

Introduzione

Flavio Cordari

Ruolo

Ingegnere Informatico – Roma

Formazione

PhD in Earth Observation – La Sapienza (*in corso*)
Laurea Magistrale in Ing. Informatica

Esperienza attuale

Research Fellow @ CESBIO (Centre)
Research Fellow @ E-geos / La Sapienza

Competenze

Geoinformation & Earth Observation · Web Development · AI · Cybersecurity

Programma di Oggi

- Natural Language Processing
- Deep Learning
- Large Language Models
- Embedding
- Attention
- Multi-Layer Perceptron
- Sviluppi Recenti

Definizione

Un **Large Language Model** (LLM) è un *tipo di algoritmo di intelligenza artificiale (IA)* che utilizza tecniche di deep learning (DL) e natural language processing (NLP) su enormi insiemi di dati per comprendere, riassumere, generare e prevedere nuovi contenuti.

Classificazione

IA

A large, light red circle is centered on the page. At the top of the circle, the letters 'IA' are written in a red, sans-serif font.

Natural Language Processing

Definizione

Il **Natural Language Processing** (NLP) è il campo di studio che si concentra sull'**interazione tra computer e esseri umani** attraverso il *linguaggio naturale*.

ELIZA

Utente: Sono infelice.

ELIZA: PENSI CHE VENIRE QUI TI AIUTERA' A NON ESSERE INFELICE

Utente: Ho bisogno di aiuto, questo sembra certo.

ELIZA: COSA SIGNIFICHEREBBE PER TE SE OTTENESSI AIUTO

Utente: Forse potrei imparare ad andare d'accordo con mia madre.

ELIZA: DIMMI DI PIU' DELLA TUA FAMIGLIA

Utente: Mia madre si prende cura di me.

ELIZA: CHI ALTRO NELLA TUA FAMIGLIA SI PRENDE CURA DI TE

Utente: Mio padre.

ELIZA: TUO PADRE

Utente: Lei è come mio padre in qualche modo.

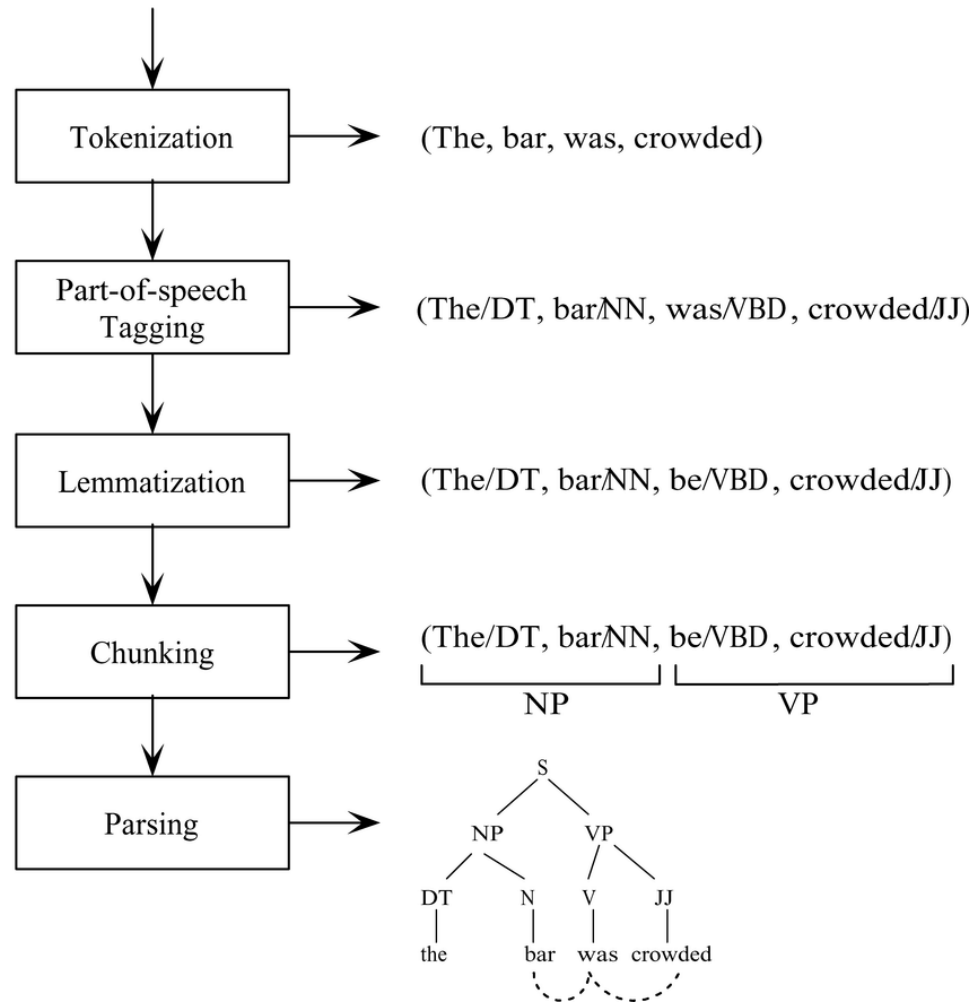
Weizenbaum, J. 1966. ELIZA - Un programma informatico per lo studio della comunicazione in linguaggio naturale tra uomo e macchina. CACM, 9(1):36-45.

