in the early twentieth century. There were a lot of results from experiments (such as the photoelectric effect, for example) that could not be explained by the existing theory — quantum mechanics was yet to be developed.

In my opinion, one key difficulty in developing a theory of deep learning stems from the fact that, on the one hand, the existing learning theory has very high standards of mathematical rigour, and on the other hand, the impressive empirical progress has so far been driven by the aim of decreasing the test error rather than by the aim of understanding what is going on.

One could perhaps compare this situation to a physics problem being attacked by a majority of researchers that were either mathematical physicists insisting on fully rigorous proofs or applied industrial-research colleagues whose primary aim is to deliver a product. While both these groups of colleagues are contributing immensely to the progress the field of physics is making, one could argue that, as for the famous example of quantum mechanics, we need the contributions of the experimentalists driven purely by the desire to understand nature and of the theoreticians aiming to explain those experiments by using scientifically sound and principled (but not necessarily fully mathematically rigorous) approaches. To understand deep learning, the machine-learning community needs to fill the gap between the mathematically rigorous works and the end-product-driven engineering progress, all while keeping the scientific rigour intact.

And this is where the physics approach and experience comes in handy. The virtue of physics research is that it strives to design and perform refined experiments that reveal unexpected (yet reproducible) behaviour, yet has a framework to critically re-examine and improve theories explaining the empirically observed behaviour.

In particular, the theoretical part of physics research is largely based on models. Models are a way of capturing the essence of a problem and stripping off the details that are not necessary to explain the experimental observations. An example would be the widely used Ising model of magnetism: it does not capture any details of the quantum mechanical nature of the magnetic interactions, and it also does not contain any details of any specific magnetic material, yet it explains the nature of the transition from a ferromagnet at low temperature to a paramagnet at high temperature.

And as it happens, physicists, particularly the community studying statistical mechanics of disordered systems, recognized the need for the modelling of machine-learning systems more than three decades ago. From a physics point of view, one aims to study a dynamical system with many interacting elements (weights of the network) evolving in structured quenched disorder (given by the data and the data-dependent network architecture). The pioneering works on the Hopfield model[10,11] and on the perceptron model[12,13] were followed by many others, reviewed in refs. [14–18]. But these early works do not provide theory of deep learning because, going back to Fig. 1, they only consider unstructured input data, only shallow networks (at most two layers) and have not analysed in a closed form the gradient-based learning algorithms.
With the successes and promises of deep learning and the theoretical questions that surround it, this direction of research has recently been picked up again. The same basic approach can be redeployed to answer the current questions, and to design experiments that would reveal more of unexpected behaviour and to explain it.

Only in the past year we have seen a range of events dedicated to these topics, where physicists and machine learning scientists have met and exchanged ideas. Notable examples include a program at the Kavli Institute for Theoretical Physics, UCSB (https://go.nature.com/2ygouAa), another program at the Institute of Pure & Applied Mathematics, UCLA (https://go.nature.com/2APixvf), or workshops hosted by the leading machine learning conferences ICML and NeurIPS[19,20]. Ongoing related work is also covered in two recent review articles[4,21], in a dedicated special issue of *Journal of Physics A*[22], or in a collection of statistical physics related articles that were accepted to one of the three leading and highly selective machine-learning conferences[23]. As physicists, let us embrace machine learning as the new tool in the box and let us use it widely and wisely. But let us also keep in mind that understanding why and how it really works requires physics methodology — we should not stand by as this formidable endeavour takes shape. So, let us embrace deep neural networks as a part of our field, and study it with the same unceasing desire that drives our quest to understand the world around us. ❐

Lenka Zdeborová ✉
Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, Gif-sur-Yvette, France.
✉*e-mail: lenka.zdeborova@cea.fr*

Published online: 26 May 2020
https://doi.org/10.1038/s41567-020-0929-2

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436 (2015).
2. Deep learning. *Wikipedia* https://go.nature.com/2XsJf4v (2020).
3. Zdeborová, L. *Nat. Phys.* **13**, 420–421 (2017).
4. Carleo, G. et al. *Rev. Mod. Phys.* **91**, 045002 (2019).
5. Vaswani, A. et al. In *Advances in Neural Information Processing Systems 30* 5998–6008 (NIPS, 2017).
6. Schmidhuber, J. *Neural Netw.* **61**, 85–117 (2015).
7. Breiman, L. In *The Mathematics of Generalization* 11–15 (Addison-Wesley, 1995).
8. Blum, A. & Rivest, R. L. In *Advances in Neural Information Processing Systems* 494–501 (1989).
9. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Preprint at https://arxiv.org/abs/1611.03530 (2016).
10. Hopeld, J. J. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
11. Amit, D. J., Gutfreund, H. & Sompolinsky, H. *Phys. Rev. Lett.* **55**, 1530–1533 (1985).
12. Gardner, E. *J. Phys. A* **21**, 257–270 (1988).
13. Gardner, E. & Derrida, B. *J. Phys. A* **21**, 271–284 (1988).
14. Mézard, M., Parisi, G. & Virasoro, M. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* Vol. 9 (World Scientific, 1987).
15. Hertz, J., Krogh, A. & Palmer, R. G. *Introduction to the Theory of Neural Computation* (Addison-Wesley, 1991).
16. Seung, S., Sompolinsky, H. & Tishby, N. *Phys. Rev. A* **45**, 6056–6091 (1992).
17. Watkin, T. L. H., Rau, A. & Biehl, M. *Rev. Mod. Phys.* **65**, 499–556 (1993).
18. Engel, A. & Van den Broeck, C. P. L. *Statistical Mechanics of Learning* (Cambridge Univ. Press, 2001).
19. Theoretical physics for deep learning. *Workshop at the 36th International Conference on Machine Learning* https://go.nature.com/36gSRDb (ICML, 2019).
20. Machine learning and the physical sciences. *Workshop at the 33rd Conference on Neural Information Processing Systems* https://go.nature.com/2Xd16w1 (NeurIPS, 2019).
21. Bahri, Y. et al. *Ann. Rev. Cond. Matt. Phys.* **11**, 501–528 (2019).
22. Agliari, E. et al. (eds). *Machine Learning and Statistical Physics: Theory, Inspiration, Application; J. Phys. A* (IOP, 2020); https://go.nature.com/2ZlvUNN
23. Mezard, M. et al. (eds). *Machine Learning 2019*; *J. Stat. Mech.* (2019); https://go.nature.com/2XkTCY2