Una frase dentro Minerva

Tutti i passaggi, con i numeri veri, da «Roma è la capitale d'» alla parola successiva.

Nei moduli 10, 11 e 12 hai visto la teoria; nel riassunto grafico hai fatto i conti a mano con numeri inventati. Qui facciamo i conti veri. Diamo a Minerva, il modello del corso, una frase cortissima e la seguiamo dentro il modello, passaggio per passaggio, fino alla parola che sceglie.

Abbiamo scelto la frase per tre motivi. È breve: solo 6 pezzi, così le tabelle restano leggibili. Ha una risposta netta, «Italia», ma anche un'alternativa credibile, «Europa». E contiene una piccola sorpresa: l'apostrofo.

I numeri sono tanti. Quando una fila è troppo lunga mostriamo i primi e scriviamo quanti ne seguono. Ogni numero è stato ricostruito con un programma separato, partendo dai pesi del modello, e confrontato con quello che produce Minerva: la prova è in fondo alla pagina. Tutti i valori sono arrotondati per la lettura: per questo una somma fatta sui numeri stampati può differire di poco da quella calcolata con tutti i decimali.

È il calcolo di un modello preciso, Minerva-350M, su una frase precisa. Altri modelli hanno dimensioni diverse e possono usare operazioni in parte diverse; le idee restano quelle del corso.

Nella pagina distinguiamo due cose. I riquadri grigi mostrano i numeri che il modello calcola: sono dati. I riquadri blu, «La nostra lettura», dicono che cosa ne ricaviamo: sono interpretazioni, e diciamo anche che cosa i numeri non bastano a dimostrare.

I NUMERI CHE INCONTRERAI

Numero	Che cosa rappresenta
7	le posizioni della frase: sei token più il segnale di inizio
1.152	i numeri della fila (il vettore) di ogni token
16	gli strati del modello e, separatamente, le teste di attenzione in ogni strato
72	i numeri di domanda, etichetta e contenuto in ogni testa
4.032	i numeri della parte allargata della rete in avanti
32.768	i token del vocabolario

LA FRASE

«Roma è la capitale d'»

LE DIECI TAPPE

- **1** Frammentazione
- **2** Identificatori
- **3** Embedding
- **4** Attenzione
- **5** Posizione
- **6** Strati
- **7** Punteggi
- **8** Probabilità
- **9** Scelta
- **10** Ripetizione

TAPPA 1 · FRAMMENTAZIONE

La frase si spezza in pezzi

Minerva non riceve direttamente il significato delle parole. Prima il testo viene trasformato in una sequenza di token, cioè pezzi di testo a cui il modello può associare un numero identificativo.

A I caratteri e i byte

La frase ha 21 caratteri, spazi compresi. Nel computer diventano 22 byte: uno per carattere, tranne la «è», che ne occupa due.

B I token

Minerva divide la frase in 6 token e mette davanti un segnale di inizio, <s>. Il puntino «·» indica lo spazio, che fa parte del token. Guarda la fine: «d» e l'apostrofo sono due token separati.

`<s>` `·Roma` `·è` `·la` `·capitale` `·d` `'`

La nostra lettura. Sette pezzi: il segnale di inizio più sei token. Un token non è sempre una parola: qui uno è una sola lettera, un altro è un apostrofo.

PER APPROFONDIRE Nel corso: modulo 10, passi 2–3 · Nel riassunto grafico: tavola 01

TAPPA 2 · IDENTIFICATORI

Ogni pezzo diventa un numero

Ogni token ha un numero nel vocabolario di Minerva, che ha 32.768 voci. Il numero serve solo a ritrovare il token in una tabella, come il numero di catalogo di un libro.

Token	Numero (ID)
`<s>`	1
`·Roma`	2.309
`·è`	413
`·la`	347
`·capitale`	8.309
`·d`	264
`'`	32.379

La nostra lettura. Da qui in poi Minerva non vede più lettere: vede 7 numeri. Il numero non dice nulla del significato: «·Roma» è il 2.309, ma un numero vicino potrebbe essere un pezzo qualsiasi.

