MAPPE DEL LABIRINTO

UNA SERIE DI PHANTOM MAZE — AI & LANGUAGE LAB

SAGGIO 18

Transformer: una manciata di pagine che hanno cambiato la forma di tutto

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.23575288

L'architettura che ha reso possibile l'IA generativa



Wenzel Hablik, Starry Sky, Attempt (1909)

Nel 2017, otto dipendenti di Google pubblicarono un articolo di poche pagine. Gli diedero un titolo da canzone: "Attention Is All You Need" strizza l'occhio a "All You Need Is Love" dei Beatles. Il documento interno di progettazione si chiamava "Transformers: autoattenzione iterativa ed elaborazione per vari compiti", e includeva un'illustrazione con sei pupazzi del franchise di giocattoli. Come è facile immaginare, il gruppo era noto come il Team Transformer. Il nome dell'architettura

rimase, così racconta la storia, perché a Jakob Uszkoreit, uno degli otto, evocava ricordi d'infanzia e si adattava bene all'idea di trasformare informazione. Nel paper precisarono che l'ordine degli autori era casuale. Oggi quell'articolo disinvolto e giocoso accumula più di duecentocinquantamila citazioni, figura tra i dieci lavori più citati del XXI secolo in qualunque disciplina e l'iniziale del transformer è finita sulla bocca di tutti attraverso la T di GPT. Gli otto, sia detto per inciso, hanno lasciato Google, quasi tutti per fondare aziende proprie. Esploriamo la storia di quell'architettura e dell'idea che l'ha resa tanto vincente.

Perfino nel mondo di Megatron e Optimus Prime, "transformer" porta con sé l'eredità del latino e del suo *transformare* (*trans*, "attraverso, al di là, da una parte all'altra", e *formare*, "dare forma", da *forma*). Trasformare è togliere qualcosa da una figura e metterlo in un'altra. Ovidio lo descrisse duemila anni fa nei due versi iniziali delle *Metamorfosi*: *In nova fert animus mutatas dicere formas / corpora*, "l'animo mi spinge a cantare le forme mutate in corpi nuovi". Seguono duecentocinquanta storie di corpi che cambiano forma, come Dafne in alloro, Aracne in ragno e Narciso in fiore, intrecciate nella tesi che nulla perisce e tutto si trasforma, cioè che la forma viaggia di corpo in corpo.

Curiosamente il nome dell'architettura si rivelò esatto anche se i suoi autori non lo cercavano, perché un transformer prende una sequenza di vettori, le rappresentazioni numeriche dei token, quei punti nello spazio del senso in cui ogni parola viene convertita in una posizione, e la trasforma, strato dopo strato, in un'altra sequenza di rappresentazioni. Nei suoi discendenti generativi, quelle trasformazioni finiscono per servire a calcolare quale token può venire dopo. Come una macchina ovidiana, muta forme.

Per capire che cosa cambiò nel 2017 bisogna guardare che cosa si faceva prima. Le reti ricorrenti elaboravano una sequenza passo dopo passo. Ogni posizione produceva uno stato interno che dipendeva da quello precedente. Le LSTM, sviluppate per conservare meglio l'informazione a lunga distanza, migliorarono molto quella memoria. Prima del transformer era già comparsa l'attenzione, che permetteva al decoder di consultare direttamente parti diverse della frase di ingresso. Il traduttore neurale di Google del 2016 combinava proprio quelle due soluzioni, i.e., profonde reti LSTM e un meccanismo di attenzione. Restava, tuttavia, un collo di bottiglia difficile da eliminare. La ricorrenza obbligava a fare i calcoli in ordine. Per arrivare alla posizione cento bisognava aver calcolato la novantanove. Era particolarmente scomodo per le GPU, progettate per eseguire enormi quantità di operazioni simultanee e qui ferme ad aspettare in coda. Il transformer fece la scommessa radicale di conservare l'attenzione ed eliminare la ricorrenza.

