Dimensione Cognitiva
5. Come le macchine comprendono il linguaggio
Il Meccanismo di Attenzione

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

A lezione di Intelligenza Artificiale

Siena - Giugno 2025

Indice

Il Meccanismo di "Attenzione"

Attenzione Selettiva

- L'**attenzione selettiva** è un meccanismo cognitivo fondamentale che consente agli esseri umani di **concentrare le risorse mentali** su un'informazione rilevante e **ignorare gli stimoli irrilevanti** presenti nell'ambiente.
- È ciò che ci permette, per esempio, di ascoltare una conversazione specifica in una stanza affollata (fenomeno noto come *cocktail party effect*), oppure di leggere un libro ignorando i rumori esterni.

Attenzione Selettiva

Dal punto di vista neurocognitivo:

- L'attenzione selettiva **filtra** l'informazione a livello sensoriale e/o percettivo.
- È **limitata**: non possiamo processare coscientemente tutto ciò che ci circonda, perciò il cervello seleziona ciò che ritiene più rilevante.
- È **dinamica**: può essere guidata da stimoli esterni (attenzione esogena) o da obiettivi interni (attenzione endogena).
- È spesso **modulata da contesto, esperienza e aspettative**.

Attenzione Selettiva

Il meccanismo di attenzione nei **modelli di deep learning**, come i transformer introdotti nel celebre paper *"Attention Is All You Need"*, si ispira (astrattamente) all'idea di attenzione umana. In particolare:

- Entrambi i sistemi **danno più "peso" alle informazioni rilevanti** rispetto a quelle irrilevanti.
- Entrambi **distribuiscono le risorse di elaborazione** in maniera non uniforme.
- L'attenzione nei transformer **filtra e valorizza dinamicamente** certi input rispetto ad altri in base al contesto, proprio come fa il cervello quando decide a cosa prestare attenzione.

Attention is all you need

Embeddings Contestuali

- 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.**

Embeddings Contestuali

- 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” (personaggio del film)?
- Anche se non c’è una risposta precisa, è molto probabile che, durante l’addestramento, la parola “arancia” sia stata principalmente associata al frutto, piuttosto che ad un film (ad esempio *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”.

Embeddings Contestuali

Ma in una frase come: **“Il film Joker mi ricorda molto 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.

Embeddings Contestuali

- 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.

Embeddings Contestuali

- 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 **"Il film Joker mi ricorda molto Arancia Meccanica"**
- oppure verso "limone" (frutto) nel caso di **"Arancia e limone contengono molta vitamica C"**

Attenzione Selettiva

- 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"*, focalizzato sulla traduzione automatica.

Attenzione Selettiva

Solo cinque o sei 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.
Mi piace questo libro → J'**aime** ce livre
Ti piace questo libro → tu **aimes** ce livre
Ci piace questo libro → nous **aimons** ce livre
...

Mi piace questo mare → J'aime **cette** mer

Attenzione Selettiva

- Per tradurre correttamente la parola "piace" in francese, il modello di traduzione ha bisogno di capire che 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!
- Il resto della frase è apparentemente inutile... attenzione selettiva!

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

Attenzione Selettiva

L'ANIMALE NON HA ATTRAVERSATO LA STRADA PERCHÉ **ERA** *TROPPO STANCO*

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

Attenzione Selettiva

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!

L'ANIMALE NON HA ATTRAVERSATO LA STRADA PERCHÉ **ERA** *TROPPO TRAFFICATA*

Attenzione Selettiva

- 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!
- Riprendiamo in esame le due frasi precedenti:
  “Il film Joker mi ricorda molto Arancia Meccanica”
  “Arancia e limone contengono molta vitamica C”
- Mappiamo tutte le parole con un *embedding* di tre dimensioni, "fruttosità", "filmosità", e una terza dimensione fittizia a caso.

Attenzione Selettiva

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)

Attenzione Selettiva

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	0.9
e	0	0	1	0
Joker	0.5	0.9	0	1

Attenzione Selettiva

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

Attenzione Selettiva

- 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".

