Artificial Cognition:
An emergent discipline to explain decision making in artificial neural networks

Alfonso Iglesias
Accenture Technology Consulting

Tipo de artículo: Actualidad, Multilingüe.
Disciplinas: Inteligencia Artificial, Psicología, Neurociencia.
Etiquetas: aprendizaje profundo, redes neuronales artificiales, conexionismo.

Machine learning based on artificial neural networks has fueled the recent rise of artificial intelligence. However, it is not easy to explain the decision making of these models, which can lead to ethical, legal and technology adoption problems. Artificial Cognition takes advantage of the methods of cognitive science to explain the decision making of the most complex systems to interpret of artificial intelligence.

(istock) ipopba.

Artificial neural networks do not follow the traditional programming logic of expert systems in which it is explicitly stated under which input conditions a given output should be returned based on a knowledge base and a series of rules. They learn by themselves from experience (supervised learning, reinforcement learning) through training, and can discover complex structures in data that vary in many dimensions. Their performance is comparable or superior to that of humans in tasks that were not computable until recently. They have revolutionized fields such as computer vision or natural language processing and are now applied to object recognition, medical diagnosis, autonomous driving, fraud detection or trend and purchase analysis.

Iglesias, A. (2022). Artificial Cognition: An emergent discipline to explain decision making in artificial neural networks. *Ciencia Cognitiva*, 16:1, 14-17.