PER APPROFONDIRE Nel corso: modulo 10, passo 4 · Nel riassunto grafico: tavola 01

TAPPA 3 · EMBEDDING

Ogni numero diventa una fila di numeri

Il numero di ogni token indica una riga della grande tabella degli embedding. Ogni riga è una fila di 1.152 numeri con la virgola: un vettore. Questi numeri non sono stati assegnati manualmente: il modello li ha appresi automaticamente durante l'addestramento, modificandoli progressivamente in base agli esempi utilizzati.

Token	I primi 8 numeri della sua riga	Lunghezza
`<s>`	`0,0099 0,0118 0,0144 −0,0271 0,0376 −0,0302 0,0137 −0,0287` *… ne seguono altri 1.144*	`0,756`
`·Roma`	`−0,0165 0,0091 −0,0120 −0,0195 −0,0067 −0,0206 0,0040 0,0354` *… ne seguono altri 1.144*	`0,851`
`·è`	`−0,0053 0,0337 −0,0063 −0,0278 −0,0075 0,0091 0,0105 0,0050` *… ne seguono altri 1.144*	`0,742`
`·la`	`−0,0006 −0,0282 0,0114 0,0292 −0,0085 0,0198 −0,0059 0,0503` *… ne seguono altri 1.144*	`0,742`
`·capitale`	`0,0079 −0,0212 −0,0439 −0,0043 −0,0295 0,0244 −0,0557 −0,0010` *… ne seguono altri 1.144*	`0,916`
`·d`	`0,0146 −0,0050 −0,0089 −0,0496 −0,0864 0,0459 0,0023 0,0053` *… ne seguono altri 1.144*	`0,897`
`'`	`0,0131 −0,0092 −0,0203 0,0079 −0,0017 0,0120 −0,0041 0,0033` *… ne seguono altri 1.144*	`0,738`

Valori reali di Minerva. La colonna «Lunghezza» misura quanto è «grande» il vettore nel suo insieme (la radice della somma dei quadrati dei suoi numeri), non quanti numeri contiene: quelli sono sempre 1.152.

La nostra lettura. La frase è diventata 7 file da 1.152 numeri: 8.064 numeri in tutto. Ogni riga è la stessa in qualunque frase: «·Roma» ha questi numeri ovunque compaia. Il contesto non c'è ancora.

PER APPROFONDIRE Nel corso: modulo 10, passi 5–7 · Nel riassunto grafico: tavola 02

TAPPA 4 · ATTENZIONE

I token si guardano: l'attenzione

Stiamo cercando di capire come l'ultimo token, l'apostrofo, possa usare le informazioni dei token precedenti per preparare la previsione della parola successiva. Seguiamo l'ultimo token, l'apostrofo, nel primo dei 16 strati, e una delle sue 16 teste di attenzione, la 12. È l'ultimo token che conta, perché è lui che dovrà dire che cosa viene dopo.

A Si regola la scala dei numeri

Prima di tutto, i numeri che rappresentano l'apostrofo vengono normalizzati: ciascuno viene diviso per 0,02198, un valore calcolato facendo la radice quadrata della media dei quadrati di tutti i numeri del vettore. Poi ogni numero viene moltiplicato per un coefficiente appreso automaticamente durante l'addestramento del modello. Queste operazioni regolano la grandezza dei numeri, rendendoli più adatti ai calcoli successivi.

	I primi 6 numeri	
prima	`0,0131 −0,0092 −0,0203 0,0079 −0,0017 0,0120`	*… ne seguono altri 1.146*
dopo	`0,5154 −0,3694 −0,7996 0,2919 −0,0644 0,4614`	*… ne seguono altri 1.146*

B Si calcolano domanda, etichetta e contenuto

Utilizzando tre tabelle di numeri appresi durante l'addestramento, il modello ricava dal vettore dell'apostrofo tre nuove serie di 72 numeri: la domanda (che cosa cerco), l'etichetta (come mi presento) e il contenuto (che cosa consegno). Ogni nuovo numero si ottiene moltiplicando a coppie i valori del vettore e quelli della tabella, e sommando i risultati. Ecco il primo numero della domanda:

Numero del vettore	× peso appreso	= prodotto
0,5154	−0,01276	−0,006575
−0,3694	0,00601	−0,002221
−0,7996	0,01587	−0,012689
0,2919	−0,00687	−0,002005
−0,0644	0,02356	−0,001518
… ne seguono altri 1.147		

Somma di tutti i 1.152 prodotti: **1,11842**. È il primo numero della domanda.

