MAPAS DEL LABERINTO

UNA SERIE DE PHANTOM MAZE — AI & LANGUAGE LAB

ENSAYO 18

Transformer: un puñado de páginas que cambiaron la forma de todo

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575288

La arquitectura que hizo posible la IA generativa



Wenzel Hablik, Starry Sky, Attempt (1909)

En 2017, ocho empleados de Google publicaron un artículo de unas pocas páginas. Le pusieron un título de canción: “Attention Is All You Need” guiña a “All You Need Is Love” de los Beatles. El documento interno de diseño se llamaba “Transformers: autoatención iterativa y procesamiento para varias tareas”, e incluía una ilustración con seis muñecos de la franquicia de juguetes. Como es fácil imaginar, al equipo le decían el Equipo Transformer. El nombre de la arquitectura quedó, según

cuenta la historia, porque a Jakob Uszkoreit, uno de los ocho, le evocaba recuerdos infantiles y se adaptaba bien a la idea de transformar información. En el paper aclararon que el orden de los autores era aleatorio. Hoy ese artículo casual y juguetón acumula más de doscientas cincuenta mil citas, figura entre los diez trabajos más citados del siglo XXI en cualquier disciplina y la inicial quedó en boca de todos a través de la T de GPT. Los ocho, dicho sea de paso, se fueron de Google, casi todos a fundar sus propias empresas. Vamos a explorar la historia de esa arquitectura y de la idea que la hizo tan exitosa.

Incluso en el mundo de Megatrón y Optimus Prime, "transformer" carga la herencia del latín y su *transformare* (*trans*, "a través, más allá, de un lado a otro", y *formare*, "dar forma", de *forma*). Transformar es sacar algo de una figura y ponerlo en otra. Ovidio lo describió hace dos mil años en los dos versos iniciales de *Metamorfosis*: *In nova fert animus mutatas dicere formas / corpora*, "mi ánimo me lleva a cantar las formas mudadas en cuerpos nuevos". Luego vienen doscientas cincuenta historias de cuerpos que cambian de forma, como Dafne en laurel, Aracne en araña y Narciso en flor, hiladas en la tesis de que nada perece y todo se transforma, es decir, que la forma viaja de cuerpo en cuerpo.

Curiosamente el nombre de la arquitectura resultó exacto incluso cuando sus autores no lo buscaban, porque un transformer toma una secuencia de vectores, las representaciones numéricas de los tokens, esos puntos en el espacio del sentido por los que cada palabra queda convertida en una posición, y la transforma, capa tras capa, en otra secuencia de representaciones. En sus descendientes generativos, esas transformaciones terminan sirviendo para calcular qué token puede venir a continuación. Como una máquina ovidiana, muda formas.

Para entender qué cambió en 2017 hay que mirar qué se hacía antes. Las redes recurrentes procesaban una secuencia paso a paso. Cada posición producía un estado interno que dependía del anterior. Las LSTM, desarrolladas para conservar mejor información a largas distancias, mejoraron mucho esa memoria. Antes del transformer ya había aparecido la atención, que permitía al decoder consultar directamente distintas partes de la oración de entrada. El traductor neuronal de Google de 2016 combinaba precisamente esas dos soluciones, i.e., profundas redes LSTM y un mecanismo de atención. Quedaba, sin embargo, un cuello de botella difícil de eliminar. La recurrencia obligaba a hacer los cálculos en orden. Para llegar a la posición cien había que haber calculado la noventa y nueve. Eso era especialmente incómodo para las GPU, diseñadas para realizar enormes cantidades de operaciones simultáneas y aquí paradas esperando en la cola. El transformer hizo la apuesta radical de conservar la atención y eliminar la recurrencia.

La idea insolente del transformer fue, en efecto, eliminar la lectura secuencial por completo. Nada de resúmenes viajando de palabra en palabra. Todas las palabras de la oración se miran entre sí a la vez, directamente, cada una decidiendo cuáles otras le importan para afinar su propio sentido. En el decoder generativo hay una restricción. Cada posición puede consultar lo anterior, pero no mirar lo que todavía no ocurrió. Eso no impide paralelizar el entrenamiento, pero obliga a que, cuando el modelo genera, la salida aparezca token por token. Este mecanismo se llama atención. Digamos aquí solamente que en lugar de un lector que avanza con el dedo ahora hay una mesa redonda donde todos

los términos de la frase se consultan simultáneamente unos a otros. Jakob Uszkoreit venía sospechando que con eso solo bastaba y la recurrencia era una muleta desechable. De ahí el título "Attention is all you need". La conjetura era tan contraria al sentido común del campo que hasta su propio padre, Hans Uszkoreit, un conocido lingüista computacional, era escéptico. Tal vez Jakob también necesitaba atención.