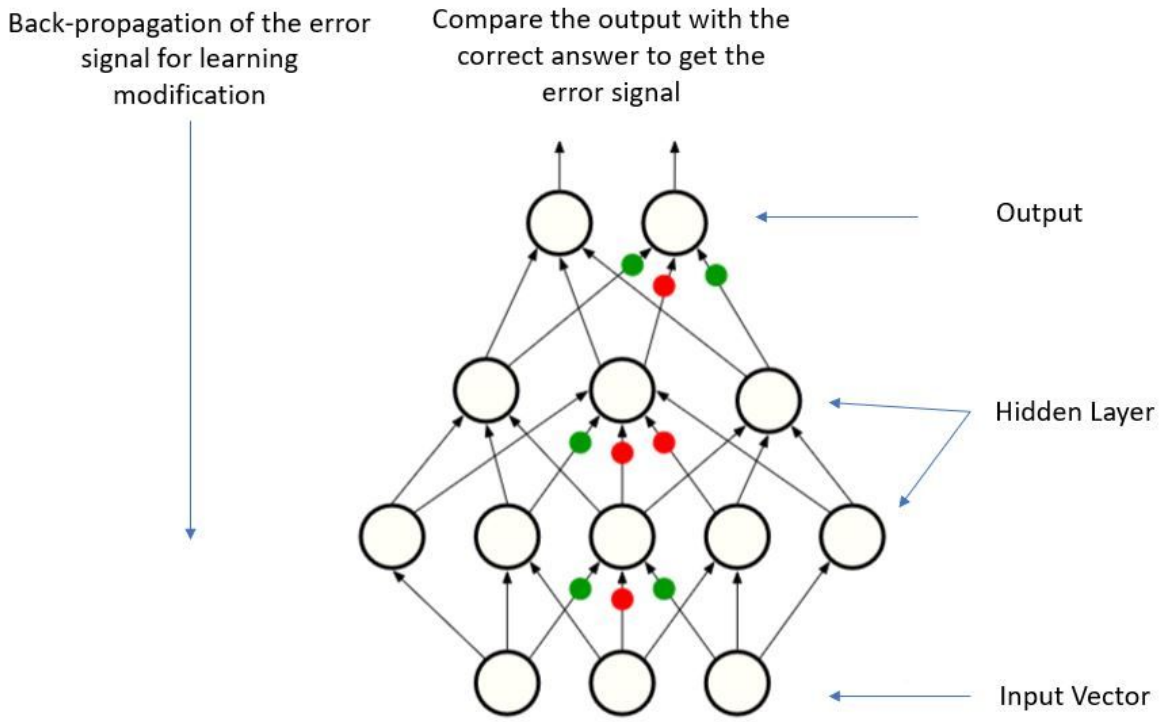

Figure 1.- Schematic figure of an artificial neural network. During an image classification training, the model is presented with an image (input), from which it creates an input vector. After propagating the information through the network, it finally produces an output in the form of a vector of scores, which identifies the categories of objects that we want it to recognize. Initially, the network sorts randomly. For the network to learn, it is necessary to compute the degree of error (or distance) between the network output and the correct pattern of scores. The model then uses a backpropagation algorithm to adjust the weights of the connections between the nodes, so that the error is reduced during training.

These networks are made up of nodes (artificial neurons) organized in layers and connections between them. Each connection, in an analogy with the synapses in a biological brain, transmits information in the form of a "signal" to other neurons. The signal at a given connection is a number, and the output of each neuron is computed by some nonlinear function of the sum of its inputs. The connections have a weight that adjusts as learning progresses thanks to a backpropagation algorithm (Figure 1).

The number of layers and parameters of deep neural networks has been growing with the continuous development of new software architectures. Each layer of these networks represents the knowledge extracted from the learning data at progressively more abstract levels. Thus, the training generates intermediate, subsymbolic representations, and the internal units can represent properties such as horizontal lines, but also more complex elements, or more complex to define, of the structure of an image.

Artificial neural networks are a “black box”. Refining themselves through learning, they make decisions based on parameters that the programmer has not defined and cannot deduce by looking at the output or the network code. Two networks with identical architecture may behave differently depending on the value of the initial random weights or the learning data.

The “explainability” of artificial neural networks becomes more important as their use spreads, hence the rise of Explainable Artificial Intelligence (XAI), a discipline that aims to improve the understanding of decision-making in artificial intelligence systems. Traditionally, the techniques used in XAI are based on the use of models that are easier to interpret or other techniques such as the visualization of the activation of layers and

neurons (see Ortiz-Tudela, 2021, http://www.cienciacognitiva.org/? p=2104) or ablation analysis that are inspired by neuroscience.

It has been proposed recently to study artificial intelligence systems not as engineering artifacts, but as "a class of actors with particular behaviour patterns and ecology" (Rahwan et al., 2019, p. 477). In this direction, Taylor et al. (2020) have proposed taking advantage of the rigor, experimental methodology and experience of psychology in the study of another black box, the human mind. This approach, Artificial Cognition, employs the experimental method of psychology, that is, controlled stimuli and measurement of behaviour in one or different network architectures to make causal inferences about the structure, architecture, and functioning of the "mind."

A good example of this is the research by Ritter et al. (2019), who tested this approach in a study using state-of-the-art neural networks on a task that consists of labelling a test image as belonging to a new category after a single example. Humans learn new concepts with very little supervision (a child can generalize the concept of "giraffe" from a single image), and research in developmental psychology shows that when learning new words, humans tend to assign the same name to objects with similar shapes, rather than to objects with other similar characteristics, such as colour, texture, or size (so-called shape bias). This and other biases help people eliminate unlikely hypotheses when inferring the meaning of new words (Marr, 1982).

The authors assumed that at least part of the hypothesis elimination theory can be extrapolated to artificial neural networks and asked: What predictive properties do the networks use? Do different networks use the same properties? Are these properties interpretable by humans?