	I primi 6 numeri di 72
domanda	1,1184 0,6907 0,8175 −2,7845 0,2639 −0,2298 *… ne seguono altri 66*
etichetta	0,2170 0,3147 0,2449 −0,3927 0,1065 −0,0854 *… ne seguono altri 66*
contenuto	−0,1945 −0,0674 −0,0037 −0,0465 0,0503 −0,0810 *… ne seguono altri 66*

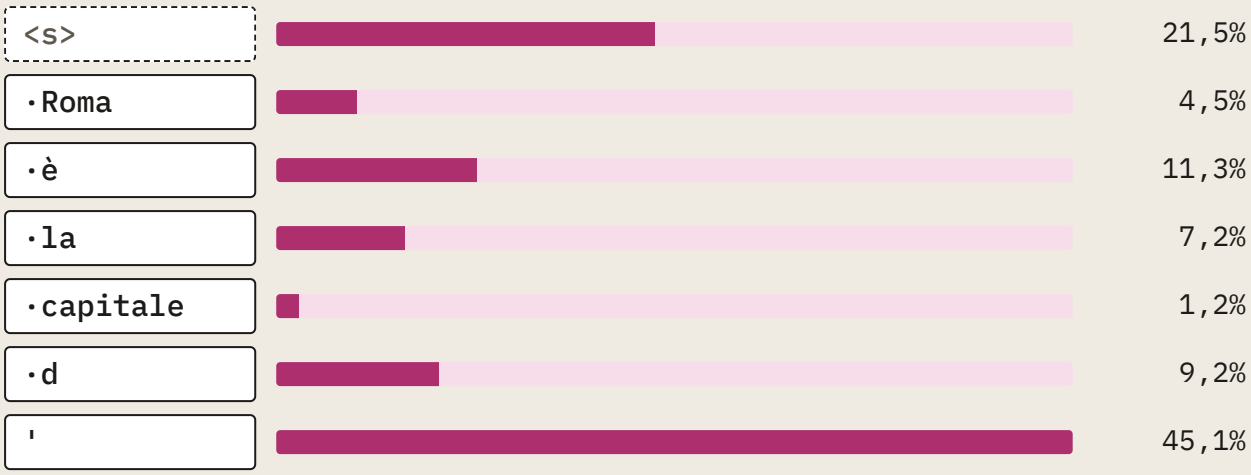
C La domanda si confronta con le etichette

La domanda dell'apostrofo si confronta con l'etichetta di ogni token, sé stesso compreso: si moltiplicano i numeri a coppie e si sommano (prodotto scalare). Poi si divide per √72, circa 8,49, per tenere i punteggi su una scala ragionevole. Questi punteggi tengono già conto della posizione: lo vedrai nella tappa 5.

Token guardato	Prodotto scalare	÷ 8,49 = punteggio
<s>	51,02	6,013
·Roma	37,78	4,452
·è	45,58	5,372
·la	41,76	4,921
·capitale	26,69	3,146
·d	43,79	5,161
'	57,32	6,755

D I punteggi diventano percentuali

La softmax trasforma i punteggi in percentuali che sommano al 100%: sono i pesi di attenzione. Non sono probabilità della parola successiva, e nemmeno una misura diretta dell'importanza linguistica delle parole: sono pesi matematici, che dicono quanto il contenuto di ogni token entra nel calcolo di questa testa.



L'apostrofo, l'ultimo token, può guardare tutti: per lui la maschera non toglie nulla. Per «·la», a metà frase, la stessa testa dà invece lo 0% a «·capitale», «·d» e all'apostrofo, che vengono dopo: <s> 49,2%, ·Roma 10,9%, ·è 20,0%, ·la 19,9%, ·capitale 0,0%, ·d 0,0%, ' 0,0%.