El transformer convirtió la lectura, que era una fila, en una asamblea lista para aprovechar el ejército de calculadoras paralelas de la GPU. El modelo original se entrenó en ocho placas durante tres días y medio, una miseria para los estándares que vendrían. Bastó para hacer época. Por supuesto, no todo es rosa. Si cada palabra consulta a todas, el costo crece con el cuadrado del largo del texto, de modo que duplicar el contexto cuadruplica la cuenta, pero vamos a concentrarnos aquí en el funcionamiento y no en los costos.

El truco de partir un trabajo secuencial para hacerlo simultáneo tiene un antecedente medieval llamativo. En el siglo XIII, alrededor de las universidades de Bolonia y París, la demanda de libros explotó. Miles de estudiantes necesitaban los mismos textos de derecho, teología y filosofía, pero un copista tardaba meses en reproducir un solo volumen, porque copiar era un trabajo de a una palabra por vez, igual que leer para una red recurrente. La solución fue el sistema de *pecia*, "pieza" en latín. El librero mantenía un ejemplar oficial del texto, revisado y aprobado por la universidad, pero partido en cuadernillos numerados de pocas hojas. Cada copista tomaba una pieza por semana y así veinte copistas podían estar trabajando en el mismo libro a la vez. De ese modo el códice completo salía en semanas al pasar de la fila a la atención simultánea. Se parece bastante a lo que el transformer le hizo a la lectura de la máquina. La red recurrente era un copista solitario encadenado al orden del texto, mientras el transformer es el taller de la pecia con el texto repartido en piezas que se procesan todas a la vez, con sistemas de numeración en los cuadernillos y marcas de posición para no perder el orden, porque al trabajar en paralelo ya no se sabe qué venía antes de qué. En ambos casos había gran demanda y un método artesanal que no escalaba.

Sin embargo, hay que ir más allá de los copistas y las redes para ilustrar el paso más interesante. En "El Aleph" de Jorge Luis Borges, el narrador baja al sótano de una casa de la calle Garay y encuentra una esfera pequeña, de dos o tres centímetros de diámetro, donde están todos los puntos del universo a la vez y sin superponerse. El problema se vuelve pronto de escritura. Vio en simultáneo lo que tiene que escribir de modo sucesivo, porque el lenguaje tiene esa estructura. El transformer vive en parte esa misma desproporción, ya que mira la frase entera de una vez, con todas las palabras consultándose al mismo tiempo y después tiene que entregar el resultado como lo entrega el lenguaje, una palabra detrás de otra. La asimetría es constitutiva y el privilegio de mirar todo junto se vuelve un dolor de cabeza a la hora de contarlo.

Wenzel Hablik, Starry Sky, Attempt (1909)

Lo que siguió fue una colonización completa. El transformer de 2017 había nacido para traducir y por eso tenía dos partes: un encoder, que recibía la oración de origen y construía una representación de ella, y un decoder, que utilizaba esa representación para producir la traducción paso a paso. Su

descendencia pronto separó las dos mitades y descubrió que podían vivir muy bien por su cuenta. BERT, presentado por Google en 2018, conservó básicamente el lado encoder. Lee el contexto en ambas direcciones y construye representaciones especialmente útiles para comprender, clasificar y buscar. Al año siguiente Google lo incorporó a su buscador para interpretar mejor ciertas consultas. GPT tomó el camino complementario al conservar el decoder y convertirlo en una máquina entrenada para continuar una secuencia, prediciendo un token a partir de los anteriores. La posibilidad generativa apareció con Łukasz Kaiser y Noam Shazeer, dos de los autores del transformer original, que participaron poco después en un sistema solo decoder capaz de producir incluso artículos completos de Wikipedia a partir de documentos de referencia. La línea de GPT convirtió esa idea en un programa de escala creciente. Su nombre todavía conserva la genealogía completa: *Generative Pre-trained Transformer*, con la arquitectura casi como apellido. Pero sería un error pensar que la herencia quedó encerrada en el lenguaje. Pronto hubo transformers que trataron una imagen como una secuencia de parches, otros que procesaron voz y música, y después video, robótica y sistemas multimodales. Una arquitectura concebida para una tarea bastante concreta, traducir de una lengua a otra, terminó convertida en un mecanismo general para procesar casi cualquier cosa que pueda representarse como una secuencia de partes relacionadas.

