Cognición Artificial: Una disciplina emergente para explicar la toma de decisiones de las redes neuronales artificiales

Alfonso Iglesias
Accenture Technology Consulting

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

El aprendizaje automático basado en redes neuronales artificiales ha propiciado el reciente auge de la inteligencia artificial. Sin embargo, no es fácil explicar la toma de decisiones de estos modelos, lo que puede conllevar problemas éticos, legales y de adopción de la tecnología. La Cognición Artificial aprovecha los métodos de la ciencia cognitiva para explicar la toma de decisiones de los sistemas más complejos de interpretar de la inteligencia artificial.

(istock) ipopba.

Las redes neuronales artificiales no siguen la lógica de programación tradicional de sistemas expertos en la que se indica explícitamente en qué condiciones de entrada ("input") se debe devolver un resultado ("output") en función de una base de conocimiento y una serie de reglas. Aprenden por sí mismas de la experiencia (aprendizaje supervisado, aprendizaje por refuerzo) mediante entrenamiento, y son capaces de descubrir estructuras complejas en datos que varían en muchas dimensiones. Su rendimiento es comparable o superior al humano en tareas que hasta hace poco no eran computables. Han revolucionado campos como la visión por ordenador o el procesamiento del lenguaje

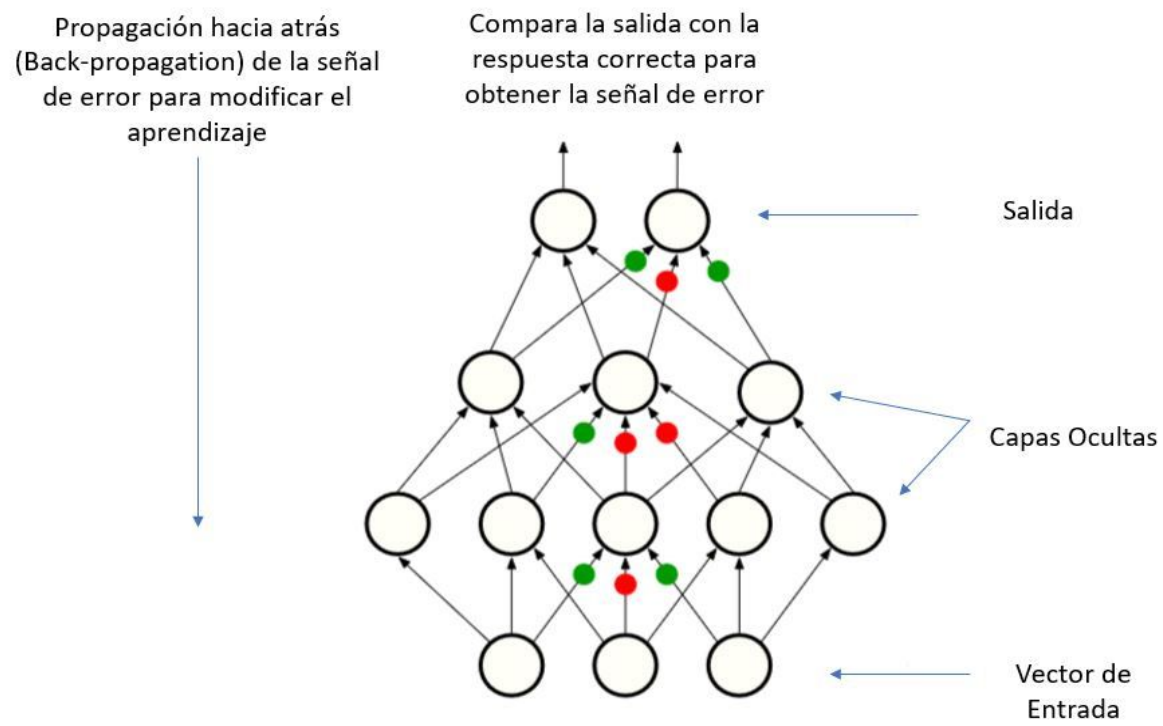

Figura 1.- Figura esquematizada de una red neuronal artificial. Durante un entrenamiento para la clasificación de imágenes de objetos, al modelo se le presenta una imagen (input), de la que crea un vector de entrada. Tras propagar la información por la red, en último lugar produce una salida en forma de vector de puntuaciones, el cual identifica las categorías de objetos que queremos que reconozca. Inicialmente, la red clasifica al azar. Para que la red aprenda, es necesario computar el grado de error (o distancia) entre la salida de la red y el patrón de puntuaciones correcto. Entonces, el modelo utiliza un algoritmo de retropropagación para ajustar los pesos de las conexiones entre los nodos, de modo que el error se reduzca durante el entrenamiento.

natural y se aplican hoy día al reconocimiento de objetos, el diagnóstico médico, la conducción autónoma, la detección del fraude o el análisis de tendencias y compras.

