El Sur

Fecha: 20/07/2025 Audiencia: 30.000 Sección: PUBLICIDAD
Vpe: $1.506.869 Tirada: 10.000 Frecuencia: 0
Vpe pág: $3.766.230 Difusión: 10.000
Vpe portada: $3.766.230 Ocupación: 40,01%

Ciencia

Cómo los científicos intentan utilizar la IA para desentrañar la mente humana

Diferentes experimentos trabajan con el uso de redes neuronales para predecir cómo se comportan los humanos y otros animales en experimentos psicológicos, utilizando esta tecnología.

El panorama actual de la Inteligencia Artificial (IA) se define por las diferencias que existen entre las redes neuronales y los cerebros humanos. Un niño aprende a comunicarse eficazmente con solo mil calorías al día y conversación constante. Mientras tanto, las empresas tecnológicas reabren centrales nucleares, contaminan comunidades marginadas y piratean terabytes de libros para entrenar y ejecutar sus grandes modelos de lenguaje (LLM).

Pero las redes neuronales son, al fin y al cabo, neuronales: están inspiradas en cerebros. A pesar de su apetito por la energía y los datos, los grandes modelos lingüísticos y los cerebros humanos tienen mucho en común. Ambos están formados por millones de subcomponentes: neuronas biológicas en el caso del cerebro, «neuronas» simuladas en el caso de las redes. Son las únicas dos cosas en la Tierra que pueden producir lenguaje con fluidez y flexibilidad. Y los científicos apenas entienden cómo funciona ninguna de las dos.

Los neurocientíficos suelen pensar que crear redes neuronales parecidas a las del cerebro es uno de los caminos más prometedores en este campo, y esa actitud ha empezado a extenderse a la psicología. La prestigiosa revista Nature publicó un par de estudios que muestran el uso de redes neuronales para predecir cómo se comportan los humanos y otros animales en experimentos psicológicos. Ambos estudios proponen que estas redes entrenadas podrían ayudar a los científicos a avanzar en su comprensión de la mente humana. Pero predecir un comportamiento y explicar cómo se produce son dos cosas muy distintas.

En uno de los estudios, los investigadores transformaron un gran modelo lingüístico en lo que denominan un «modelo básico de la cognición humana». De entrada, los grandes modelos lingüísticos no imitan muy bien el comportamiento humano: se comportan de forma lógica en entornos en los que los humanos abandonan la razón, como los casinos. Así que los investigadores perfeccionaron Llama 3.1, uno de los LLM de código abierto de Meta, con datos de 160 experimentos psicológicos que incluían tareas como elegir entre un conjunto de «máquinas tragamonedas» para obtener el máximo premio o recordar secuencias de letras. Al modelo resultante lo llamaron Centaur.

En comparación con los modelos psicológicos convencionales, que utilizan simples ecuaciones matemáticas, Centaur hizo un trabajo mucho mejor a la hora de predecir el comportamiento. Las predicciones precisas de cómo responden los humanos en experimentos psicológicos son valiosas en sí mismas: por ejemplo, los científicos podrían utilizar Centaur para pilotar sus experimentos en un ordenador antes de contratar y pagar a participantes humanos. En su artículo, sin embargo, los investigadores proponen que Centaur podría ser algo más que una máquina de predicción. Al examinar los mecanismos que permiten a Centaur reproducir con eficacia el comportamiento humano, los científicos podrían desarrollar nuevas teorías sobre el funcionamiento interno de la mente.

Pero algunos psicólogos dudan de que Centaur pueda decirnos mucho sobre la mente. Por supuesto, es mejor que los modelos psicológicos convencionales a la hora de predecir el comportamiento humano, pero también tiene mil millones de parámetros más. Y que un modelo se comporte como un ser humano por fuera no significa que funcione como tal por dentro. Olivia Guest, profesora adjunta de Ciencia Cognitiva computacional en la Universidad Radboud de los Países Bajos, compara a Centaur con una calculadora, que puede predecir eficazmente la respuesta que dará un genio de las matemáticas cuando se le pide que sume dos números. "No sé qué se puede aprender sobre la suma humana estudiando una calculadora", afirma.

Incluso si Centaur capta algo importante sobre la psicología humana, los científicos pueden tener dificultades para extraer información de los millones de neuronas del modelo. Aunque los investigadores de IA se esfuerzan por averiguar cómo funcionan los grandes modelos lingüísticos, apenas han conseguido abrir la caja negra. Comprender un enorme modelo de la mente humana basado en una red neuronal puede no resultar mucho más fácil que comprender la propia cosa.

Una alternativa es ir a lo pequeño. El segundo de los dos estudios de Nature se centra en redes neuronales minúsculas -algunas con una sola neurona- que, sin embargo, pueden predecir el comportamiento de ratones, ratas, monos e incluso seres humanos. Como las redes son tan pequeñas, es posible seguir la actividad de cada neurona y utilizar esos datos para averiguar cómo produce la red sus predicciones de comportamiento. Aunque no hay garantía de que estos modelos funcionen como los cerebros que se entrenaron para imitar, al menos pueden generar hipótesis comprobables sobre la cognición humana y animal.

La comprensibilidad tiene un coste. A diferencia de Centaur, que se entrenó para imitar el comportamiento humano en docenas de tareas diferentes, cada pequeña red solo puede predecir el comportamiento en una tarea específica. Una red, por ejemplo, está especializada en hacer predicciones sobre cómo la gente elige entre distintas máquinas tragamonedas. "Si el comportamiento es realmente complejo, se necesita una red grande", dice Marcelo Mattar, profesor adjunto de Psicología y Ciencias Neuronales en la Universidad de Nueva York, quien dirigió el estudio de las redes diminutas y también contribuyó a Centaur. "El compromiso es que ahora entenderla es muy, muy difícil".

Esta diatriba entre predicción y comprensión es una característica clave de la ciencia basada en redes neuronales. Estudios como el de Mattar están haciendo algunos progresos para cerrar esa brecha: por diminutas que sean sus redes, pueden predecir el comportamiento con más precisión que los modelos psicológicos tradicionales. Lo mismo ocurre con la investigación sobre la interpretabilidad de los LLM en sitios como Anthropic. Por ahora, sin embargo, nuestra comprensión de los sistemas complejos -desde los humanos hasta los sistemas climáticos o las proteínas- va cada vez más a la zaga de nuestra capacidad para hacer predicciones sobre ellos. (MIT Technology Review)