E Si raccolgono i contenuti

Ora ogni token consegna il suo contenuto, in proporzione al peso che ha ricevuto, e i contributi si sommano: è il messaggio di questa testa per l'apostrofo.

	I primi 6 numeri di 72	
messaggio	`−0,0771 −0,0495 −0,0095 −0,0925 0,0559 −0,0582`	*... ne seguono altri 66*

F Le altre 15 teste fanno lo stesso

Ogni testa ha le sue tabelle e quindi la sua domanda: ciascuna distribuisce l'attenzione a modo suo. Ecco i pesi dell'apostrofo in tutte le 16 teste dello strato 1.

Testa	<s>	·Roma	·è	·la	·capitale	·d	'
1	33	9	12	13	6	13	14
2	37	6	10	10	4	18	14
3	29	10	12	12	6	14	17
4	32	9	12	9	5	16	17
5	34	6	7	23	3	8	20
6	44	6	10	13	4	9	13
7	45	5	9	13	3	9	16
8	53	5	8	11	3	6	13
9	34	3	11	9	2	8	33
10	41	4	8	10	2	12	25
11	41	4	8	11	2	14	19
12	21	5	11	7	1	9	45
13	27	6	11	13	3	12	27
14	12	4	15	16	2	14	37
15	21	6	12	25	3	19	16
16	32	5	9	16	2	14	21

Percentuali arrotondate. In questo primo strato molte teste assegnano la quota maggiore al segnale di inizio. È un comportamento che può essere interpretato come una forma di «parcheggio» dell'attenzione, già illustrata nel modulo 11.

G I messaggi si uniscono e si sommano alla fila

I 16 messaggi, ciascuno formato da 72 numeri, vengono uniti in un vettore di 1.152 numeri (16 × 72). Una nuova tabella di pesi appresi durante l'addestramento li trasforma ulteriormente. Il risultato si somma alla fila di partenza dell'apostrofo, che non viene buttata via: è la connessione residua.

	I primi 6 numeri	Lunghezza
fila di partenza	`0,0131 −0,0092 −0,0203 0,0079 −0,0017 0,0120`	0,738
+ risultato dell'attenzione	`−0,0121 0,0302 0,0320 −0,0037 −0,0073 0,0047`	2,768
= nuova fila	`0,0010 0,0210 0,0117 0,0042 −0,0090 0,0167`	2,911

La nostra lettura. L'apostrofo ha guardato i token che lo precedono e ha raccolto un po' del loro contenuto. Nella testa 12 ha ascoltato soprattutto sé stesso (45,1%) e il segnale di inizio (21,5%): nel primo strato è presto per collegare l'apostrofo a «Roma».

PER APPROFONDIRE Nel corso: modulo 11, passi 2–8 · Nel riassunto grafico: tavola 05, tavola 06, tavola 07, tavola 08

TAPPA 5 · POSIZIONE

Conta anche la posizione

Torniamo per un momento indietro. Prima di calcolare i punteggi di attenzione che abbiamo appena visto, Minerva modifica domanda ed etichetta per rappresentare la posizione dei token: le ruota di un angolo che dipende dal posto del token. I 72 numeri sono divisi in 36 coppie: ogni coppia è come una lancetta che gira.

A Una lancetta della domanda dell'apostrofo

L'apostrofo è al posto 6 (si conta da 0: il segnale di inizio è al posto 0). In Minerva le coppie non sono formate da numeri vicini: il 1º numero va in coppia con il 37º, il 2º con il 38º, e così via. Per questo la coppia qui sotto contiene il primo numero della domanda (1,1184, lo stesso della tappa 4) e il 37º, che nella tappa 4 non era mostrato. È la coppia che ruota più velocemente: per ogni posizione avanza di 1 radiante, cioè circa 57,3 gradi. Alla posizione 6 la rotazione complessiva è di 6 radianti, circa 343,8 gradi. Altre coppie ruotano molto più lentamente: la più lenta di circa 0,0074 gradi per posizione.

