I limiti dell’intelligenza artificiale

Parlare e dialogare sulla migliore approssimazione di Arte Artificiale realizzata sin qui dall’uomo, quella che comunemente viene definita Intelligenza Artificiale. Per leggere tutti gli articoli clicca qui.

Sono due le cose che fanno potente una macchina, velocità e memoria. Dalla macchina astratta di Turing, fino all’ultima meraviglia della tecnica **NVidia Blackwell**, la capacità di una macchina si valuta in velocità di calcolo e memoria disponibile.

La velocità di calcolo fino a poco tempo fa si stabiliva in **Hertz**, **GigaHertz** e numero di *Cores* che li eseguivano. La memoria di una macchina, da un po’ di anni si misurava in Terabyte di disco, Gigabyte di RAM e Megabyte di cache del Processore.

Per coloro che non avessero dimestichezza con la struttura interna delle macchine, la memoria su disco, quella su RAM e quella cache del Processore sono più capienti, ma più lente, più costose dal punto di vista costruttivo, nell’ordine in cui le abbiamo elencate.

L’Intelligenza Artificiale ed il suo impiego massiccio comportano un cambio di paradigma. Alimentare neural networks Feed Forward con quantità colossali di informazioni nel corso del loro training, o verificarle verso modelli giganteschi in fase di inferenza, richiede enormi quantità di memoria (40-80GB, nel 2024) ad altissime prestazioni, paragonabili a quelle della memoria cache dei vecchi processori, vicinissime alle unità di calcolo ovvero alle GPU.

Quindi, ricapitolando, le vecchie architetture con memorie più grandi, ma più lente all’allontanarsi dall’unità di calcolo,

sono inutili per l’AI? **La risposta è sì.**

La motivazione profonda di questa situazione, risiede nel II teorema di Shannon, nell’enunciato del termine entropia (dal greco: en tropos, trasformazione interna) e da una celeberrima conversazione che Shannon scambiò con Von Neumann: “*La mia più grande preoccupazione era come chiamarla. Pensavo di chiamarla informazione, ma la parola era fin troppo usata, così decisi di chiamarla incertezza. Quando discussi della cosa con John Von Neumann, lui ebbe un’idea migliore. Mi disse che avrei dovuto chiamarla entropia, per due motivi: “Innanzitutto, la tua funzione d’incertezza è già nota nella meccanica statistica con quel nome. In secondo luogo, e più significativamente, nessuno sa cosa sia con certezza l’entropia, così in una discussione sarai sempre in vantaggio*“.

In soldoni, quando si scambiano messaggi ad alta velocità si realizza un’autotrasformazione, un effetto indesiderato, caotico e pernicioso che comunemente chiamiamo Errore, ma che in realtà è il fenotipo dell’’entropia alimentata dal calore che la corrente sta generando nel trasferimento dei messaggi stessi. Per minimizzare questo errore, le cui origine ricorda Von Neumann non sono così chiare e si relazionano con la III legge della Termodinamica e con il calore prodotto e dissipato durante questi processi, è fondamentale che questi messaggi percorrano meno strada possibile autotrasformando meno energia possibile. Questo il motivo per cui si spinge su processi costruttivi a minimo numero di nanometri e per cui un dissipatore da GPU è più voluminoso della GPU stessa.

Nota: L’Entropia, una misura straordinaria su tutti gli stati possibili di un sistema, sul minimo stato energetico delle cose e sul concetto stesso di tempo. Ma anche il principale limite fisico con cui combattiamo dal 1868 in tutte le sue forme.

La digressione sopra serve per fissare dei punti fermi che l’AI porta con sé. Essendo in finale, logica costruita dai dati, l’AI ha bisogno di movimentare volumi di informazioni molto maggiori di altre tecnologie che invece trasferivano la logica sottoforma di istruzioni ad un processore, il quale elaborava pochi dati per volta. Ciò, per il II teorema di Shannon, ed ancora indietro per la III legge della termodinamica, implica che le architetture GPU necessitino anche di tantissima memoria vicinissima alla parte di calcolo.

In questo passaggio emerge una problematica che stiamo

affrontando nelle architetture server, GPU come la nuova NVidia Blackwell forniscono un boost 10-100X rispetto alle precedenti portando più memoria, più veloce, direttamente nei pressi dell'unità di calcolo e puntando su bus di comunicazione ad altissime prestazioni che consolidano, entro certi limiti, la memoria disponibile su nodi diversi. Ma per tutto il resto, come faremo?

Ecco, attualmente non lo sappiamo. Operiamo devices come gli smartphone che hanno necessità di consumare poca corrente, non possono montare dissipatori su cui cucinare una bistecca perché risiedono nelle nostre tasche e li manipoliamo con le mani non guantate, dunque non possiamo risolvere con la strategia adottata per i server. Ok, negli smartphone come nei PC di casa (ed ARM sta puntando tutto su questo) stiamo inserendo dei co-processori per AI che sono specializzati nel calcolo matriciale, ma non risolvono il problema descritto sopra, quello della memoria e dell'entropia che funesta il trasferimento delle informazioni quando si parla di Gigabyte/s.

Per avere un riferimento, un LLM decente nel 2024, richiede 8-12GB di memoria ed un iPhone 15 PRO ne ha disponibile di 8 per tutto il sistema operativo. Per ovviare a questo problema, Google e Microsoft sono impegnate nello sviluppo di modelli LLM con footprint da 1.5-2.0GB, ma ciò non risolve il problema dell'entropia perché su smartphone non abbiamo una vera architettura GPU, siamo ancora su CPU + memoria RAM. Inoltre, questi modellini leggeri, producono anche risultati opinionati. Senza sorprese, l'AI estrae logica dai dati, dunque meno dati, modelli più piccoli, comportano meno logica.

Esiste sempre anche una via puramente commerciale, ad esempio quella che vorrebbe convincerci sull'utilizzo di tecnologia Small Language Model. Gli SLM sono prodigiosi per task estremamente specifici, ma la vita di tutti i giorni e la varietà di task che affrontiamo con il nostro smartphone hanno poco o nulla di specifico.

Attualmente, il limite è teoretico, scolpito nel granito che ci hanno lasciato quei burloni di Shannon e Von Neumann. Servirà svoltare nella scienza dei superconduttori o nella comunicazione quantistica per superare l'attuale limitazione. Una considerazione da fare quando si acquisterà il prossimo ultimo modello di smartphone. Una considerazione che ha già fatto chi sta acquistando decine e centinaia di migliaia di GPU perché sa già che, nei lustri a venire, saranno l'unica fonte di AI, via Cloud, per tutti i device mobili sul terreno.

Tutto ciò, resterà vero, almeno finché l'AI resterà sulle Neural Networks. Ma la discussione su un AI diversa sarà per un altro articolo, con quella attuale abbiamo per l'appunto grossi problemi di accesso alla memoria e sono invalicabili.

https://www.key4biz.it/i-limiti-dellintelligenza-artificiale/4

85257/