Intelligenza artificiale. Origini e modelli linguistici dell'IA

L'espressione 'intelligenza artificiale' rappresenta oggi un insieme di hardware e software in grado di simulare capacità umane quali il ragionamento, l'apprendimento, la pianificazione e la creatività, grazie all'analisi degli effetti delle azioni precedenti.

Usata per la prima volta in maniera ufficiale, nel 1956 da **John McCarthy** durante una conferenza a Dartmouth, negli USA, la terminologia 'intelligenza artificiale' indicava inizialmente software dedicati alla sola funzione di *problem solver*.

Efficaci per svolgere alcune operazioni, tali software non erano in grado di interpretare un ambiente reale, né di risolvere problemi al di fuori di quelli consentiti dai dati immessi.

Negli anni '70 comparirono i sistemi esperti, che riuscivano a elaborare dati tramite il calcolo delle probabilità.

Con l'intelligenza artificiale nacquero anche i primi studi sulla realizzazione di una rete simile a quella dei neuroni biologici e lo sviluppo delle prime reti neurali non lineari.

Negli anni '80 e '90 i computer registrarono una notevole crescita nella capacità di memoria e nella potenza di calcolo computazionale. Consentirono il superamento di limiti incontrati negli anni precedenti e il conseguente sviluppo di programmi sempre più vicini alla definizione odierna di

intelligenza artificiale generativa.

Gli esordi dell'intelligenza artificiale generativa

Era il 1950 quando **Alan Turing**, uno dei padri dell'informatica, introdusse il suo articolo ***Computing machinery and intelligence*** con la domanda "*Can machines think?*" intendendo la capacità di un computer di avere un comportamento intelligente, alla stregua di un essere umano.

Per rispondere a questa domanda realizzò un test definito "***Test di Turing***" o "***Imitation Game***" con il quale un pc e un uomo avrebbero risposto alle stesse domande di un giudice.

Il computer sarebbe stato 'intelligente' solo se il giudice non fosse stato in grado di riconoscere quali risposte aveva dato la macchina e quali l'uomo.

Oggi, i progressi tecnologici, velocissimi, nel campo dell'intelligenza artificiale, stanno trasformando diversi aspetti del vivere 'umano' e ponendo un grande punto interrogativo sul significato stesso di 'essere umano'.

Intelligenza artificiale generativa e modelli linguistici

L'intelligenza artificiale generativa si basa su modelli linguistici di grandi dimensioni (LLM, dall'acronico inglese Large Language Models): categorie di elaborazione del linguaggio con parametri molto ampi.

Presente e futuro della tecnologia, gli LLM sono largamente utilizzati per la ricerca online (basti pensare che se si inserisce una parola su una barra di ricerca essi suggeriscono le parole successive), rispondono a una vasta gamma di richieste e assumono forme diverse.

Sono utilizzati per la traduzione delle lingue, per la scrittura di un testo a partire da un'indicazione generale, per la realizzazione di un riassunto a partire da un elaborato complesso, per la classificazione di un testo, per la generazione e il completamento di codici, di programmazione, per il riconoscimento di un testo all'interno di un'immagine, del parlato e della grafia e per l'analisi di emozioni e per la composizione di un brano musicale.

Intelligenza artificiale generativa: i compiti dei modelli linguistici dell'AI

- segmentazione di informazioni in frasi
- tokenizzazione delle parole
- derivazione delle parole
- lemmatizzazione delle parole
- etichettatura delle parti del discorso
- individuazione ed eliminazione delle stopword (congiunzioni, articoli, preposizioni)
- riconoscimento dell'entità denominata
- classificazione del testo
- frammentazione del testo in frasi di senso compiuto
- ricerca delle espressioni che si riferiscono alla stessa entità all'interno di un testo

Gli LLM sono diversi dai modelli linguistici tradizionali, perché sono basati su un trasformatore, una rete neurale di deep learning, milioni di parametri (o pesi) e un ampio addestramento che li ottimizza per la realizzazione di un modello con un tasso di errore che sia il minore possibile.

Per fare ciò, sono necessari set di dati di grandi dimensioni (da cui deriva il termine 'grande modello linguistico') che rendono i modelli più accurati, ma che richiedono più memoria e che hanno una velocità di esecuzione inferiore.

I set di dati di grandi dimensioni comportano due ipotetici problemi: la violazione del copyright e la spazzatura. La spazzatura può essere ripulita, ma la violazione del copyright è oggetto di cause legali. Da qui la necessità di una regolamentazione per la sicurezza dell'AI.

Per ovviare a questi due limiti, diverse aziende sviluppano e migliorano costantemente LLM. Ogni mese Meta AI, Google AI, OpenAI e Anthropic rilasciano aggiornamenti o nuove versioni. Eppure la percentuale di aziende italiane che usano l'AI è ancora residuale.