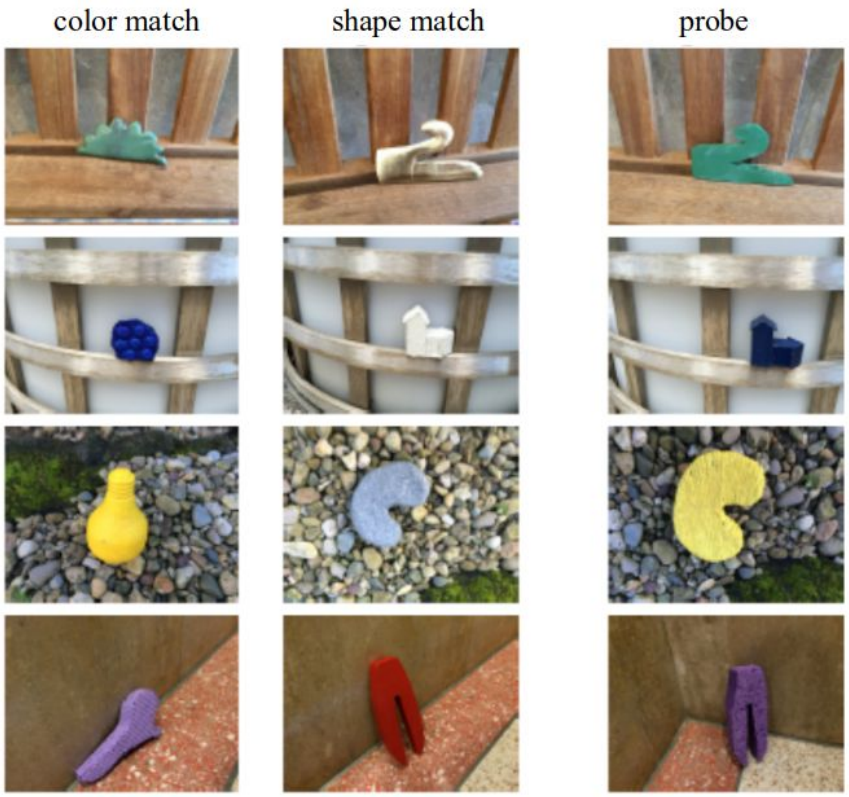

Figure 2.- Example images of the data set. The data consists of image triplets, each of the rows containing an image that matches colour (left column), an image that matches shape (middle column), and the test image (right column). These combinations were used to calculate the shape bias, based on the proportion of times a model assigns the shape-matching class to the test image. (c) Samuel Ritter (DeepMind). Reproduced with permission.

Ritter and collaborators investigated two network architectures: one called “matching networks”, which has shown the best performance in this task, and an initial reference model. Following in the footsteps of the original studies with children, they instructed the models to identify the image most similar to the learning image from a new set that included shape-matched stimuli and colour-matched stimuli (Figure 2; Landau et al., 1988). The results showed that the network was much more likely to identify the shape-matching novel object as belonging to the same category, confirming a shape bias similar to that exhibited by humans. There was also a high variability in this bias: a) between different networks, b) during the training process, and c) in the same networks

initialized with different random weights, showing that otherwise identical networks converge to qualitatively different solutions.

What we aim to highlight here is "the ability of cognitive psychology tools to expose hidden computational properties of deep neural networks" (Ritter et al, 2019). In this case, the shape of the object can refer to tumour nodes or any medical imaging problem, hence its importance. As the applications and use of artificial neural networks spread, their explainability is more urgent for society, which demands to understand how and on what basis artificial intelligence makes decisions that affect it.

References

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, 521, 436–444.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.

Rahwan, I., Cebrian, M., Obradovich, N., et al. (2019). Machine behaviour. *Nature*, 568, 477-486.

Taylor, J., & Taylor, G. (2020). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28, 454-475.

Ritter, S., Barrett, D., Santoro, A., & Botvinick, M. (2017). Cognitive psychology for Deep Neural Networks: A shape bias case study. *Proceedings of the 34th International Conference on Machine Learning,* 70, 2940-2949.

Manuscript received on November, 15th, 2021.
Accepted on March 29th, 2022.

This is the English version of

Iglesias, A. (2022). Cognición Artificial: Una disciplina emergente para explicar la toma de decisiones de las redes neuronales artificiales. *Ciencia Cognitiva*, 16:1, 10-13.