volerelaluna LA POLITICA PUNTOACAPO

La bolla dell'intelligenza artificiale e un gol segnato in contropiede

Autore: Franco Marra

Contropiede: dice la Treccani: «azione velocissima e improvvisa di contrattacco condotta mentre l'avversario si trova proteso all'attacco […] prendere, cogliere uno in c. (o di c.), coglierlo alla sprovvista, sorprenderlo in un momento critico». **L'Intelligenza Artificiale si basa sulla forza bruta matematica. Ogni cosa viene misurata e rappresentata in macchina secondo un gran numero di parametri (la cui scelta è già un atto politico)**. Questo porta ad elenchi molto lunghi di numeri ("vettori", uno per ogni cosa). La classificazione di un oggetto o di una relazione tra oggetti passa attraverso un calcolo sui loro vettori, che misura la loro affinità con i vettori di campioni noti. **Cose simili, con affinità alta, sono etichettate in una stessa categoria. Il calcolo è fatto da una struttura nota come "rete neurale", per affinità (appunto), in gran parte solo formale, con le corrispondenti strutture biologiche. A differenza del cervello umano però, che consuma come una lampadina da 20W, gli algoritmi di calcolo vettoriale delle reti neurali divorano spropositate quantità di energia**: quasi a celebrare la consacrazione del primato del biologico sul digitale!

Le reti neurali artificiali funzionano grazie a sequenze di micro scelte su miliardi di parametri, noti, nel loro insieme, come "modello", i cui valori sono tarati con una lunga e costosa serie di prove (*training*). Ad esempio, con il *training* si possono passare ad una rete ancora "ignorante" innumerevoli vettori di gattini (tipicamente popolati con i valori numerici dei *pixel* di un'immagine che li rappresenta), aggiustando per ogni gattino "certo" il valore dei parametri del modello in modo da avere sempre la stessa uscita (1: è un gattino). Quando è "addestrata" la rete determina con buona approssimazione se il prossimo vettore con cui viene alimentata sia quello di un gattino o no (0: è un'altra cosa), ed è in grado, per così dire, di porre la corrispondente immagine in un cassetto etichettato con la scritta "gattini". **Un classificatore automatico di immagini, che opera su base statistica. In tutto questo processo i conti si sprecano, il contatore elettrico impazzisce e tutto si scalda come una stufa**.

Nei sistemi cosiddetti "generativi", **il prossimo oggetto "predetto" dal calcolo verrà preso dal cassetto che ha l'etichetta "il più probabile" tra tutti quelli possibili nell'ambito del contesto corrente e del dialogo con l'umano con cui si interagisce**. Nel caso dei linguaggi, gli LLM (*Large Language Model*) che scrivono e traducono, altro non sono quindi che "estrusori di stringhe di testo probabili", come dice Daniela Tafani, del Dipartimento di Scienze Politiche dell'Università di Pisa, o "pappagalli stocastici", come definiti dal famoso articolo che provocò il licenziamento delle ricercatrici di Google Timnit Gebru e Margareth Mitchell. Pappagalli che non hanno nessuna possibilità di conoscere il significato di quello che generano. Si dice che talvolta delirano, la realtà è che producono sempre "stronzate" (il termine usato nella letteratura specialistica è letteralmente *bullshit*) senza senso, che per caso sono molto spesso vere.

Il contesto è stato per lungo tempo lo scoglio su cui sono naufragati traduttori automatici e generatori di testo. La parola “riso” si riferisce al cereale coltivato nel vercellese o all’omonima espressione emotiva? È una questione di affinità statistica con i termini che circondano quello sotto indagine, dove **il vettore associato al termine questa volta esprimerà le frequenze del suo accoppiamento con gli altri termini nei suoi dintorni**, come “mondina” o “diserbante” (nel caso di “riso”). Basta quindi solo un po’ di “attenzione” (*Attention is all you need* è titolo di un famoso articolo di otto ricercatori di Google). A questo punto è un gioco da ragazzi per le reti neurali eliminare l’ambiguità. **Qui i conti si sprecano ancora di più. La stufa diventa rovente e bisogna raffreddarla con fiumi di acqua. L’energia non basta più, bisogna tornare al nucleare**. La buona notizia è che si tratta di conti eseguibili in parallelo. Non bisogna aspettare un’eternità come nel caso della domanda fondamentale sulla vita, l’universo e tutto quanto posta nel romanzo *Guida galattica per gli autostoppisti* di Douglas Adams.

Negli anni ‘90 del secolo scorso molti di noi giocavano agli “sparatutto” come *Quake*. In questi giochi le scene sullo schermo vengono aggiornate in tempo reale in ogni loro minimo dettaglio grafico (*texture*). Per coerenza scenografia, tutti assieme. Un chiaro requisito di calcolo parallelo, che veniva eseguito da particolari schede (GPU: *Graphics Processing Unit*), sviluppate per la prima volta da una start-up che per parecchi anni boccheggiò nel limbo dei profitti: NVIDIA. Questa a sua volta acquistava i chip dalla TSMC, un’azienda di Taiwan specializzata nello sviluppo di chip su misura (*custom*), a cui passava le specifiche di progetto. TSMC oggi è così strategica da muovere i bombardieri cinesi nel suo cielo. Siccome il calcolo vettoriale è un calcolo parallelo, chi progettava intelligenza artificiale si buttò con entusiasmo sulle GPU, facendo lievitare i valori di capitalizzazione di mercato dell’intero settore a livelli fuori di testa.