I principali modelli linguistici di grandi dimensioni oggi utilizzati

Famiglia GPT

GPT (acronimo di Generative Pretrained Transformer) è il modello di Chatbot probabilmente più conosciuto. Produce testi, crea domande e risposte e funge anche da traduttore tra diverse lingue. Nato nel 2018 sull'architettura della rete neurale Transformer di **Google**, oggi utilizza circa **117 milioni di parametri**. È un trasformatore unidirezionale pre-addestrato sul Toronto Book Corpus con un obiettivo di modellazione del linguaggio causale. Prevede quindi il token successivo all'interno di una sequenza.

Nel 2019 viene rilasciato GPT-2, composto da 1,5 miliardi di parametri e addestrato su un set di dati di 8 milioni di pagine web, per un totale di circa 40 GB di dati di testo.

È del 2020 GPT-3, modello linguistico autoregressivo (ogni variabile dipende dalla precedente) composto da 175 miliardi di parametri e addestrato su una combinazione di Common Crawl, Webtext2, Books1, Books2 e Wikipedia inglese. Usa la rete neurale di GPT-2 con qualche blocco aggiuntivo.

Nel 2021 esce CODEX, diretto discendente di GPT-3, ottimizzato per la generazione di codice su 54 milioni di repository GitHub open source.

Nel 2022 GPT-3.5 ingloba gli aggiornamenti di GPT-3 e CODEX e aggiunge un modello ottimizzato per la chat.

GPT-4, il modello multimodale di grandi dimensioni del 2023, accetta input di immagini e testo ed emette output di testo con prestazioni paragonabili a quelli di un essere umano su determinati parametri professionali e accademici.

Della famiglia GPT fanno parte ChatGPT e BingGPT, chatbot nati con GPT-3.5 e aggiornati a marzo 2023 per utilizzare GPT-4. Attualmente, per usare ChatGPT basata su GPT-4 è necessario avere un abbonamento a ChatGPT Plus.

La versione gratuita, ChatGPT standard, è basata su GPT-3.5 e addestrata su dati interrotti a settembre 2021, così come BingGPT, "The New Bing", al quale si accede direttamente se si utilizza il browser Microsoft Edge.

Curiosità sulla famiglia GPT

I modelli GPT fanno capo a **OpenAI**, società di ricerca e distribuzione di intelligenza artificiale che ha come mission "garantire che l'intelligenza generale artificiale avvantaggi tutta l'umanità".

OpenAI inizialmente limitò l'accesso a GPT-2 perché avrebbe potuto generare notizie false. Con il trascorrere dei mesi l'azienda cedette, e i potenziali problemi dichiarati si acuirono con GPT-3, versione che tende molto ad "allucinare", a realizzare testi credibili su fatti non accaduti.

Le versioni GPT-3.5 e GPT-4 hanno avuto lo stesso problema, anche se in misura minore, e si sono evolute notevolmente nell'arco di un paio di mesi.

OpenAI per ragioni competitive non ha spiegato come è stato addestrato GPT-4, data la concorrenza tra Microsoft (che ha finanziato OpenAI per 13 miliardi di dollari) e **Google**.

LLaMA

Altro concorrente di ChatGPT, è un modello linguistico di grandi dimensioni 'grezzo' da 65 miliardi di parametri.

LLaMA (Large Language Model Meta AI), chatbot rilasciato da **Meta AI**, richiede meno potenza di calcolo e risorse perché addestrato su un ampio set di dati senza etichetta, adatti quindi a una varietà di attività.

Llama2, l'evoluzione di LLaMA, ha il 40% di dati in più (2 trilioni di token da fonti disponibili al pubblico) e il doppio della lunghezza del contesto.

È disponibile per paramenti di 7, 13 o 70 miliardi.

Bard

Concorrente di ChatGPT, è il servizio di intelligenza artificiale conversazionale di Google del 2023. Genera testi su argomenti generali e fornisce spunti di idee quando vengono richiesti.

Aggiornato diverse volte dal suo rilascio, ha la capacità di generare codici in 20 diversi linguaggi di programmazione. Nel luglio 2023 ha ottenuto l'input di 40 lingue, ha incorporato **Google Lense** e ha aggiunto funzionalità di sintesi vocale in oltre 40 lingue.

Si basa sul modello PaLM-E, che a sua volta si basava sul modello LaMDA (Language Model for Dialogue Applications), un modello linguistico realizzato da Google nel 2021 basato su Transformer addestrato sul dialogo e messo a punto per migliorare significativamente la sensibilità e la specificità delle sue risposte.

PaLM-E è un modello linguistico multimodale incorporato realizzato nel 2023 da Google. I ricercatori hanno prima realizzato la versione PaLM e poi hanno incorporato i dati dei sensori aggiungendo l'agente robotico (e quindi la E nel nome). È anche un modello di visione e linguaggio con pochi margini di errore. Svolge diverse attività, dal ragionamento basato sul buon senso, al calcolo aritmetico, dalla spiegazione di barzellette, alla traduzione di testi e alla generazione di codici per la programmazione.

PaLM è il modello Transformer di sola decodifica di **Google Research** del 2022 addestrato con il sistema Pathways con 540 miliardi di parametri.

PaLM (Pathways Language Model) ha set di dati in lingua inglese e multilingue, che includono documenti web, libri, Wikipedia, conversazioni e codice GitHub di alta qualità.

