Transformers, la magia detrás del ChatGPT

Transformers, the magic behind ChatGPT

Por Asdrúbal López Chau

Resumen: El ChatGPT (Chat Transformador Generativo Preentrenado) es un modelo de la empresa OpenAI, con capacidades que incluyen traducir textos a varios idiomas, resumir documentos y generar textos originales. ¿Cómo logra realizar estas tareas aparentemente complejas? En este artículo, nos adentraremos en el funcionamiento de este modelo.

Palabras clave: ChatGPT, IA, modelo de lenguaje, red neuronal, datos

Abstract: ChatGPT (Pre-Trained Generative Transformational Chat) is a model from the company OpenAI, with capabilities that include translating texts into several languages, summarizing documents and generating original texts. How does it manage to perform these seemingly complex tasks? In this article we will take a closer look at how this model works.

Keywords: ChatGPT, AI, language model, neural network, data.

Los chats conversacionales, como ChatGPT, Bard y BingAI, se clasifican en una categoría de modelos denominada LLM (Large Language Models, Modelos de Lenguaje de Gran Tamaño). En términos generales, los LLM son una forma particular de la inteligencia artificial generativa. La figura 1 ilustra la relación de esta categoría con otros campos afines.

Los LLM se nutren de una vasta cantidad de documentos, con frecuencia abarcan una gran parte del contenido de internet. Una vez que estos modelos han procesado los documentos, cada elemento esencial de los textos, conocidos como *tokens* (por ejemplo, palabras), se convierte en un vector denso de números reales. Estos se calculan de manera que los que representan elementos similares o relacionados se encuentren más próximos entre sí que los vectores de elementos sin ninguna relación. En la arquitectura de los LLM, esta operación se realiza en capas denominadas *embedding*. En aprendizaje profundo, el término *arquitectura* se refiere a la estructura, organización y configuración de una red neuronal.

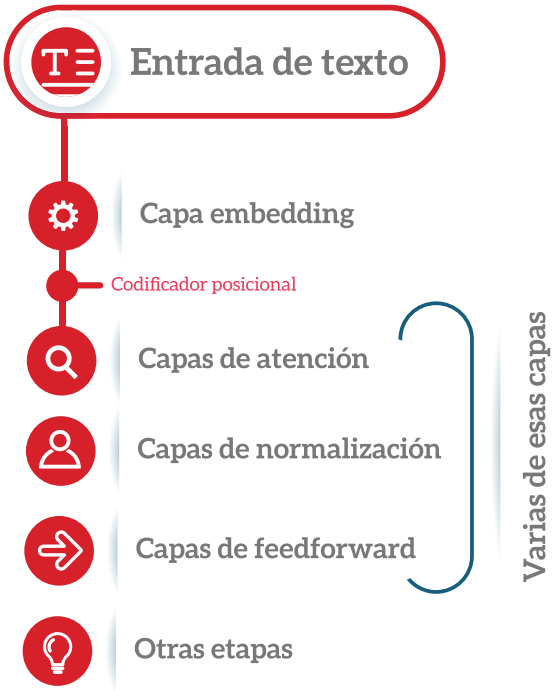


Figura 1. Mapa simplificado de la IA para chats conversacionales. Fuente: Elaboración propia.



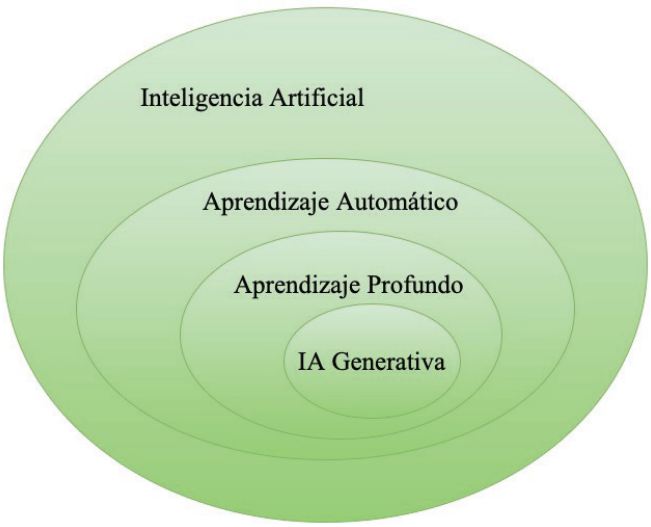
Con el propósito de capturar de manera más precisa las relaciones entre las palabras, por ejemplo en un párrafo, se llevan a cabo modificaciones en los vectores que representan cada palabra mediante operaciones matemáticas sencillas. Es decir, cada vector ejerce una influencia sobre los demás presentes en el mismo párrafo. Este mecanismo, que fue introducido en 2017, se conoce como *Atención* (Vaswani *et al.*, 2017). Este último, habilita al modelo para calcular la relevancia relativa de cada palabra dentro de una secuencia en relación con todas las demás palabras que conforman un documento. El efecto resulta en mayor relevancia a algunas palabras en función de las otras.

