MAPPE DEL LABIRINTO

UNA SERIE DI PHANTOM MAZE — AI & LANGUAGE LAB

SAGGIO 15

L’attenzione: l’arco e l’attesa

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575278

Che cosa significa “attenzione” nell’intelligenza artificiale?



Artemide di Stabia, Villa Arianna, Castellammare di Stabia, I sec. d.C.

L’intelligenza artificiale parla spesso con parole prese in prestito dalla vita mentale. Impara, ricorda, ragiona, allucina e anche presta attenzione. “Attention Is All You Need”, recitava il titolo che nel 2017 presentò l’architettura transformer, e da allora quella parola che Simone Weil aveva posto al centro della sua etica e della sua mistica designa un’operazione matematica che avviene a velocità vertiginosa nei sotterranei dell’industria.

Il termine dice molto. “Attenzione” viene da *ad* più *tendere*, protendersi verso qualcosa o qualcuno. Prestare attenzione è puntare l’animo come si punta un arco, con concentrazione e direzione. Il francese introdusse una deviazione e *attendre* finì per significare aspettare, perché tendere verso qualcosa è un modo di attenderlo. In italiano, come in spagnolo, l’attenzione si presta, sottraendola a tutto il resto ma solo per un tempo, e in inglese si paga: *to pay attention*, come in *pay respects* o *pay homage*, un dare dal tono cerimoniale che è andato sfilacciandosi. Fra l’arco e l’attesa, il nuovo senso finì per conservarne uno solo.

Artemide di Stabia, Villa Arianna, Castellammare di Stabia, I sec. d.C.

Ma facciamo un passo indietro. “Tutti sanno che cos’è l’attenzione”, scrisse William James nel 1890 con una sicurezza eccessiva. In *The Principles of Psychology*, opera fondativa della psicologia moderna, la definì come la presa di possesso chiara, da parte della mente, di uno fra diversi oggetti possibili. Prestare attenzione significa focalizzare, concentrarsi e soprattutto lasciare qualcosa fuori. La sua essenza, diceva James, consiste nel ritirarsi da alcune cose per trattare efficacemente con altre. In casa James l’attenzione era un grande tema. Il fratello di William, Henry, il grande romanziere americano, divenne celebre per aver ristretto il punto di vista a una coscienza centrale, così che il lettore sa soltanto ciò che quel personaggio riesce a percepire o a immaginare, come in *What Maisie Knew* (1897) o *The Ambassadors* (1903).

L’idea che prestare attenzione sia selezionare, illuminare un punto del campo spegnendo il resto, filtrò nell’ingegneria. Nel 2014 il gruppo di Dzmitry Bahdanau cercava di migliorare la traduzione automatica. I sistemi di allora tentavano di comprimere l’intera frase di partenza in una rappresentazione unica e poi di ricostruire la traduzione a partire da quel riassunto. Il nuovo meccanismo permetteva che, nel produrre ogni parola, il modello tornasse sulla frase originale dando più peso alle parti che risultavano pertinenti. Chiamarono “attenzione” quella distribuzione disuguale di rilevanza. A differenza dell’attenzione psicologica, che ritira elementi dal campo, il meccanismo matematico assegna gradi. Un elemento conta molto, poco o quasi nulla. La selezione avviene per gradazione, non per esclusione. Tre anni dopo il *transformer* fece di questa operazione il principio organizzatore, mettendo in relazione le posizioni di una sequenza e calcolando quanto ciascuna contribuisse alla rappresentazione delle altre.

Come funziona il meccanismo? Ogni token arriva rappresentato da un vettore e da questo il modello calcola tre versioni distinte mediante trasformazioni apprese durante l’addestramento. Si chiamano *query*, *key* e *value*. Possiamo immaginarle come un’interrogazione, una chiave e un contenuto. La prima rappresenta quale informazione risulti pertinente da una posizione; la seconda permette di valutare che cosa apportino le altre; la terza contiene ciò che effettivamente viene incorporato. Ricordiamo che ci sono soltanto proiezioni numeriche apprese, senza alcuna intenzione né ricerca. La *query* di una posizione viene confrontata matematicamente con le *key* delle altre e quanto maggiore è la compatibilità, tanto maggiore sarà il peso che riceverà il loro contenuto. Un’operazione chiamata *softmax* converte tutti quei punteggi in proporzioni e con queste si combinano i *value*. Il risultato è

una nuova rappresentazione in cui l'informazione di alcune posizioni contribuisce più di quella di altre.