L'idea insolente del transformer fu, in effetti, eliminare del tutto la lettura sequenziale. Niente riassunti che viaggiano di parola in parola. Tutte le parole della frase si guardano tra loro in una volta, direttamente, ciascuna decidendo quali altre le importano per affinare il proprio senso. Nel decoder generativo c'è una restrizione. Ogni posizione può consultare ciò che precede, ma non guardare ciò che ancora non è accaduto. Questo non impedisce di parallelizzare l'addestramento, ma fa sì che, quando il modello genera, l'uscita debba comparire token dopo token. Questo meccanismo si chiama attenzione. Diciamo qui soltanto che al posto di un lettore che avanza col dito ora c'è una tavola

rotonda dove tutti i termini della frase si consultano simultaneamente l'uno con l'altro. Jakob Uszkoreit sospettava da tempo che quello da solo bastasse e che la ricorrenza fosse una stampella di cui si poteva fare a meno. Da qui il titolo "Attention is all you need". La congettura era così contraria al senso comune del settore che perfino suo padre, Hans Uszkoreit, noto linguista computazionale, era scettico. Forse anche Jakob aveva bisogno di attenzione.

Il transformer convertì la lettura, che era una fila, in un'assemblea pronta a sfruttare l'esercito di calcolatrici parallele della GPU. Il modello originale fu addestrato su otto schede per tre giorni e mezzo, una miseria per gli standard che sarebbero venuti. Bastò per fare epoca. Naturalmente, non è tutto rose e fiori. Se ogni parola consulta tutte le altre, il costo cresce con il quadrato della lunghezza del testo, sicché raddoppiare il contesto quadruplica il conto, ma qui ci concentreremo sul funzionamento e non sui costi.

Il trucco di spezzare un lavoro sequenziale per renderlo simultaneo ha un precedente medievale sorprendente. Nel XIII secolo, attorno alle università di Bologna e Parigi, la domanda di libri esplose. Migliaia di studenti avevano bisogno degli stessi testi di diritto, teologia e filosofia, ma un copista impiegava mesi a riprodurre un solo volume, perché copiare era un lavoro di una parola alla volta, proprio come leggere per una rete ricorrente. La soluzione fu il sistema della *pecia*, "pezzo" in latino. Il libraio custodiva un esemplare ufficiale del testo, rivisto e approvato dall'università, ma diviso in fascicoli numerati di poche carte. Ogni copista prendeva un pezzo alla settimana e così venti copisti potevano lavorare sullo stesso libro contemporaneamente. In questo modo il codice completo usciva in settimane, passando dalla fila all'attenzione simultanea. Somiglia parecchio a ciò che il transformer ha fatto alla lettura della macchina. La rete ricorrente era un copista solitario incatenato all'ordine del testo, mentre il transformer è la bottega della pecia, con il testo ripartito in pezzi che si elaborano tutti insieme, con sistemi di numerazione dei fascicoli e segni di posizione per non perdere l'ordine, perché lavorando in parallelo non si sa più che cosa veniva prima di che cosa. In entrambi i casi c'erano una grande domanda e un metodo artigianale che non scalava.

Tuttavia, bisogna andare oltre i copisti e le reti per illustrare il passaggio più interessante. In "L'Aleph" di Jorge Luis Borges, il narratore scende nella cantina di una casa di via Garay e trova una sfera piccola, di due o tre centimetri di diametro, dove stanno tutti i punti dell'universo insieme e senza sovrapporsi. Il problema diventa presto un problema di scrittura. Ha visto in simultanea ciò che deve scrivere in successione, perché il linguaggio ha quella struttura. Il transformer vive in parte quella stessa sproporzione, dato che guarda la frase intera in una volta, con tutte le parole che si consultano nello stesso momento, e poi deve consegnare il risultato come lo consegna il linguaggio, una parola dietro l'altra. L'asimmetria è costitutiva e il privilegio di guardare tutto insieme diventa un mal di testa al momento di raccontarlo.

Wenzel Hablik, Starry Sky, Attempt (1909)

Quel che seguì fu una colonizzazione completa. Il transformer del 2017 era nato per tradurre e perciò aveva due parti: un encoder, che riceveva la frase di origine e ne costruiva una rappresentazione, e un decoder, che usava quella rappresentazione per produrre la traduzione passo dopo passo. La sua

