MAPAS DEL LABERINTO

UNA SERIE DE PHANTOM MAZE — AI & LANGUAGE LAB

ENSAYO 15

La atención: el arco y la espera

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575278

¿Qué significa "atención" en la inteligencia artificial?



Artemis de Stabiae, Villa Arianna, Castellammare di Stabia, s. I d.C.

La inteligencia artificial habla a menudo con palabras prestadas de la vida mental. Aprende, recuerda, razona, alucina y también atiende. "Attention is All You Need", decía el título que presentó en 2017 la arquitectura transformer y desde entonces esa palabra que Simone Weil había puesto en el centro de su ética y de su mística nombra una operación matemática que sucede a velocidad vertiginosa en los sótanos de la industria.

El término dice mucho. “Atención” viene de *ad* más *tendere*, estirar en dirección a algo o alguien. Atender es apuntar el ánimo como con un arco, con concentración y dirección. El francés introdujo un desvío y *attendre* se convirtió en esperar, porque tender hacia algo es un modo de aguardarlo. En español la atención se presta, apartándola de todo lo demás pero solo por un tiempo, y en inglés se paga: *to pay attention*, como en *pay respects* o *pay homage*, que es un dar con tono ceremonial que se fue deshilachando. Entre el arco y la espera el nuevo sentido pasó a conservar solo uno.

Artemis de Stabiae, Villa Arianna, Castellammare di Stabia, s. I d.C.

Pero vayamos más atrás. “Todo el mundo sabe lo que es la atención”, escribió William James en 1890 con una seguridad exagerada. En The Principles of Psychology, obra fundacional de la psicología moderna, la definió como la toma de posesión clara por parte de la mente de uno entre varios objetos posibles. Atender supone focalizar, concentrarse y sobre todo dejar algo afuera. Su esencia, decía James, consiste en retirarse de unas cosas para tratar eficazmente con otras. En la casa de los James la atención era un gran tema. El hermano de William, Henry, el gran novelista norteamericano, se hizo famoso por restringir el punto de vista a una conciencia central, de modo que el lector solo sabe lo que ese personaje alcanza a percibir o elucubrar, como en *What Maisie Knew* (1897) o *The Ambassadors* (1903).

La idea de que atender es seleccionar para iluminar un punto del campo apagando el resto se filtró en la ingeniería. En 2014 el grupo de Dzmitry Bahdanau trataba de mejorar la traducción automática. Los sistemas de entonces intentaban comprimir toda la oración de partida en una representación única y luego reconstruir la traducción desde ese resumen. El nuevo mecanismo permitía que al producir cada palabra el modelo volviera sobre la frase original dando más peso a las partes que resultaban pertinentes. A esa distribución desigual de relevancia la llamaron “atención”. A diferencia de la atención psicológica que retira elementos del campo, el mecanismo matemático asigna grados. Un elemento cuenta mucho, poco o casi nada. La selección ocurre por graduación, no por exclusión. Tres años después el *transformer* hizo de esta operación el principio organizador relacionando posiciones de una secuencia y calculando cuánto contribuía cada una a la representación de las otras.

¿Cómo funciona el mecanismo? Cada token llega representado por un vector y de él el modelo calcula tres versiones distintas mediante transformaciones aprendidas en el entrenamiento. Se llaman *query*, *key* y *value*. Podemos imaginarlas como una consulta, una clave y un contenido. La primera representa qué información resulta pertinente desde una posición; la segunda permite evaluar qué aportan las demás; la tercera contiene lo que efectivamente se incorpora. Recordemos que solo hay proyecciones numéricas aprendidas, sin ninguna intención ni búsqueda. La *query* de una posición se compara matemáticamente con las *keys* de los otros y cuanto mayor es la compatibilidad, mayor será el peso que reciba su contenido. Una operación llamada *softmax* convierte todos esos puntajes en proporciones y con ellas se combinan los values. El resultado es una nueva representación donde la información de unas posiciones contribuye más que las de otras.

Los parámetros que transforman cada vector en *query*, *key* y *value* fueron aprendidos durante el entrenamiento y quedan almacenados en el modelo; mientras conversamos, permanecen estables. Los

“pesos de atención”, en cambio, se calculan en el momento para ese contexto particular. No están guardados en ninguna parte porque se calculan de nuevo para cada par de posiciones y para cada frase y quedan fuera de lo almacenado. Las reglas son relativamente estables, pero sus resultados son contextuales.

