Ecco come funziona l'IA di DeepSeek che fa tremare i colossi statunitensi

di Tongliang Liu/The Conversation

La notizia che i modelli di intelligenza artificiale realizzati dalla piccola azienda cinese sono competitivi con quelli statunitensi, seppur realizzati con molte meno risorse finanziare e di calcolo, scuote i mercati ed entusiasma i ricercatori

L'azienda cinese di intelligenza artificiale (IA) DeepSeek ha fatto tremare la comunità tecnologica con il rilascio di modelli di IA estremamente efficienti, in grado di competere con i prodotti all'avanguardia di aziende statunitensi come OpenAI e Anthropic. Fondata nel 2023, DeepSeek ha ottenuto i suoi risultati con solo una frazione del denaro e della potenza di calcolo dei suoi concorrenti.

Il modello R1 "ragionante" di DeepSeek, rilasciato la scorsa settimana, ha suscitato l'entusiasmo dei ricercatori, lo sgomento degli investitori e le reazioni dei grandi dell'intelligenza artificiale. Il 28 gennaio l'azienda ha presentato un modello in grado di lavorare sia con le immagini sia con il testo.

Che cosa ha fatto DeepSeek e come lo ha fatto?

A dicembre DeepSeek ha rilasciato il modello V3. Si tratta di un modello linguistico "standard" molto potente che ha prestazioni simili a quelle di GPT-4o di OpenAI e del Claude 3.5 di Anthropic. Anche se questi modelli sono soggetti a errori e a volte inventano le risposte, possono svolgere compiti come rispondere a domande, scrivere saggi e generare codice informatico. In alcuni test di risoluzione dei problemi e di ragionamento matematico, ottengono punteggi migliori rispetto alla media degli esseri umani.

V3 è stato addestrato a un costo dichiarato di circa 5,58 milioni di dollari. Si tratta di un costo nettamente inferiore a quello di GPT-4, per esempio, il cui sviluppo è costato più di 100 milioni di dollari. DeepSeek sostiene inoltre di aver addestrato V3 usando circa 2000 chip di computer specializzati, in particolare le GPU H800 di NVIDIA. Anche in questo caso si tratta di un numero molto inferiore rispetto ad altre aziende, che potrebbero aver utilizzato fino a 16000 chip H100 più potenti.

Il 20 gennaio DeepSeek ha rilasciato un altro modello, chiamato R1. Si tratta di un modello cosiddetto "di ragionamento", che cerca di risolvere problemi complessi passo dopo passo. Questi modelli sembrano essere migliori in molti compiti che richiedono un contesto e hanno più parti interconnesse, come la comprensione della lettura e la pianificazione strategica. Il modello

R1 è una versione ottimizzata di V3, modificata con una tecnica chiamata apprendimento per rinforzo. R1 sembra funzionare a un livello simile a quello di o1 di OpenAI, rilasciato l'anno scorso.
DeepSeek ha usato la stessa tecnica per realizzare versioni "ragionanti" di piccoli modelli open source che possono essere eseguiti su computer domestici.
Questo rilascio ha scatenato un'enorme ondata di interesse per DeepSeek, facendo aumentare la popolarità della sua app chatbot alimentata da V3 e innescando un massiccio crollo dei prezzi dei titoli tecnologici, mentre gli investitori rivalutano il settore dell'IA. Al momento in cui scriviamo, il produttore di chip NVIDIA ha perso circa 600 miliardi di dollari di valore.
Come ha fatto DeepSeek
Le scoperte di DeepSeek sono consistite nel raggiungimento di una maggiore efficienza: ottenere buoni risultati con meno risorse. In particolare, gli sviluppatori di DeepSeek sono stati pionieri di due tecniche che potrebbero essere adottate dai ricercatori di IA in modo più ampio.
La prima ha a che fare con un'idea matematica chiamata "sparsità". I modelli di intelligenza artificiale hanno molti parametri che determinano le loro risposte agli input (V3 ne ha circa 671 miliardi), ma solo una piccola parte di questi parametri viene usata per ogni dato input. Tuttavia, prevedere quali parametri saranno necessari non è facile. DeepSeek ha utilizzato una nuova tecnica per farlo e poi ha addestrato solo quei parametri. Di conseguenza, i suoi modelli hanno richiesto un addestramento di gran lunga inferiore rispetto a un approccio convenzionale.
L'altro trucco ha a che fare con il modo in cui V3 memorizza le informazioni nella memoria del computer. DeepSeek ha trovato un modo intelligente per comprimere i dati rilevanti, in modo da facilitarne la memorizzazione e l'accesso rapido.
Che cosa significa
I modelli e le tecniche di DeepSeek sono stati rilasciati sotto licenza MIT, il che significa che chiunque può scaricarli e modificarli. Se da un lato questa può essere una cattiva notizia per alcune aziende di IA – i cui profitti potrebbero essere erosi dall'esistenza di modelli potenti e liberamente disponibili – dall'altro è un'ottima notizia per la più ampia comunità di ricerca sull'IA.
Attualmente, molte ricerche sull'IA richiedono l'accesso a enormi quantità di risorse informatiche. I ricercatori come me che lavorano nelle università (o in

qualsiasi altro posto che non sia una grande azienda tecnologica) hanno avuto una capacità limitata di effettuare test ed esperimenti. Modelli e tecniche più efficienti cambiano la situazione. La sperimentazione e lo sviluppo potrebbero ora essere significativamente più facili per noi.

Anche per i consumatori l'accesso all'IA potrebbe diventare più economico. Un numero maggiore di modelli di IA potrebbe essere eseguito sui dispositivi degli utenti, come computer portatili o telefoni, anziché essere eseguito "nel cloud" a pagamento. Per i ricercatori che dispongono già di molte risorse, una maggiore efficienza potrebbe avere un effetto minore. Non è chiaro se l'approccio di DeepSeek contribuirà a creare modelli con prestazioni complessivamente migliori o semplicemente modelli più efficienti.

L'autore Tongliang Liu è professore associato di apprendimento automatico e direttore del Sydney AI Centre dell'Università di Sydney, in Australia.