Estas redes están formadas por nodos (neuronas artificiales) organizados en capas y conexiones entre ellas. Cada conexión, en una analogía con las sinapsis de un cerebro biológico, transmite información en forma de "señal" a otras neuronas. La señal en una conexión dada es un número y la salida de cada neurona se calcula mediante alguna función no lineal de la suma de sus entradas. Las conexiones tienen un peso que se ajusta a medida que avanza el aprendizaje gracias a un algoritmo de retropropagación (Figura 1).

El número de capas y parámetros de las redes neuronales profundas ha ido creciendo con el desarrollo continuo de nuevas arquitecturas de software. Cada capa de estas redes representa el conocimiento extraído de los datos de aprendizaje a niveles progresivamente más abstractos. Así, el entrenamiento genera representaciones intermedias, subsimbólicas, y las unidades internas pueden representar propiedades como líneas horizontales, pero también elementos más complejos, o más complejos de definir, de la estructura de una imagen.

Las redes neuronales artificiales son una "caja negra". Al refinarse a sí mismas a lo largo del aprendizaje, toman decisiones basándose en parámetros que el programador no ha definido y no puede deducir mediante la observación del resultado o el código de la red. Dos redes con idéntica arquitectura se pueden comportar de manera diferente en función del valor de los pesos aleatorios de inicio o de los datos de aprendizaje.

La interpretabilidad o “explicabilidad” de las redes neuronales artificiales cobra mayor importancia conforme su uso se extiende, de ahí el auge de la Inteligencia Artificial Explicable (IAE), disciplina que pretende mejorar la comprensión de la toma de decisiones de los sistemas de inteligencia artificial. Tradicionalmente, las técnicas usadas en IAE se basan en el uso de modelos más fáciles de interpretar u otras técnicas como la visualización de la activación de capas y neuronas (véase Ortiz-Tudela, 2021, http://www.cienciacognitiva.org/?p=2104) o el análisis de ablación, que están inspiradas en la neurociencia.

Recientemente, se ha propuesto estudiar los sistemas de inteligencia artificial no como artefactos de ingeniería, sino como “una clase de actores con patrones de comportamiento y ecología particulares” (Rahwan y col., 2019, p. 477). En esta dirección, Taylor y col. (2020) propusieron aprovechar el rigor, la metodología experimental y la experiencia de la psicología en el estudio de otra caja negra, la mente humana. Este enfoque, la Cognición Artificial, emplea el método experimental de la psicología, es decir, estímulos controlados y medición de la conducta en una o diferentes arquitecturas de red para hacer inferencias causales sobre la estructura, la arquitectura y el funcionamiento de la “mente” artificial.

Un buen ejemplo de ello es la investigación de Ritter y col. (2019), quienes pusieron a prueba este enfoque en un estudio con redes neuronales de última generación en una tarea que consiste en etiquetar una imagen de prueba como perteneciente a una nueva categoría después de un único ejemplo. Los humanos aprendemos nuevos conceptos con muy poca supervisión (un niño puede generalizar el concepto de "jirafa" a partir de una sola imagen) y la investigación en psicología del desarrollo muestra que, al aprender nuevas palabras, los humanos tendemos a asignar el mismo nombre a objetos con formas similares, en lugar de a objetos con otras características similares, como color, textura o tamaño (el llamado sesgo de forma). Este y otros sesgos ayudan a las personas a eliminar hipótesis improbables al inferir el significado de nuevas palabras (Marr, 1982).