Tras pasar la capa de *Atención*, los vectores se utilizan como entradas en una red neuronal completamente conectada, diseñada para emular el funcionamiento de las neuronas. A esta capa se le denomina red neuronal *FeedForward*.

El vector resultante de esta red neuronal, una vez normalizado, puede interpretarse como una representación modificada de los tokens. El orden o posición relativa de cada uno de estos vectores normalizados se tiene en cuenta para otorgar mayor coherencia a las secuencias de palabras, es decir, para permitir entender la estructura de la información en el texto. Esta última operación es realizada por capas llamadas *Codificadores posicionales*.

En el contexto de los LLM, el término *transformers* hace referencia a múltiples capas que incluyen mecanismos de atención, redes neuronales feedforward, normalización y codificadores posicionales (ver Figura 2). Estas capas están dispuestas una después de otra, de modo que la salida de una capa se convierte en la entrada de la siguiente. Los transformers tienen algunos componentes más que no se abordan en este artículo de divulgación (revisar Wang, 2023). Actualmente, permiten a los LLM estar en el pináculo de la creación automática de contenido textual. Para hacer que ChatGPT tuviera las capacidades actuales, el modelo fue afinado mediante la ayuda de expertos, quienes relacionaron manualmente preguntas o instrucciones (llamados *promts*) con respuestas correctas.

Figura 2. Estructura simplificada de un transformer



Fuente: Elaboración propia basada en Binxu Wang (2023).

Antes de la irrupción de los transformers, los métodos más comunes para la generación automática de texto se apoyaban en redes neuronales con memoria, como las redes neuronales recurrentes (RNN) y las redes neuronales con memoria a largo plazo (LSTM). No obstante, los transformers sobrepasan con soltura a estos modelos y a numerosos otros enfoques.

Actualmente, nos enfrentamos a dos desafíos fundamentales: obtener buenos resultados con una cantidad mínima de texto, y permitir que los modelos aprendan y olviden de manera continua, emulando la capacidad de los seres humanos. Como mencioné, los modelos de lenguaje de gran escala (LLM) requieren una vasta cantidad de datos, en el rango de los terabytes. Además, debe considerarse que el proceso de entrenamiento de los LLM puede prolongarse durante semanas e incluso meses.

Predecir cuál será la próxima arquitectura disruptiva en el campo de la inteligencia artificial es un desafío arduo. No obstante, es intrigante observar que los cimientos de los modelos más revolucionarios en este ámbito residen en artículos científicos que fueron publicados hace varios años, y aguardan a que una mente creativa conecte aparentemente ideas disímiles para desplegar su potencia artificial en problemas nacionales.

Referencias

Vaswani, Ashish *et al.* (2017). "Attention Is All You Need", *en NIPS'17:*, *Proceedings of the 31st International Conference on Neural Information Processing* Systems. Red Hook, NY, USA: Curran Associates Inc.

Wang, Binxu (2023). "From Transformer to LLM: Architecture, Training and Usage", en *Retrieved.* <https://scholar.harvard.edu/binxuw/classes/machine-learning-scratch/materials/transformers>.



Asdrúbal López-Chau es profesor de tiempo completo en el Centro Universitario Zumpango, de la Universidad Autónoma del Estado de México. Es integrante del Sistema Nacional de Investigadores Nivel 1 y miembro del Research Laboratory in Engineering and Applied Sciences, del CU UAEM Zumpango. Sus líneas de investigación actuales son el diseño y aplicación de modelos de inteligencia artificial en problemas nacionales.