Coppia di numeri	Prima della rotazione	Dopo la rotazione
1º e 37º numero	`(1,1184; 0,4575)`	`(1,2017; 0,1267)`
... ne seguono altre 35		

La nostra lettura. La rotazione introduce nel confronto tra domanda ed etichetta un'informazione sulla posizione relativa dei token, cioè su quanti posti li separano. La rotazione era già dentro i punteggi della tappa 4: l'abbiamo spiegata dopo, come nel corso, per non complicare tutto insieme.

PER APPROFONDIRE Nel corso: modulo 11, passi 9–11 · Nel riassunto grafico: tavola 04, tavola 06

TAPPA 6 · STRATI

Strato dopo strato

Uno strato non fa solo attenzione. Dopo, ogni token viene rielaborato da solo, senza guardare gli altri, da una rete in avanti. Poi tutto si ripete nello strato successivo, con tabelle diverse: in Minerva gli strati sono 16.

A La rete in avanti dello strato 1

I numeri che rappresentano l'apostrofo vengono nuovamente normalizzati. La rete in avanti li trasforma prima in 4.032 numeri, applica ulteriori calcoli che modificano il contributo di ciascun valore e infine li riporta a 1.152. Il risultato viene sommato al vettore precedente, come nell'attenzione.

	I primi 6 numeri	Lunghezza
fila dopo l'attenzione	`0,0010 0,0210 0,0117 0,0042 −0,0090 0,0167`	`2,911`
+ risultato della rete in avanti	`0,1190 0,1776 −0,1099 −0,0881 −0,1429 −0,1523`	`8,540`
= uscita dello strato 1	`0,1199 0,1986 −0,0982 −0,0839 −0,1519 −0,1356` *… ne seguono altri 1.146*	`10,366`

Nella parte allargata, 808 numeri su 4.032 superano 0,05 in valore assoluto. Mostra che alcuni valori sono molto più grandi di altri, ma non vuol dire che gli altri siano inattivi: la rete non funziona come una serie di interruttori accesi o spenti.

B Ne seguono ancora 15 strati

Gli altri 15 strati ripetono gli stessi calcoli. Non li mostriamo numero per numero: ecco che cosa succede all'apostrofo, strato dopo strato. L'ultima colonna mostra un esperimento: prendiamo il vettore prodotto da ciascuno strato, applichiamo la normalizzazione finale e la tabella di uscita, e osserviamo quale token riceverebbe la probabilità maggiore se il calcolo terminasse in quel punto. È la «lente sui logit» del modulo 12, una tecnica di analisi e una lettura approssimata: il modello, normalmente, non fa una previsione a ogni strato.

Strato	Lunghezza della fila	Testa che guarda di più «Roma»	Se si fermasse qui
1	10,4	testa 3 · 9,7%	`s` 3%
2	15,2	testa 6 · 12,0%	`s` 11%
3	18,4	testa 14 · 6,1%	`·behavior` 8%
4	22,9	testa 9 · 6,9%	`s` 6%
5	29,3	testa 11 · 4,1%	`app` 10%
6	36,9	testa 14 · 7,8%	`obbligo` 10%
7	37,9	testa 4 · 2,2%	`acc` 13%
8	45,3	testa 10 · 6,2%	`origine` 15%
9	52,1	testa 13 · 7,3%	`origine` 12%
10	61,5	testa 12 · 7,0%	`Italia` 38%
11	68,5	testa 9 · 45,7%	`Italia` 59%
12	79,3	testa 2 · 13,7%	`Italia` 92%
13	87,3	testa 8 · 28,8%	`Italia` 91%
14	101,0	testa 8 · 14,9%	`Italia` 84%
15	118,7	testa 1 · 7,5%	`Italia` 79%
16	138,7	testa 13 · 13,4%	`Italia` 84%

Valori reali per l'apostrofo. Nei primi strati la sbirciata dà pezzi senza senso; dallo strato 10 compare «Italia». Nello strato 11 la testa 9 dà all'apostrofo il 45,7% della sua attenzione verso «Roma». La lunghezza crescente della fila non è una misura della conoscenza raccolta.