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

Attenzione Selettiva

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;

Attenzione Selettiva

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:

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;

Attenzione Selettiva

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:

frequenza d'uso: alcune parole sono molto comuni ('è', '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.

Transformer

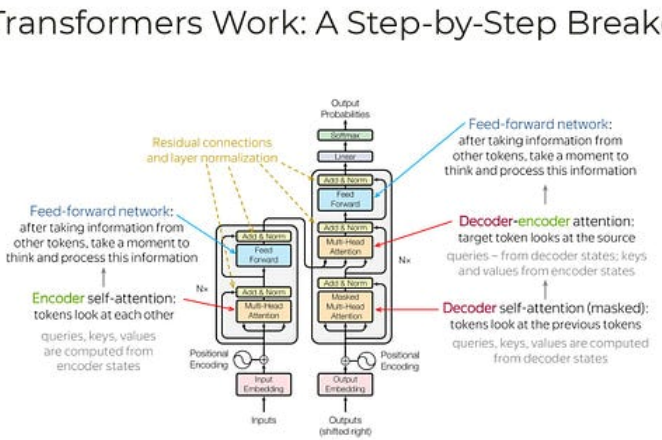
Transformer

- 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**.

Transformer

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**.


How Transformers Work: A Step-by-Step Breakdown
Output Probabilities
Softmax
Linear
Residual connections and layer normalization
Feed-forward network: after taking information from other tokens, take a moment to think and process this information
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Feed-forward network: after taking information from other tokens, take a moment to think and process this information
Decoder-encoder attention: target token looks at the source
queries – from decoder states; keys and values from encoder states
Encoder self-attention: tokens look at each other
queries, keys, values are computed from encoder states
Decoder self-attention (masked): tokens look at the previous tokens
queries, keys, values are computed from decoder states
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Input Sequence
A sequence of words or tokens that the model will process.

Transformer

- 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.

Transformer

- 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.

Transformer

In Conclusione...

Geometria e Probabilità

- Lo spazio dei dati è come un territorio immenso
- Immaginate tutte le possibili immagini, testi, musiche o video che potrebbero mai esistere come punti in un territorio multidimensionale gigantesco.
- Ogni contenuto reale occupa una piccola zona di questo territorio - per esempio, tutte le foto di gatti stanno raggruppate in una "regione gatti", tutte le sinfonie di Mozart in una "regione Mozart", e così via.

Geometria e Probabilità

Esempio pratico:Quando ChatGPT scrive un testo, non assembla parole a caso.

Ha imparato che certe combinazioni di concetti "vivono" in zone specifiche dello spazio dei dati.

Quando gli chiedi di scrivere una poesia, si muove nella "regione poesia" e crea un nuovo punto che rispetta le regole di quella zona.

L'AI generativa è quindi un sistema che ha imparato la "geografia" dello spazio dei dati e può creare nuovi contenuti esplorando intelligentemente le zone più promettenti.

Geometria e Probabilità

Conoscere la "densità di popolazione": Qui entra la distribuzione di probabilità - l'AI impara non solo dove si trovano i contenuti validi, ma anche quanto sono comuni in ogni zona

Creare nuovi punti: Genera contenuti posizionandoli strategicamente nelle zone più "popolate" e probabili

Geometria e Probabilità

- Quando ChatGPT scrive, non solo sa che certe parole vanno insieme, ma sa anche quanto spesso vanno insieme. Sa che dopo "buongiorno" è molto probabile trovare "come va?" piuttosto che "elefante viola". L'AI ha imparato questa "mappa delle probabilità" dai dati di training.
- In sostanza, l'AI generativa crea contenuti che rispettano anche le "regole statistiche" di quanto certi tipi di contenuto sono comuni o rari nella realtà.

L'IA generativa segue un percorso in questo spazio astratto secondo regole statistiche e di ottimizzazione, non possiede alcun concetto di ≪verità≫!!!


CHATGPT È IN GRADO DI SOSTENERE UN COLLOQUIO CON UN INTERLOCUTORE UMANO
SÌ, MA OGNI TANTO SBAGLIA LE RISPOSTE...