Google ha anche creato un vocabolario "senza perdite", che preserva tutti gli spazi bianchi fondamentali per il codice, divide i caratteri Unicode fuori dal vocabolario in byte e divide i numeri in singoli token, uno per ogni cifra. Esiste anche PaLM-Coder, versione di PaLM540B ottimizzata su un set di codice solo Python.

Claude 2

Rilasciato da Anthropic, accetta fino a 100mila token (circa 75 mila parole) in un singolo prompt, con una singola indicazione.

Tra i suoi compiti rientrano la modifica, la riscrittura e il riassunto di testi, la classificazione e l'estrazione di dati strutturati e la possibilità di creare domande e dare risposte in base all'input che riceve.

Ha una vasta conoscenza dei linguaggi di programmazione. Lavora benissimo in inglese, ma ha un alto livello in diverse altre lingue comuni e riesce a comunicare anche in lingue meno

comuni. Addestrato per essere utile, onesto e innocuo, difficilmente realizza testi offensivi o pericolosi.

Per gli utenti statunitensi e britannici è disponibile con una versione beta gratuita ed è stato adottato da partner commerciali quali Jasper (una piattaforma di intelligenza artificiale generativa), **Sourcegraph Cody** (una piattaforma di intelligenza artificiale di codice) e Amazon Bedrock.

Non è ancora disponibile in italiano.

T5

Con circa 220 milioni di parametri totali, è un modello di chatbot Text-To-Text Transfer Transformer (T5) di Google che trasforma tutte le attività dell'elaborazione del linguaggio naturale (NLP: Natural Language Processing) in un formato testo-testo unificato, per cui input e output sono stringhe di testo.

Usa l'apprendimento di trasferimento, per cui un modello viene prima pre-addestrato su un'attività ricca di dati e poi perfezionato su un'attività specifica.

BERT

LLM di Google realizzato nel 2018 sull'architettura di rete neurale Transfomer dell'azienda. È stato addestrato su Wikipedia in inglese e Toronto Books Corpus. Utilizza il Masked Language Modeling (MLM) in cui il 15% circa dei token viene corrotto per l'addestramento. Progettato per pre-addestrare rappresentazioni bidirezionali profonde da testo senza etichetta, condizionando il contesto sia destro che sinistro a tutti i livelli. La chat bot iniziale usava 100 e 340 milioni di parametri totali.

ELMo

Modello linguistico di grandi dimensioni, definisce sia le caratteristiche complesse dell'uso delle parole, sia il modo in cui questo uso varia in base al contesto di riferimento.

Il suo acronimo, Embedding for Language Model, rappresenta una sequenza di parole come una sequenza di vettori. È stato creato nel febbraio 2018 dai ricercatori dell'Allen, Institute for Artificial Intelligence e dell'Università di Washington.

Probabilmente, nessuno di questi LLM raggiunge ancora la vera intelligenza artificiale generativa, dato che tutti sono soggetti a errori e quindi alla produzione di testi imprecisi e informazioni distorte o errate per tre motivi principali:

- gli LLM inventano fatti, chiamati allucinazioni, che possono sembrare credibili ma che non sono mai esistiti
- le traduzioni proposte dagli LLM non vengono corrette da un madrelingua, fatta eccezione per le frasi più comuni
- i codici di programmazione generati dagli LLM spesso presentano bug o non possono essere eseguiti

Origine e storia dei modelli linguistici di grandi dimensioni

I modelli linguistici nascono con **Andrey Markov** che nel 1913 applicò la matematica alla poesia. Dimostrò che nell'opera ***Eugene Onegin*** del poeta **Aleksandr Sergeevič Puškin** la probabilità della comparsa di un carattere dipendeva dal carattere precedente, con un'alternanza tra consonanti e vocali.

Questa applicazione viene oggi utilizzata per descrivere una sequenza di eventi in cui la probabilità di un evento dipende dal precedente.

Nel 1948 **Claude Elwood Shanon** riprese il lavoro di Markov e lo

estese realizzando la sua teoria delle comunicazioni.

Nel 1985 **Fred Jelinek** e **Robert Mercer** dell'IBM ripresero il lavoro di Markov per produrre un modello linguistico basato sulla convalida incrociata, che chiamarono stime cancellate e che prevede che un modello linguistico statistico assegni probabilità a sequenze di parole.

In un articolo del 2000, la cui prima firma è **Yoshua Bengio**, è stato descritto nel dettaglio il modello linguistico probabilistico neurale, con il quale le reti neurali sostituiscono le probabilità in un modello linguistico statistico.

I modelli linguistici hanno continuato a ingrandirsi nel tempo, con l'obiettivo di migliorare le prestazioni, ma gli svantaggi di tale crescita hanno portato, nel 2021, alla stesura del documento '***On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?***', con il quale si insinua il dubbio che la tendenza al 'più grande è e migliore è' non sia sempre corretta se si considerano la valutazione dei costi ambientali e finanziari e l'investimento di risorse per la cura e per la documentazione dei set di dati.

https://www.key4biz.it/intelligenza-artificiale-origini-e-modelli-linguistici-dellia/474598/