einstein
Official Publication of the Instituto Israelita de Ensino e Pesquisa Albert Einstein

ISSN: 1679-4508 | e-ISSN: 2317-6385

EDITORIAL

Unpacking the black box

Desembalando a caixa preta

Thiago Gonçalves dos Santos Martins[1], Paulo Schor[2]

[1] Universidade de Coimbra, Coimbra, Portugal.
[2] Universidade Federal de São Paulo, São Paulo, SP, Brazil.

DOI: 10.31744/einstein_journal/2021ED6037

Today there are large databases (Big Data) of electronic records and digital images that allow pattern recognition in large volumes of information in a short period of time, contributing to the creation of a personalized Medicine with the aid of artificial intelligence.[1,2]

The algorithms used in medical diagnosis software are formed by a deep neural network, consisting of several layers that resemble the biological processes of the human cortex. Each neuron corresponds to a specific stimulus for a region within an image, similar to the way the brain neuron would respond to visual stimuli, activating a specific region of visual space. The neural network learns automatically by conferring values to these filters, even though specific parameters such as number of filters, filter size, and network architecture still need to be defined before the training phase.

The phase in which the device trains to improve its predictions is that of learning, which can be divided into two types: supervised and unsupervised. In supervised learning, information is assigned to the training data as they are entered into the computer, while in unsupervised learning, the device creates its own input algorithm. The training phase is followed by data validation, in which the figures should ideally be different from those used in the previous phase. Now the machine is able to classify the data and start making predictions. For this training phase, an algorithm needs thousands of data to reach acceptable accuracy.[3,4]

However, many researchers still cannot explain how the algorithms reached certain conclusions, and this slightly diminishes the confidence of the scientific world in the use of artificial intelligence in Medicine, since we tend to refute what we cannot explain. This term is known as "black box," in which we do not have access to the information of the internal design and implementation of the algorithm. This term is opposed to "white box," in which the component is completely exposed to the user. Among these algorithms, there is the "gray box," which is when we only have access to some data.[5]

The most developed algorithms are formed by a non-linear structure with several layers in its architecture, enabling high complexity predictions. Nevertheless, this makes it difficult to explain how we get a result, which is easier in simpler and linear algorithms.[6]

Access to information of the algorithms is difficult, since it is not stored in a specific place, as is the human memory that - during learning - forms several synapses. Learning of the machines ends up forming these learning paths, attributing different weights to the filters during the training phase.

Unveiling the black box could create the possibility of seeing Medicine in a different way. Some animals, like the hummingbird, have a number and

How to cite this article:
Martins TG, Schor P. Unpacking the black box. einstein (São Paulo). 2021;19:eED6037.

Corresponding author:
Thiago Gonçalves dos Santos Martins
Rua Botucatu, 821 – Vila Clementino
Zip code: 04023-062 –São Paulo, SP, Brazil
Phone: (55 11) 5521-2571
E-mail: thiagogsmartins@yahoo.com.br

type of cones different from human beings, and develop the ability to identify colors imperceptible to our eyes, enabling them to see the world differently.[7] Similarly, we can mention the function of algorithms, whose processing logic is not the same as that of human beings. From this point of view, even attesting to the results achieved by the algorithms, we are reluctant to trust the results since we do not understand the path taken. When we can explain the function of the algorithms, we can justify their decisions and minimize errors, thus making their vulnerability more evident. Learning with the machine can be exemplified with the development of programs such as AlphaGo, which allowed the teaching of new techniques of playing Go. This board game has been unveiled by machine learning, having a performance superior to that of humans.[8] In this way, by learning from the machines, we can expand our knowledge in other areas of science.

Some studies were carried out trying to explain the functioning of algorithms. Caruana et al.,[9] described an algorithm for diagnosis of pneumonia and studied its behavior with cases of real patients. One must bear in mind that the algorithms created for the diagnosis and monitoring of diseases should preferably be validated with data from different populations, which were used in the training stage and with the participation of researchers who were independent, not taking part in development.[10]

Thus, the knowledge of how the human brain works can help us better understand the black box. A better understanding of machine learning can improve our ways of comprehending the world, enabling us to see Medicine with "hummingbird eyes", and to begin to appreciate the mysteries of Medicine that, until then, have not been solved. In addition, it allows the development of better algorithms that need fewer examples for their learning.

AUTHORS´ INFORMATION

Martins TG: http://orcid.org/0000-0002-3878-8564
Schor P: http://orcid.org/0000-0002-3999-4706

REFERENCES

1. Martins TG, Costa AL. A new way to communicate science in the era of Big Data and citizen science. einstein (São Paulo). 2017;15(4):523.
2. Martins TG, Costa AL, Martins TG. Big Data use in medical research. einstein (São Paulo). 2018;16(3):eED4087.
3. Martins TG, Francisco Kuba MC, Martins TG. Teaching ophthalmology for machines. Open Ophthalmol J. 2018;12:127-9.
4. Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Applications of artificial intelligence in ophthalmology: general overview. J Ophthalmol. 2018;2018:5278196. Review.
5. Suman RR, Mall R, Sukumaran S, Satpathy M. Extracting state models for Black-Box software components. J Object Technol. 2010;9(3):79-103.
6. Tan S, Sim KC, Gales M. Improving the interpretability of deep neural networks with stimulated learning. IEEE Workshop Autom, Speech Recognit. 2015;617-23.
7. Herrera G, Zagal JC, Diaz M, Fernández MJ, Vielma A, Cure M, et al. Spectral sensitivities of photoreceptors and their role in colour discrimination in the green-backed firecrown hummingbird (Sephanoides sephaniodes). J Comp Physiol A Neuroethol Sens Neural Behav Physiol. 2008;194(9):785-94.
8. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. 2017; 550(7676):354-9.
9. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015. Pages 1721-1730 [cited 2020 Dec 3]. Sydney, NSW, Australia; 10-13 Aug. Available from: https://dl.acm.org/doi/10.1145/2783258.2788613
10. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. Transl Vis Sci Technol. 2020;9(2):7. Erratum in: Transl Vis Sci Technol. 2020; 9(9):33.