I parametri che trasformano ogni vettore in *query*, *key* e *value* sono stati appresi durante l'addestramento e restano immagazzinati nel modello; mentre conversiamo, rimangono stabili. I "pesi di attenzione", invece, si calcolano al momento, per quel contesto particolare. Non sono conservati da nessuna parte, perché vengono ricalcolati per ogni coppia di posizioni e per ogni frase e restano fuori da ciò che è immagazzinato. Le regole sono relativamente stabili, ma i loro risultati sono contestuali.

Nei modelli generativi compare un'altra restrizione. Ogni posizione può prestare attenzione a ciò che è disponibile, ma non a ciò che ancora non si è dato. Una maschera blocca le posizioni ancora inesistenti e per questo si parla di "attenzione causale" o "mascherata". È questa la ragione per cui il modello scrive in avanti, vincolato a ciò che ha già prodotto. Se seguiamo la metafora dell'inizio, il modello si tende come un arco verso ciò che c'è, ma non aspetta. Un esempio minimo: in "dopo aver salito le scale si fermò al secondo piano", "piano" arriva con la sua ambiguità fra il livello di un edificio e lo strumento musicale, ma anche con l'informazione che permette di risolverla. "Scale" e "salito" risultano rilevanti e le loro rappresentazioni contribuiranno più di altre. Non c'è nessuna etichetta che dica "edificio" e nemmeno le parole funzionali restano sempre relegate, perché a volte sono cruciali. Dopo molte operazioni di questo tipo, la rappresentazione di "piano" può spostarsi verso una configurazione compatibile con quel senso. È uno dei macchinari che contribuiscono a fabbricare il suo *embedding* contestuale.

Lo schema generale del blocco appare piuttosto semplice. Prima, ogni posizione raccoglie informazione dalle altre mediante l'attenzione; poi la sua rappresentazione passa per una piccola rete neurale che la trasforma posizione per posizione. Guardarsi intorno, elaborare localmente e ripetere nello strato successivo, con le connessioni residuali e le normalizzazioni che aiutano a mantenere stabile il percorso. Il *transformer* ripete questa alternanza molte volte. Ora viene la svolta che rende tutto un po' mitologico. In ogni strato lavorano in parallelo diverse "teste di attenzione" (*attention heads*), ciascuna con le proprie trasformazioni di query, key e value. Questo permette di calcolare simultaneamente relazioni diverse sullo stesso testo. Alcune teste sembrano sensibili a posizioni specifiche, a dipendenze sintattiche o a relazioni di correferenza; altre sono ridondanti o molto più difficili da interpretare, e in determinati modelli una proporzione importante può essere eliminata senza che le prestazioni crollino. Fra le più studiate ci sono le cosiddette "teste di induzione" (*induction heads*), che in certi modelli riconoscono sequenze del tipo A-B ... A e favoriscono che dopo compaia B. Non bisogna immaginare, tuttavia, funzioni fisse.

L'eco mitologica viene da Argo Panoptes, il guardiano che Era mise a sorvegliare Io. Aveva cento occhi e non dormivano mai tutti insieme. La sentinella perfetta. Sapendo che non avrebbe potuto vincerlo con la forza, Ermes gli suonò la siringa mentre gli raccontava storie, finché i cento occhi, uno dopo l'altro, si chiusero. Quegli occhi Era li raccolse e li mise sulla coda del pavone, dove restano aperti senza sorvegliare. Il transformer sembra un Argo amministrativo, con centinaia di occhi sullo stesso testo,

che calcola relazioni diverse allo stesso tempo. Non c'è racconto che lo addormenti, non si distrae mai, anche se a rigore nemmeno sta guardando.