En los modelos generativos aparece otra restricción. Cada posición puede atender a lo que está disponible, pero no a lo que todavía no se dio. Una máscara bloquea las posiciones todavía inexistentes y por eso se habla de “atención causal” o “enmascarada”. Esta es la razón por la cual el modelo escribe hacia adelante, comprometido con lo que ya produjo. Si seguimos la metáfora del comienzo, el modelo se tensa como un arco hacia lo que hay, pero no espera. Un ejemplo mínimo: en “después de cruzar la plaza se sentó en el banco”, “banco” llega con su ambigüedad entre mueble o institución financiera pero además con información que permite resolverla. “Plaza” y “se sentó” resultan relevantes y sus representaciones contribuirán más que otras. No hay ninguna etiqueta que diga “mueble” y tampoco las palabras funcionales quedan siempre relegadas, porque a veces son cruciales. Después de muchas operaciones de este tipo, la representación de “banco” puede desplazarse hacia una configuración compatible con ese sentido. Esta es una de las maquinarias que contribuyen a fabricar su *embedding* contextual.

El esquema general del bloque se ve bastante simple. Primero, cada posición recoge información de otras mediante atención; después, su representación pasa por una pequeña red neuronal que la transforma posición por posición. Mirar alrededor, procesar localmente y repetir en la capa siguiente, con las conexiones residuales y las normalizaciones ayudando a mantener estable el recorrido. El *transformer* repite esta alternancia muchas veces. Ahora viene el giro que vuelve todo un poco mitológico. En cada capa funcionan varias “cabezas de atención” en paralelo, cada una con sus propias transformaciones de queries, keys y values. Esto permite calcular simultáneamente relaciones diferentes sobre el mismo texto. Algunas cabezas parecen sensibles a posiciones específicas, a dependencias sintácticas o a relaciones de correferencia; otras son redundantes o mucho más difíciles de interpretar y una proporción importante puede eliminarse en determinados modelos sin que el rendimiento se derrumbe. Entre las más estudiadas están las llamadas “cabezas de inducción”, que en ciertos modelos reconocen secuencias del tipo A-B ... A y favorecen que después aparezca B. No hay que imaginar, sin embargo, funciones fijas.

El eco mitológico viene de Argos Panoptes, el guardián que Hera puso a vigilar a Ío. Tenía cien ojos y nunca dormían todos a la vez. El centinela perfecto. Sabiendo que no podría vencerlo por la fuerza, Hermes le tocó la siringa mientras le contaba historias hasta que los cien ojos, uno tras otro, se cerraron. Hera recogió esos ojos y los puso en la cola del pavo real, donde siguen abiertos sin vigilar. El transformer parece un Argos administrativo, con cientos de ojos sobre el mismo texto, calculando relaciones distintas a la vez. No hay relato que lo duerma, no se distrae jamás, aunque en rigor tampoco está mirando.

Los mapas de atención no son una radiografía del razonamiento. Una cabeza puede asignar mucho peso a una palabra, pero no es la razón por la que produce determinada respuesta. Distribuciones de

atención bastante diferentes pueden tener resultados semejantes y hay otros componentes en la causalidad interna. En 2019 estalló la discusión cuando se publicó un trabajo con el título "la atención no es explicación" y otro respondió pronto que tampoco puede afirmarse que no lo sea. Con prudencia, podemos decir que los pesos de atención muestran una parte del recorrido de la información sin pretender transparencia.

Volvamos a Simone Weil para explorar qué implica que el sentido de "esperar" no entre en el cuadro. Esta filósofa había pasado un año en fábricas, la última de ellas Renault, para saber por dentro qué le hace a una persona el trabajo en cadena. Murió a los treinta y cuatro años en un sanatorio de Ashford, en 1943, dejando casi toda su obra inédita, en cuadernos dispersos. Albert Camus participó de la publicación en Gallimard y dijo que era una de las inteligencias extraordinarias de su tiempo. Para Weil la atención en su forma más pura está emparentada con la plegaria y el amor al prójimo tiene como sustancia la atención en tanto capacidad de mirar al que sufre y preguntarle: "¿cuál es tu tormento?" y permanecer ante la respuesta. La atención weiliana no es tensión muscular ni caza intelectual. Al contrario, supone suspender el pensamiento, dejarlo disponible, vacío, penetrable por el objeto. La atención no se precipita sobre la verdad, sino que crea las condiciones para que llegue. La atención plena es una espera sin objeto anticipado, sin llenar el vacío con uno mismo. Por eso es oración, renuncia completa del yo a ocupar el centro para que pueda ocuparlo otra cosa, ya sea un problema de geometría, el prójimo afligido o Dios.