La nostra lettura. Strato dopo strato la fila dell'apostrofo cambia, finché la tabella finale vi legge «Italia». Nello strato 11 una testa assegna un peso particolarmente alto a «Roma», e nello stesso strato la previsione di «Italia» diventa più forte. È un indizio interessante del modo in cui il modello mette in relazione le parole, ma non basta, da solo, a dimostrare quali calcoli abbiano determinato la risposta: un peso alto dice quanto un contenuto entra nel calcolo di una testa, non che ne sia la causa.

PER APPROFONDIRE Nel corso: modulo 11, passi 12–13, modulo 12, passo 3 · Nel riassunto grafico: tavola 09

TAPPA 7 · PUNTEGGI

Un voto per ogni token del vocabolario

Finiti i 16 strati, la fila dell'apostrofo viene normalizzata un'ultima volta e confrontata con le 32.768 righe di una seconda tabella, una per ogni token che Minerva conosce. Ogni confronto dà un punteggio, il logit.

A Il punteggio di «Italia»

Si moltiplicano a coppie i 1.152 numeri della fila con quelli della riga di «Italia» e si sommano.

	I primi 6 numeri	
fila finale dell'apostrofo	`0,0116 −2,4033 −2,2718 0,8319 0,3164 −1,0204`	*… ne seguono altri 1.146*
riga di «Italia»	`−0,0410 −0,0378 −0,0135 0,0894 −0,0125 −0,0608`	*… ne seguono altri 1.146*

Somma dei 1.152 prodotti: **22,304**. È il punteggio di «Italia».

B Tutti i 32.768 punteggi

Lo stesso calcolo, ripetuto per ogni riga. In testa alla classifica:

Token	Punteggio
`Italia`	22,30
`Europa`	20,56
`italia`	16,65
`·Italia`	16,10
`arte`	15,26
`O`	14,64
… ne seguono altri 32.762	
`zionati`	−12,09
`rows`	−12,53
`zionato`	−13,01

In fondo, gli ultimi tre. Il punteggio medio è −2,64.

La nostra lettura. Minerva non ha ancora scelto: ha dato un voto a tutti i 32.768 token. «Italia» è primo, «Europa» secondo, staccato di 1,75 punti.

PER APPROFONDIRE Nel corso: modulo 12, passi 1–2 · Nel riassunto grafico: tavola 10

TAPPA 8 · PROBABILITÀ

I voti diventano probabilità

La softmax, la stessa formula dell'attenzione, trasforma i 32.768 punteggi in probabilità che sommano al 100%. Qui però le percentuali hanno un altro scopo: indicano la probabilità che il modello assegna a ciascun token come prossimo elemento del testo.



«Italia» e «Europa» insieme fanno il 99,1%. Gli altri 32.766 token si dividono le briciole.

La sorpresa dell'apostrofo

Nella classifica c'è due volte «Italia»: «Italia» attaccata, al 84,4%, e «·Italia» con lo spazio davanti, appena al 0,17%. Per Minerva sono due token diversi, con due numeri diversi (3.065 e 1.796). Dopo un apostrofo non si mette lo spazio: il modello ha appreso anche regolarità di scrittura come questa durante l'addestramento.

> **La nostra lettura.** Minerva assegna un'alta probabilità a «Italia» nella propria distribuzione: l'incertezza misurata è di appena 0,71 bit. È una misura interna al modello, non una verifica indipendente della correttezza: un modello può dare un'alta probabilità anche a una risposta sbagliata.

PER APPROFONDIRE Nel corso: modulo 12, passi 4–5 · Nel riassunto grafico: tavola 10

TAPPA 9 · SCELTA

Si sceglie il token

Il modo più semplice è prendere il primo della classifica, la scelta golosa: «Italia». Con l'estrazione a sorte, invece, «Europa» uscirebbe in media circa 15 volte su 100.

Italia scelta golosa: il primo della classifica, con il 84,4%

La nostra lettura. La scelta è un passaggio separato dalle probabilità: con le stesse probabilità, modi diversi di scegliere possono dare parole diverse.