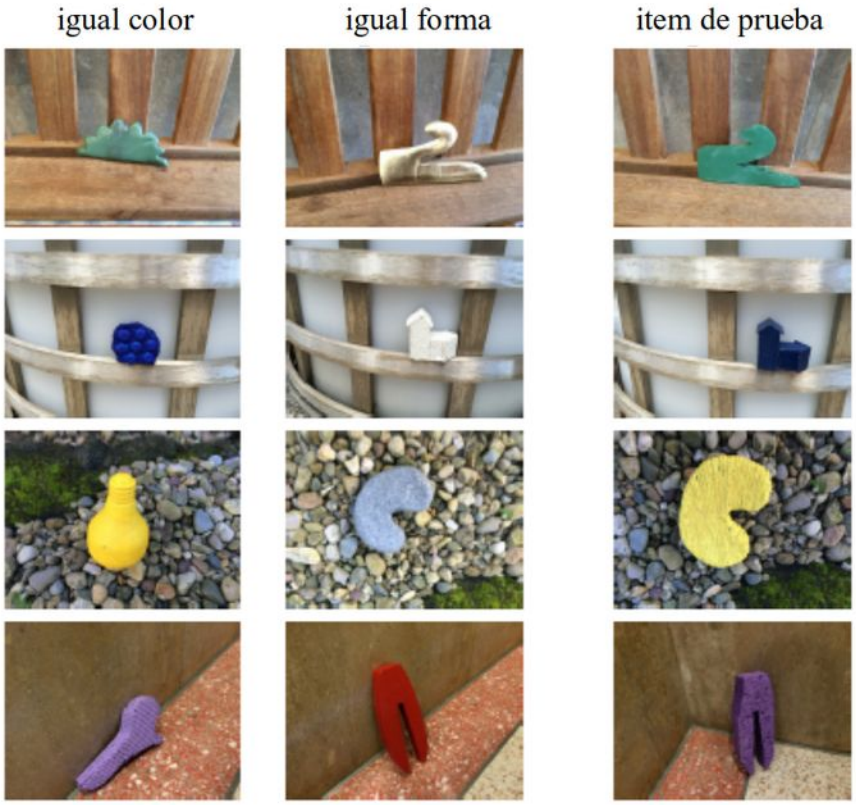

Figura 2.- Imágenes de ejemplo del conjunto de datos. Los datos consisten en tripletes de imágenes, cada una de las filas contiene una imagen que coincide en color (columna izquierda), una imagen que coincide en forma (columna central) y la imagen de prueba (columna derecha). Se utilizaron estas combinaciones para calcular el sesgo de forma, según la proporción de veces que un modelo asigna la clase que coincide en forma a la imagen de prueba. (c) Samuel Ritter (DeepMind). Reproducido con permiso.

Los autores asumieron que, al menos parte, de la teoría de la eliminación de hipótesis se puede extrapolar a las redes neuronales artificiales y se preguntaron: ¿qué propiedades predictivas usan las redes? ¿Diferentes redes usan las mismas propiedades? ¿Son estas propiedades interpretables para los humanos?

Ritter y colaboradores investigaron dos arquitecturas de red: una llamada “matching networks”, que ha mostrado el mejor rendimiento en esta tarea, y un modelo inicial de referencia.

Siguiendo los pasos de los estudios originales con niños, instruyeron a los modelos a identificar la imagen más similar a la imagen de aprendizaje de entre un nuevo conjunto que incluía estímulos con coincidencia de forma y estímulos con coincidencia de color (Figura 2; Landau y col., 1988). Los resultados mostraron que era mucho más probable que la red identificase el objeto novedoso que coincide en la forma como perteneciente a la misma categoría, lo que confirma un sesgo de forma similar al que presentan los humanos. Hubo también una alta variabilidad en este sesgo: a) entre distintas redes, b) durante el proceso de entrenamiento, y c) en las mismas redes inicializadas con pesos aleatorios diferentes, demostrando que redes, por lo demás idénticas, convergen en soluciones cualitativamente diferentes.

Lo que se pretende destacar aquí es "la capacidad de las herramientas de la psicología cognitiva para exponer propiedades computacionales ocultas de las redes neuronales profundas" (Ritter y col., 2019). En este caso, la forma del objeto puede hacer referencia a los ganglios tumorales o cualquier problema de imagen médica, de ahí su importancia. A medida que las aplicaciones y uso de las redes neuronales artificiales se extienden, su explicabilidad es más urgente para la sociedad, que demanda entender cómo y en base a qué la inteligencia artificial toma decisiones que le afectan.

Referencias

LeCun, Y., Bengio, Y., 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., y col. (2019). Machine behaviour. *Nature*, 568, 477-486.

Taylor, J., y 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., y 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.

Manuscrito recibido el 15 de noviembre de 2021.
Aceptado el 29 de marzo de 2022.

Esta es la versión en español de

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