Le mappe di attenzione non sono una radiografia del ragionamento. Una testa può assegnare molto peso a una parola, ma non è questa la ragione per cui il modello produce una determinata risposta. Distribuzioni di attenzione piuttosto diverse possono dare risultati simili e ci sono altri componenti nella causalità interna. Nel 2019 la discussione esplose quando uscì un lavoro intitolato "Attention Is Not Explanation", l'attenzione non è spiegazione, e un altro rispose presto, "Attention Is Not Not Explanation", che nemmeno si può affermare che non lo sia. Con prudenza, possiamo dire che i pesi di attenzione mostrano una parte del percorso dell'informazione senza pretendere trasparenza.

Torniamo a Simone Weil per esplorare che cosa implichi che il senso di "aspettare" non entri nel quadro. Questa filosofa aveva passato un anno in fabbriche, l'ultima delle quali la Renault, per sapere dall'interno che cosa faccia a una persona il lavoro alla catena. Morì a trentaquattro anni in un sanatorio di Ashford, nel 1943, lasciando quasi tutta la sua opera inedita, in quaderni dispersi. Albert Camus partecipò alla pubblicazione presso Gallimard e disse che era una delle intelligenze straordinarie del suo tempo. Per Weil l'attenzione nella sua forma più pura è imparentata con la preghiera, e l'amore del prossimo ha come sostanza l'attenzione in quanto capacità di guardare chi soffre e domandargli: "Qual è il tuo tormento?" e restare davanti alla risposta. L'attenzione weiliana non è tensione muscolare né caccia intellettuale. Al contrario, suppone di sospendere il pensiero, lasciarlo disponibile, vuoto, penetrabile dall'oggetto. L'attenzione non si precipita sulla verità, ma crea le condizioni perché arrivi. L'attenzione piena è un'attesa senza oggetto anticipato, senza riempire il vuoto con se stessi. Per questo è preghiera, rinuncia completa dell'io a occupare il centro perché possa occuparlo qualcos'altro, sia esso un problema di geometria, il prossimo afflitto o Dio.

Dunque l'attenzione del transformer porta all'estremo matematico la selezione e la ponderazione che William James aveva visto. Senza fatica né perdita di concentrazione esegue sempre la stessa operazione di distribuzione dei pesi con una precisione che ci supera. Tuttavia, precisione non equivale a pertinenza. Il meccanismo può ponderare pessimamente, ignorando ciò che è decisivo o appoggiandosi ad associazioni spurie. Per Weil l'attenzione corregge le nostre aspettative e ci obbliga a rivedere che cosa è importante. Il transformer non ha questa relazione con il suo oggetto. Se sbaglia, non c'è frustrazione né responsabilità di rettificare nulla.

Chiudiamo con un'ironia d'epoca. Nel 1971 Herbert Simon sostenne che l'abbondanza di informazione produce scarsità di ciò che l'informazione consuma, vale a dire attenzione. A che cosa dedicare tempo? A spese di che cosa? Quanto più cresce l'offerta di informazione, tanto più preziosa diventa la risorsa umana che la seleziona. Mezzo secolo dopo costruiamo macchine il cui meccanismo centrale porta il nome di "attenzione" e le usiamo per abbattere il costo della produzione di testo, immagini, audio e codice. L'attenzione computazionale è diventata abbondante e scalabile, ma l'attenzione umana continua ad avere la stessa capacità limitata. La tecnologia ha acuito la scarsità di Simon moltiplicando quasi senza limite ciò che reclama attenzione, senza moltiplicare l'attenzione di chi deve concederla. Abbiamo costruito una creatura straordinariamente abile nel tendere l'arco e

incapace di aspettare, l'attenzione dell'arciere senza arciere, i cento occhi di Argo sulla coda del pavone. L'attenzione della macchina funziona e ci ricorda che l'*attente*, almeno per ora, resta affar nostro.

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.

COME CITARE

Marsico, Claudia e Hernán Inverso. 2026. "L'attenzione: l'arco e l'attesa". *Mappe del labirinto*, saggio 15. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575278

Mappe del labirinto è una serie di Phantom Maze, AI & Language Lab. La collezione vive su **phantommaze.cloud**. L'iscrizione è gratuita.

Alcune delle idee che inquadrano questo testo sono sviluppate in *Il fantasma senza ego. Una mappa nel labirinto dell'intelligenza artificiale*, di Hernán Inverso e Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-19-5).