discendenza separò presto le due metà e scoprì che potevano vivere benissimo per conto proprio. BERT, presentato da Google nel 2018, conservò sostanzialmente il lato encoder. Legge il contesto in entrambe le direzioni e costruisce rappresentazioni particolarmente utili per comprendere, classificare e cercare. L'anno seguente Google lo integrò nel suo motore di ricerca per interpretare meglio certe richieste. GPT prese la strada complementare conservando il decoder e trasformandolo in una macchina addestrata a continuare una sequenza, predicendo un token a partire dai precedenti. La possibilità generativa comparve con Łukasz Kaiser e Noam Shazeer, due degli autori del transformer originale, che parteciparono poco dopo a un sistema solo decoder capace di produrre perfino articoli completi di Wikipedia a partire da documenti di riferimento. La linea di GPT fece di quell'idea un programma di scala crescente. Il suo nome conserva ancora la genealogia completa: *Generative Pre-trained Transformer*, con l'architettura quasi come cognome. Ma sarebbe un errore pensare che l'eredità sia rimasta chiusa nel linguaggio. Presto ci furono transformer che trattarono un'immagine come una sequenza di tasselli, altri che elaborarono voce e musica, e poi video, robotica e sistemi multimodali. Un'architettura concepita per un compito piuttosto concreto, tradurre da una lingua all'altra, ha finito per diventare un meccanismo generale per elaborare quasi qualunque cosa possa essere rappresentata come una sequenza di parti in relazione tra loro.

Ha vinto perché si lasciava addestrare in parallelo e permetteva di mettere in relazione posizioni lontane in una sequenza. Nell'era del calcolo massivo, l'architettura che meglio si distribuisce tra mille calcolatrici si prende tutto. È una lezione scomoda e ripetuta. Una parte importante dell'IA si spiega più con il costo di scalare che con un'astratta nozione di intelligenza. La macchina dal nome di giocattolo, quella che muta forme strato dopo strato come il poema di Ovidio, ha finito per essere protagonista della metamorfosi dell'intero campo. Finora lei stessa non è quasi cambiata. I modelli attuali hanno cambiato molti pezzi attorno a quell'architettura, ma la combinazione fondamentale di autoattenzione, strati feed-forward e connessioni residuali resta riconoscibile. Ci sono pretendenti al trono che promettono di schivare il suo famoso pedaggio quadratico. Nel frattempo, conviene ricordare che otto pagine del 2017 hanno deciso la forma di quasi tutto ciò che la macchina dice e che le forme, come insegnava Ovidio, sono solo di passaggio nei corpi.

COME CITARE

Marsico, Claudia e Hernán Inverso. 2026. "Transformer: una manciata di pagine che hanno cambiato la forma di tutto". *Mappe del labirinto*, saggio 18. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.23575288

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).

Fonti. L'articolo fondativo: 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 sull'ordine casuale degli autori e i dettagli dell'addestramento (8 GPU P100, 3,5 giorni per il modello grande). I dettagli di colore sul titolo beatlesiano, il nome scelto da Uszkoreit "perché gli piaceva come suonava", il documento di progettazione "Transformers: Iterative Self-Attention and Processing for Various Tasks" con i sei pupazzi del franchise, e lo scetticismo di Hans Uszkoreit, in Steven Levy, "8 Google Employees Invented Modern AI. Here's the Inside Story" (*Wired*, 2024), e in Madhumita Murgia, "Transformers: The Google Scientists Who Pioneered an AI Revolution" (*Financial Times*, 2023), che documenta anche l'uscita degli otto autori da Google. Il posto del paper tra i più citati del secolo, in Helen Pearson et al., "Exclusive: The Most-Cited Papers of the Twenty-First Century" (*Nature*, 2025). Ovidio, *Metamorfosi*, I, 1-2, e la dottrina della trasformazione universale nel discorso di Pitagora del libro XV. Jorge Luis Borges, "L'Aleph", pubblicato su *Sur* nel 1945 e raccolto in *L'Aleph* (1949): la sfera cangiante di due o tre centimetri nella cantina della casa di via Garay, e la disperazione del narratore davanti alla simultaneità di ciò che ha visto contro la successione che gli impone il linguaggio. Il traduttore neurale di Google del 2016: Yonghui Wu et al., "Google's Neural Machine Translation System" (2016). Sulla memoria a breve termine di lunga durata: Sepp Hochreiter e Jürgen Schmidhuber, "Long Short-Term Memory" (*Neural Computation*, 1997). Sul sistema della pecia: Jean Destrez, *La pecia dans les manuscrits universitaires du XIIIe et du XIVe siècle* (1935), lo studio classico che identificò esemplari e pezzi di Bologna, Parigi, Oxford e Napoli; sintesi nella voce "Pecia system" dell'*Oxford Companion to the Book* (Oxford University Press, 2010). Sulla discendenza: Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers" (2018); Alec Radford et al., "Improving Language Understanding by Generative Pre-Training" (2018), primo GPT; Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2021), il transformer di visione.