PER APPROFONDIRE Nel corso: modulo 12, passi 6–8 · Nel riassunto grafico: tavola 10

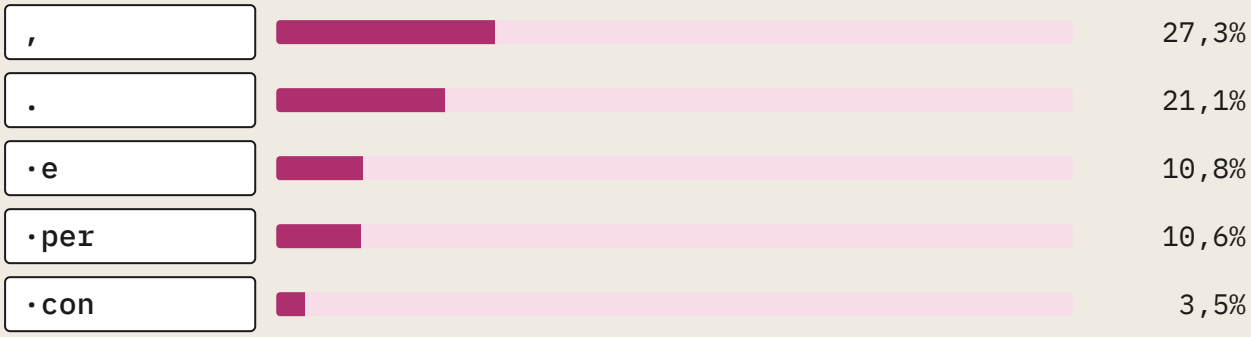
TAPPA 10 · RIPETIZIONE

Si aggiunge e si ricomincia

Il token «Italia» viene aggiunto alla sequenza. Il modello utilizza quindi il contesto aggiornato per calcolare le probabilità del token successivo. La cache conserva le etichette (key) e i contenuti (value) già calcolati per le posizioni precedenti, in ogni strato: così il modello non deve ripetere tutti quei calcoli, e calcola solo il token nuovo.

A Il secondo giro

Dopo «Roma è la capitale d'Italia», ecco i candidati per il token successivo:



Qui la probabilità è molto meno concentrata: la virgola e il punto sono quasi alla pari.

B Dodici giri

Ripetendo la scelta golosa per 12 giri, Minerva scrive:

`·Roma` `·è` `·la` `·capitale` `·d` `'` `Italia` `,` `·ma` `·anche` `·la` `·capitale` `·del` `·mondo` `.` `⏎` `La` `·città`

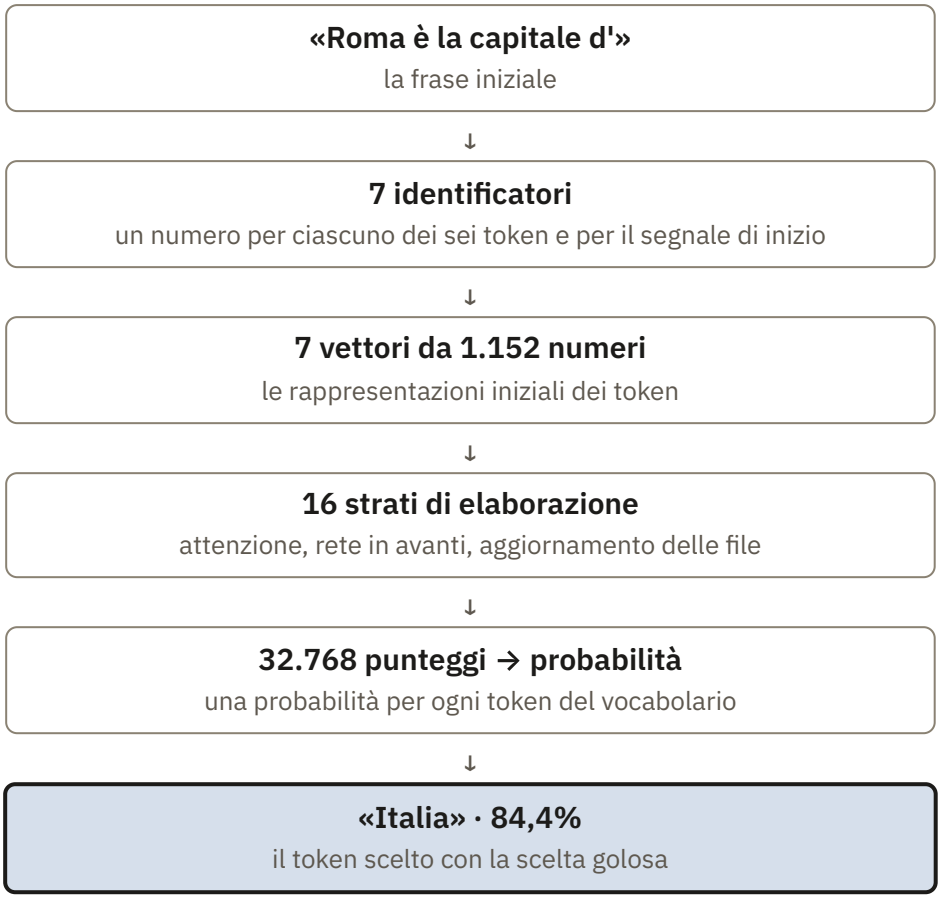
«*Roma è la capitale d'Italia, ma anche la capitale del mondo. ⏎La città*»

