COME FUNZIONA L'IA · PER COMINCIARE · IN PAROLE SEMPLICI

Come fa un'IA a continuare una frase?

Il percorso di un modello linguistico raccontato senza numeri e senza tecnicismi.

Immaginiamo di dare a un'intelligenza artificiale l'inizio di questa frase:

`Domani` `vado` `al` `?`

Il suo compito è trovare la parola che viene dopo. A noi basta un attimo per pensare «mare». L'IA, invece, deve compiere una lunga serie di passaggi. Non perché sia più lenta o più stupida di noi, ma perché non vede le parole e i significati come li vediamo noi: per lei tutto deve diventare qualcosa che si può calcolare.

Seguiamo il percorso tappa per tappa. Le tappe hanno gli stessi numeri, da 0 a 12, che ritroverai nel riassunto grafico e nella presentazione animata: lì gli stessi passaggi li approfondiremo con i numeri e le operazioni.

TAPPA 0

Prima di tutto: ha letto moltissimo

Prima di poter rispondere, l'IA è stata addestrata. Ha letto un'enorme quantità di testi e, a ogni frase, ha provato a indovinare la parola successiva. Quando sbagliava, si correggeva di pochissimo. Questo è successo miliardi di volte.
Alla fine non conserva i testi che ha letto, ma tutto ciò che ha imparato da essi, sotto forma di moltissimi numeri regolati con cura. Quando ti risponde non impara più: usa ciò che ha già imparato.

Nel riassunto grafico: passo 0, da dove arrivano i numeri.

TAPPA 1

Ritagliare la frase in pezzi

È come se prendesse una frase scritta e la ritagliasse in tanti cartoncini separati.

`Domani` `vado` `al`

Questi pezzi si chiamano **token**. Spesso sono parole intere, come qui; ma una parola rara o lunga può essere spezzata in due o tre pezzi, e anche un apostrofo o una virgola possono essere un pezzo a sé.
Ogni pezzo viene poi cercato in un elenco, il **vocabolario** dell'IA, dove ha un suo numero, come il numero di una pagina. Attenzione: il vocabolario non è un dizionario con le definizioni. Non dice che «vado» è un verbo o che «Domani» indica il futuro: è soltanto l'elenco dei pezzi che l'IA conosce.
Nel riassunto grafico: passo 1, tokenizzazione.

TAPPA 2

Richiamare ciò che sa di ogni pezzo

Per ogni pezzo del vocabolario l'IA conserva una specie di «scheda» che ha costruito durante l'addestramento. Non è una definizione: è il ricordo di come quel pezzo viene usato nei testi.
Pezzi usati in frasi simili hanno schede simili. La scheda di «mare» assomiglia a quelle di «spiaggia», «estate», «vacanza», «onde». È un po' come succede a noi: se qualcuno dice «neve», ci vengono in mente «freddo», «inverno», «montagna».
Ma «simile» non vuol dire «con lo stesso significato»: «caldo» e «freddo» sono contrari, eppure hanno schede vicine, perché si usano nelle stesse frasi («oggi fa…»). E la scheda di un pezzo è sempre la stessa, in qualunque frase compaia: sarà il lavoro delle tappe successive ad adattarla al contesto.
Nel riassunto grafico: passo 2, embedding e mappa delle somiglianze.

TAPPA 3

Ricordare la posizione

Non basta sapere quali pezzi ci sono: bisogna sapere anche in che ordine sono. Pensiamo a queste due frasi:

Il cane morde l'uomo	*L'uomo morde il cane*

Le parole sono le stesse, il significato no. Per questo l'IA aggiunge a ogni pezzo un'informazione sulla sua posizione: chi viene prima, chi dopo.
Nel riassunto grafico: passo 3, la posizione.

TAPPA 4

Ogni pezzo prepara una domanda, un'etichetta e un contenuto

Ora i pezzi devono scambiarsi informazioni. Per farlo, ognuno prepara tre cose, come a una riunione in cui ciascuno arriva con:

- una **domanda**: «che cosa mi servirebbe sapere?»;
- un'**etichetta**: «ecco di che cosa mi occupo»;
- un **contenuto**: «ecco che cosa posso raccontare, se mi chiedono».

L'IA prepara queste tre informazioni con regole che ha imparato durante l'addestramento, le stesse per ogni frase.

Nel riassunto grafico: passo 4, Q, K e V.

TAPPA 5

Cercare i collegamenti

Seguiamo l'ultimo pezzo, «al»: è da lui che dipende la parola successiva. «Al» confronta la sua domanda con l'etichetta di ogni pezzo della frase, compresa la propria:

`Domani` `vado` `al`

Da ogni confronto esce un punteggio: quanto quel pezzo sembra utile ad «al», secondo ciò che l'IA ha imparato. Non è una misura di quanto due parole si somigliano, e non è un ragionamento come il nostro: è il risultato di un confronto imparato dai testi.

Nel riassunto grafico: passo 5, il punteggio di attenzione.

TAPPA 6

Non guardare avanti

Durante l'addestramento l'IA legge frasi intere, e deve imparare a indovinare ogni parola usando soltanto quelle che vengono prima. È come quando un insegnante copre con la mano l'ultima parola di una frase e chiede agli studenti di indovinarla: se si potesse sbirciare, non si imparerebbe niente.