Ganó porque se dejaba entrenar en paralelo y permitía relacionar posiciones lejanas en una secuencia. En la era del cómputo masivo, la arquitectura que mejor se reparte entre mil calculadoras se lleva todo. Es una lección incómoda y repetida. Una parte importante de la IA se explica más por costo de escalado que por una abstracta noción de inteligencia. La máquina con nombre de juguete, la que muda formas capa tras capa como el poema de Ovidio, terminó protagonizando la metamorfosis del campo entero. Hasta ahora ella misma casi no cambió. Los modelos actuales han cambiado muchas piezas alrededor de aquella arquitectura, pero la combinación fundamental de autoatención, capas feed-forward y conexiones residuales sigue siendo reconocible. Hay pretendientes al trono que prometen esquivar su famoso peaje cuadrático. Mientras tanto, conviene recordar que ocho páginas de 2017 decidieron la forma de casi todo lo que la máquina dice y que las formas, como enseñaba Ovidio, solo están de paso por los cuerpos.

◊

CÓMO CITAR

Marsico, Claudia y Hernán Inverso. 2026. “Transformer: un puñado de páginas que cambiaron la forma de todo”. *Mapas del laberinto*, ensayo 18. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575288

Mapas del laberinto es una serie de Phantom Maze, AI & Language Lab. La colección vive en **phantommaze.cloud**; los ensayos aparecen también en phantommazeailabes.substack.com. La suscripción es gratuita.

Algunas de las ideas que enmarcan este texto se desarrollan en *El fantasma sin ego. Un mapa en el laberinto de la inteligencia artificial*, de Hernán Inverso y Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-10-2).

Fuentes. El artículo fundacional: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser e Illia Polosukhin, "Attention Is All You Need" (NeurIPS, 2017), con la nota sobre el orden aleatorio de los autores y los detalles de entrenamiento (8 GPU P100, 3,5 días para el modelo grande). Los datos de color sobre el título beatle, el nombre elegido por Uszkoreit "porque le gustaba cómo sonaba", el documento de diseño "Transformers: Iterative Self-Attention and Processing for Various Tasks" con los seis muñecos de la franquicia, y el escepticismo de Hans Uszkoreit, en Steven Levy, "8 Google Employees Invented Modern AI. Here's the Inside Story" (*Wired*, 2024), y en Madhumita Murgia, "Transformers: The Google Scientists Who Pioneered an AI Revolution" (*Financial Times*, 2023), que documenta también la partida de los ocho autores de Google. El lugar del paper entre los más citados del siglo, en Helen Pearson y col., "Exclusive: The Most-Cited Papers of the Twenty-First Century" (*Nature*, 2025). Ovidio, *Metamorfosis*, I, 1-2, y la doctrina de la transformación universal en el discurso de Pitágoras del libro XV. Jorge Luis Borges, "El Aleph", publicado en *Sur* en 1945 y recogido en *El Aleph* (1949): la esfera tornasolada de dos o tres centímetros en el sótano de la casa de la calle Garay, y la desesperación del narrador ante la simultaneidad de lo visto frente a la sucesión que le impone el lenguaje. El traductor neuronal de Google de 2016: Yonghui Wu y col., "Google's Neural Machine Translation System" (2016). Sobre la memoria larga de corto plazo: Sepp Hochreiter y Jürgen Schmidhuber, "Long Short-Term Memory" (*Neural Computation*, 1997). Sobre el sistema de pecia: Jean Destrez, *La pecia dans les manuscrits universitaires du XIIIe et du XIVe siècle* (1935), el estudio clásico que identificó ejemplares y piezas de Bolonia, París, Oxford y Nápoles; síntesis en la entrada "Pecia system" del *Oxford Companion to the Book* (Oxford University Press, 2010). Sobre la descendencia: Jacob Devlin y col., "BERT: Pre-training of Deep Bidirectional Transformers" (2018); Alec Radford y col., "Improving Language Understanding by Generative Pre-Training" (2018), primer GPT; Alexey Dosovitskiy y col., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2021), el transformer de visión.