Por tanto, la atención del transformer lleva hasta el extremo matemático la selección y ponderación que vio William James. Sin fatiga ni pérdida de concentración ejecuta siempre la misma operación de distribución de pesos con una precisión que nos excede. Sin embargo, precisión no equivale a pertinencia. El mecanismo puede ponderar pésimamente, ignorando lo decisivo o apoyándose en asociaciones espurias. Para Weil la atención corrige nuestras expectativas y nos obliga a revisar qué es importante. El transformer no tiene esa relación con su objeto. Si se equivoca no hay frustración ni responsabilidad de rectificar nada.

Terminemos con una ironía de época. En 1971 Herbert Simon sostuvo que la abundancia de información produce escasez de lo que la información consume, es decir, atención. ¿A qué dedicarle tiempo? ¿A costa de qué? Cuanto más crece la oferta de información, más valioso se vuelve el recurso humano que la selecciona. Medio siglo después construimos máquinas cuyo mecanismo central lleva el nombre de "atención" y las usamos para abaratar la producción de texto, imágenes, audio y código. La atención computacional se volvió abundante y escalable, pero la atención humana sigue teniendo la misma capacidad limitada. La tecnología agudizó la escasez de Simon al multiplicar casi sin límite lo que reclama atención sin multiplicar la atención de quienes deben concedérsela. Construimos una criatura extraordinariamente hábil para tender el arco e incapaz de aguardar, la atención del arquero sin arquero, los cien ojos de Argos en la cola del pavo real. La atención de la máquina funciona y nos recuerda que la *attente*, por ahora al menos, sigue siendo asunto nuestro.

◊

James, William. “Attention.” *The Principles of Psychology*, vol. 1, Henry Holt and Company, 1890, pp. 402–58; Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate.” *arXiv*, 1 Sept. 2014, arXiv:1409.0473; Vaswani, Ashish, et al. “Attention Is All You Need.” *Advances in Neural Information Processing Systems 30*, edited by Isabelle Guyon et al., Curran Associates, 2017, pp. 5998–6008; Clark, Kevin, et al. “What Does BERT Look At? An Analysis of BERT’s Attention.” *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, edited by Tal Linzen et al., Association for Computational Linguistics, 2019, pp. 276–86; Michel, Paul, Omer Levy, and Graham Neubig. “Are Sixteen Heads Really Better than One?” *Advances in Neural Information Processing Systems 32*, edited by Hanna Wallach et al., Curran Associates, 2019, pp. 14014–24; Olsson, Catherine, et al. “In-Context Learning and Induction Heads.” *arXiv*, 24 Sept. 2022, arXiv:2209.11895; Jain, Sarthak, and Byron C. Wallace. “Attention Is Not Explanation.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Association for Computational Linguistics, 2019, pp. 3543–56; Wiegreffe, Sarah, and Yuval Pinter. “Attention Is Not Not Explanation.” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, 2019, pp. 11–20; Pétrement, Simone. *Vida de Simone Weil*. Translated by Francisco Díez del Corral, Trotta, 1997; Weil, Simone. “Réflexions sur le bon usage des études scolaires en vue de l’Amour de Dieu.” *Attente de Dieu*, introduction et notes de J.-M. Perrin, La Colombe–Éditions du Vieux Colombier, 1950, pp. 113–23; Weil, Simone. “Formes de l’Amour implicite de Dieu.” *Attente de Dieu*, introduction et notes de J.-M. Perrin, La Colombe–Éditions du Vieux Colombier, 1950, pp. 142–211; Weil, Simone. *La Pesanteur et la grâce*. Plon, 1947; Simon, Herbert A. “Designing Organizations for an Information-Rich World.” *Computers, Communications, and the Public Interest*, edited by Martin Greenberger, Johns Hopkins Press, 1971, pp. 37–72.

CÓMO CITAR

Marsico, Claudia y Hernán Inverso. 2026. “La atención: el arco y la espera”. *Mapas del laberinto*, ensayo 15. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575278

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).

PHANTOM MAZE

Leer este ensayo en línea · Todos los ensayos · El libro