`Domani` `vado` `al` 

Per questo ogni pezzo può guardare solo sé stesso e quelli che lo precedono; quelli che vengono dopo sono coperti. Quando poi l'IA scrive davvero, la parola successiva non esiste ancora: lavora nelle stesse condizioni in cui ha imparato.

Nel riassunto grafico: passo 6, la maschera causale.

TAPPA 7

Dividere l'attenzione

I punteggi vengono trasformati in percentuali che, sommate, fanno il cento per cento: è l'**attenzione** che «al» dedica a ciascun pezzo. Nell'esempio del riassunto grafico, «al» dedica circa la metà della sua attenzione a «Domani», un terzo a «vado» e il resto a sé stesso.
Queste percentuali non dicono quali parole sono «importanti» per noi: dicono quanta parte dell'informazione di ciascun pezzo arriverà ad «al».

Nel riassunto grafico: passo 7, la softmax.

TAPPA 8

Mettere insieme le informazioni

«Al» raccoglie i contenuti degli altri pezzi, ognuno nella sua percentuale: molto da «Domani», un po' meno da «vado». Dopo questo passaggio «al» non rappresenta più soltanto sé stesso: porta con sé anche qualcosa di «andare» e di «domani».
È come se ogni pezzo si arricchisse grazie al contesto che lo precede.

Nel riassunto grafico: passo 8, la somma pesata.

TAPPA 9

Più punti di vista insieme

Quando leggiamo una frase possiamo fare attenzione a cose diverse: chi fa l'azione, quando, dove. L'IA fa qualcosa di simile: ripete i collegamenti delle tappe 4–8 più volte in parallelo, ciascuna con regole diverse, poi unisce i risultati.
Una precisazione: nessuno assegna a questi «punti di vista» un compito come «la grammatica» o «il tempo». Le specializzazioni nascono da sole durante l'addestramento, e spesso non sono così nette.

Nel riassunto grafico: passo 9, più teste di attenzione.

TAPPA 10

Rielaborare ogni pezzo

Dopo aver raccolto informazioni dagli altri, ogni pezzo viene rielaborato per conto suo, senza guardare gli altri, con altre regole imparate durante l'addestramento.
Non è una riflessione come la nostra: è un nuovo calcolo, sempre lo stesso per ogni pezzo, che trasforma ciò che il pezzo ha raccolto in una forma più utile per le tappe successive.

Nel riassunto grafico: passo 10, la rete in avanti.

TAPPA 11

Ripetere molte volte

Le tappe dalla 4 alla 10 non si fanno una volta sola: si ripetono, una dopo l'altra, molte volte. Ogni giro si chiama **strato** e ha regole sue. Minerva, il piccolo modello italiano che useremo nel corso, ne ha 16; i modelli più grandi ne hanno decine.

Possiamo immaginare una serie di insegnanti che lavorano uno dopo l'altro sullo stesso compito: ciascuno parte da ciò che ha lasciato il precedente e aggiunge qualcosa. Alla fine del percorso, l'ultimo pezzo, «al», porta con sé ciò che serve per scegliere la parola successiva.

Nel riassunto grafico: passo 11, gli strati.

TAPPA 12

Fare una classifica e scegliere

Ora l'IA dà un voto a ogni pezzo del suo vocabolario, decine di migliaia, e trasforma i voti in probabilità. Ne esce una classifica, per esempio:

1. **mare**
2. lavoro
3. parco
4. cinema

… e così via, fino all'ultimo pezzo del vocabolario

Poi sceglie. Di solito prende uno dei primi della classifica, ma non sempre il primo: molti assistenti estraggono a sorte fra i più probabili, per non scrivere sempre le stesse cose.

`Domani` `vado` `al` `mare`

La parola scelta viene aggiunta alla frase, e tutto il percorso ricomincia da capo per la parola successiva. Una parola alla volta, fino a quando la risposta è finita.

Nel riassunto grafico: passo 12, dai punteggi alla parola successiva.

IN SINTESI

Le tredici tappe

0. ha imparato leggendo moltissimi testi;
1. ritaglia la frase in pezzi e dà a ciascuno un numero;
2. richiama la scheda di ogni pezzo;
3. ricorda la posizione di ogni pezzo;
4. prepara per ogni pezzo una domanda, un'etichetta e un contenuto;
5. cerca i collegamenti con i pezzi precedenti;
6. non guarda avanti;
7. divide l'attenzione;
8. mette insieme le informazioni;
9. osserva da più punti di vista;
10. rielabora ogni pezzo;
11. ripete tutto molte volte;
12. fa una classifica e sceglie, poi ricomincia.

Ciò che a noi sembra comprensione, per l'IA è il risultato di moltissimi calcoli imparati dai testi. Per questo spesso trova la parola giusta, ma non controlla se ciò che scrive è vero: può sbagliare, o inventare, con lo stesso tono sicuro (Modulo 17).

> E ADESSO
>
> Gli stessi passaggi, con i conti fatti a mano su numeri piccoli e inventati: il riassunto grafico e la presentazione animata. Poi, dal Modulo 10, ogni tappa in profondità.