La nostra lettura. Ogni pezzo è stato scelto con le dieci tappe che hai visto. La frase finale unisce due cose diverse. «Roma è la capitale d'Italia» è un'affermazione verificabile. «La capitale del mondo» è una formula retorica, con radici storiche e culturali (Roma caput mundi), che non si verifica allo stesso modo. Il modello produce entrambe le espressioni con la stessa scorrevolezza, perché genera il testo sulla base delle regolarità apprese. Questa capacità, però, non garantisce che sappia distinguere correttamente un'affermazione verificabile da una formula retorica.

PER APPROFONDIRE Nel corso: modulo 12, passi 9–10, modulo 13, passo 11 · Nel riassunto grafico: tavola 11

RICAPITOLANDO

Come cambia la forma dei dati



IL VIAGGIO INTERO, IN MOVIMENTO

SPIEGAZIONE ANIMATA

Dalla frase alla risposta

Tutte le tappe, dalla frase al token scelto, una scena alla volta: i conti si fanno davanti ai tuoi occhi.

cultor.net/corso-ia/percorso-animato

LABORATORIO · 142 FRASI

Passo passo dentro Minerva

Scegli una frase tra 142, dai proverbi ai numeri alle altre lingue, e seguila in 15 passi con i numeri veri del modello.

cultor.net/corso-ia/passo-passo

DA DOVE VENGONO I NUMERI

La prova del calcolo

Tutti i numeri presentati in questo documento provengono dai calcoli eseguiti con Minerva-350M (pesi originali, 16 strati, 1.152 numeri per vettore, 16 teste, 32.768 token nel vocabolario). Non li abbiamo presi dal modello già pronti. Abbiamo ricostruito separatamente, con un programma di calcolo, le operazioni matematiche che Minerva esegue, partendo dai suoi pesi: normalizzazione, domanda, etichetta e contenuto, rotazione, punteggi, softmax, messaggi, rete in avanti, punteggi finali.

Poi abbiamo confrontato i risultati con quelli ottenuti dall'esecuzione originale del modello. La differenza più grande, su tutti i passaggi e tutti gli strati, è di **`0,00005`**: sono solo gli arrotondamenti dei calcoli del computer. Anche il secondo giro calcolato con la cache dà gli stessi numeri di quello calcolato da capo. Vuol dire che le operazioni descritte nel corso riproducono fedelmente il calcolo del modello. Le immagini usate per spiegarle (la biblioteca, le lancette) restano invece semplificazioni didattiche.

Chi vuole ripetere la verifica trova il programma accanto a questa pagina, nel file verifica_minerva.py: con i pesi pubblici di Minerva-350M produce tutti i numeri della pagina e la differenza massima.

Calcolo: PyTorch 2.14.1, Transformers 5.21.0, numeri a 32 bit; i valori sono arrotondati per la lettura. Modello: Minerva-350M-base-v1.0 (Sapienza NLP, 2024). Scelta golosa nei giri successivi.