Dopo quella *dotcom* dell’inizio del millennio, **abbiamo ora la bolla della IA**. **Sam Altman, CEO di OpenAI (dove nasce ChatGPT) soffia nella bolla promettendo, per attirare i *round* di investimenti, la AGI (*Artificial General Intelligence*), con capacità umane o quasi**, portando, a ottobre del 2024, la capitalizzazione della sua società a 157 miliardi di $, contro dei profitti di 3,7 miliardi di $ e una perdita nel corso dell’anno di 5 miliardi di $. Per uno spaventoso valore di 42 del rapporto tra capitalizzazione e fatturato! **Poco dopo Elon Musk annuncia che gli sviluppi IA di X si gioveranno di un *cluster* di 100.000 GPU NVIDIA H100 aumentabili fino a 1.000.000 unità**, e il programma *Stargate* da poco annunciato da OpenAI, Oracle e Softbank parla di 500 miliardi di $ da investire in infrastruttura di calcolo.

Le autovetture in USA negli anni 60 erano più simili a pachidermi che a mezzi di trasporto familiare. Grandi consumi, poca efficienza. *Bigger is better* (“più grande è meglio”) sembra lo slogan che caratterizza ogni fase infantile del capitalismo USA, che si traduce, nel caso della IA, in GPU sempre più potenti, raggruppate in enormi cluster e in consumi energetici sempre più alti. Allora arrivarono i giapponesi, e ora giungono i cinesi, a dare la sveglia. Con un momento *sputnik*, come lo definisce Marc Andreessen, papà del

primo browser della storia, fidato consigliere di Trump e vate della tecno-autocrazia. **Gli ingegneri cinesi, soggetti all'embargo delle GPU più potenti, hanno infatti fatto leva sulle caratteristiche della loro cultura – frugalità, cooperazione, e inventività (ma anche concentrazione del potere e poca protezione dei dati personali) – per entrare, con il botto, in un mercato popolato da brontosauri**, **segnando un clamoroso gol in contropiede con un agile ed economico *velociraptor*** (o forse un *cigno nero*, come si definisce in economia un evento improvviso che fa saltare tutte le previsioni)**: DeepSeek**. E pungendo la bolla finanziaria. Le azioni di NVIDIA negli ultimi giorni di gennaio perdono il 17%, e non si sa come andrà a finire in un mercato pieno di promesse di ritorni economici ancora lungi dall'essere mantenute. Ma l'indicatore più significativo è forse quello delle compagnie che si occupano di energia: Constellation, il più grande operatore USA nel campo dell'energia nucleare, è crollato del 20%.

Il costo per l'addestramento di DeepSeek sembra parecchio più basso di quello dei suoi diretti concorrenti. È un costo in gran parte energetico: **la IA cinese consuma dal 50% al 75% in meno (fonte DeepSeek) di quanto impiegato dalle più recenti GPU NVIDIA (H100)**, che infatti DeepSeek non usa, impiegando i "vecchi" H800. Un cluster di 100.000 unità come quello di Elon Musk consumerebbe così circa la metà dei corrispondenti 150MW. In realtà il tema è complesso e dibattuto, perché, in un'ottica di costo totale assai difficile da determinare, oltre al *training* andrebbe considerato anche quello legato alle risposte. Che a sua volta dipende dal "tipo" di ragionamento adottato. E qua per gli analisti e per le loro analisi sui posizionamenti competitivi tra le diverse piattaforme si aprono verdi praterie. Ancora, il modello della creatura di Liang Wenfeng è aperto, al contrario, ad esempio, di quello di ChatGPT, che perfino il suo papà Sam Altman afferma essere stato sviluppato dalla parte sbagliata della storia. Questo vuol dire che **il modello può essere scaricato da chiunque ed essere utilizzato, grazie anche ai ridotti consumi energetici e alla minore potenza richiesta, su molte macchine**. L'era del *downsizing*, come capitò ai computer con l'avvento del PC, è forse iniziata.

Gli americani reagiscono scompostamente: "ci hanno *succhiato* via il modello" (come se loro non stessero fregando dati da Internet da tutta una vita), "DeepSeek *non* risponde alle domande su Tienanmen", "DeepSeek *non* è sicuro" ma le loro reazioni non cambiano la sostanza del problema: **come in Vietnam, ha vinto chi mangia riso invece di *hot dog***, e chi calza sandali con la suola fatta con copertoni usati invece di stivaletti tattici. Ma soprattutto, di fronte alle difficoltà, usa il cervello (umano). E i servizi *premium* delle IA americane cominciano a essere erogati gratis, in un clima da guerra dei prezzi, ridimensionando sempre di più le speranze di un ritorno a breve degli investimenti.

E poi arrivano gli altri. Ad esempio, sempre dalla Cina arriva Qwen. Ma anche in Italia Almawave lancia Velvet, che gira sul supercalcolatore Leonardo di Cineca. Abbiamo cercato la sua chat, per provarlo, con risultati ambigui. Purtroppo Velvet è un nome spesso usato in altri ambiti, e le chat sono strumento frequentemente usato per altri scopi. Lasciamo al lettore giudicare, se ha la pazienza di cliccare su questo link. Ma noi

pensiamo che in fondo avesse ragione Fabrizio De André: «È mai possibile, porco d'un cane, / che le avventure in codesto reame / debban risolversi tutte con grandi puttane?».