Pipeline NLP

- 1 **Raccolta Dati:** Raccogliere dati testuali grezzi.
- 2 **Text Preprocessing:** Pulizia e preparazione dei dati testuali (Tokenizzazione, stemming, lemmatizzazione, rimozione delle stop word, ...)
- 3 **Feature Extraction:** Conversione del testo in un formato comprensibile dagli algoritmi di machine learning
- 4 **Modellazione:** Applicazione di algoritmi per apprendere dai dati.
- 5 **Valutazione:** Valutazione delle prestazioni del modello.
- 6 **Distribuzione:** Integrazione del modello nelle applicazioni.

Text Preprocessing

The bar was crowded



Feature Extraction

Technique	Main Features	Use Cases	Size and Complexity
CountVectorizer	Converts text to matrix of word counts	Text classification, topic modeling	Simple and fast, suitable for small to medium-sized datasets
TF-IDF	Assigns weights to words based on importance	Information retrieval, text classification	More complex and computationally expensive, suitable for medium to large-sized datasets
Word embeddings	Vector representation of words based on semantics and syntax	Text classification, information retrieval	Can handle large datasets, computationally expensive to train
Bag of words	Represents text as a vector of word frequencies	Text classification, sentiment analysis	Simple and fast, suitable for small to medium-sized datasets
Bag of n-grams	Captures frequency of sequences of n words	Text classification, sentiment analysis	Size and complexity depend on the size of the n-grams and the dataset
Hashing Vectorizer	Maps words to fixed-size feature space using hashing function	Large-scale text classification, online learning	Suitable for large datasets, memory-efficient, may suffer from hash collisions
Latent Dirichlet Allocation (LDA)	Identifies topics in corpus and assigns probability distribution to each document	Topic modeling, content analysis	Suitable for medium to large-sized datasets, computationally expensive
Non-negative matrix factorization (NMF)	Decomposes document-term matrix into lower-dimensional parts	Topic modeling, content analysis	Suitable for medium-sized datasets, computationally expensive
Principal component analysis (PCA)	Reduces dimensionality of document-term matrix	Text visualization, text compression	Suitable for large datasets, computationally expensive
Part-of-speech (POS) tagging	Assigns part of speech tag to each word in text	Named entity recognition, text classification	Requires additional processing, suitable for small to medium-sized datasets

Exploring Feature Extraction Techniques for Natural Language Processing

Cos'è un Modello Linguistico?

Un modello linguistico è un *algoritmo statistico e computazionale* che consente a un computer di comprendere, interpretare e generare il linguaggio umano basandosi sulla probabilità di occorrenza delle parole e delle sequenze di parole.

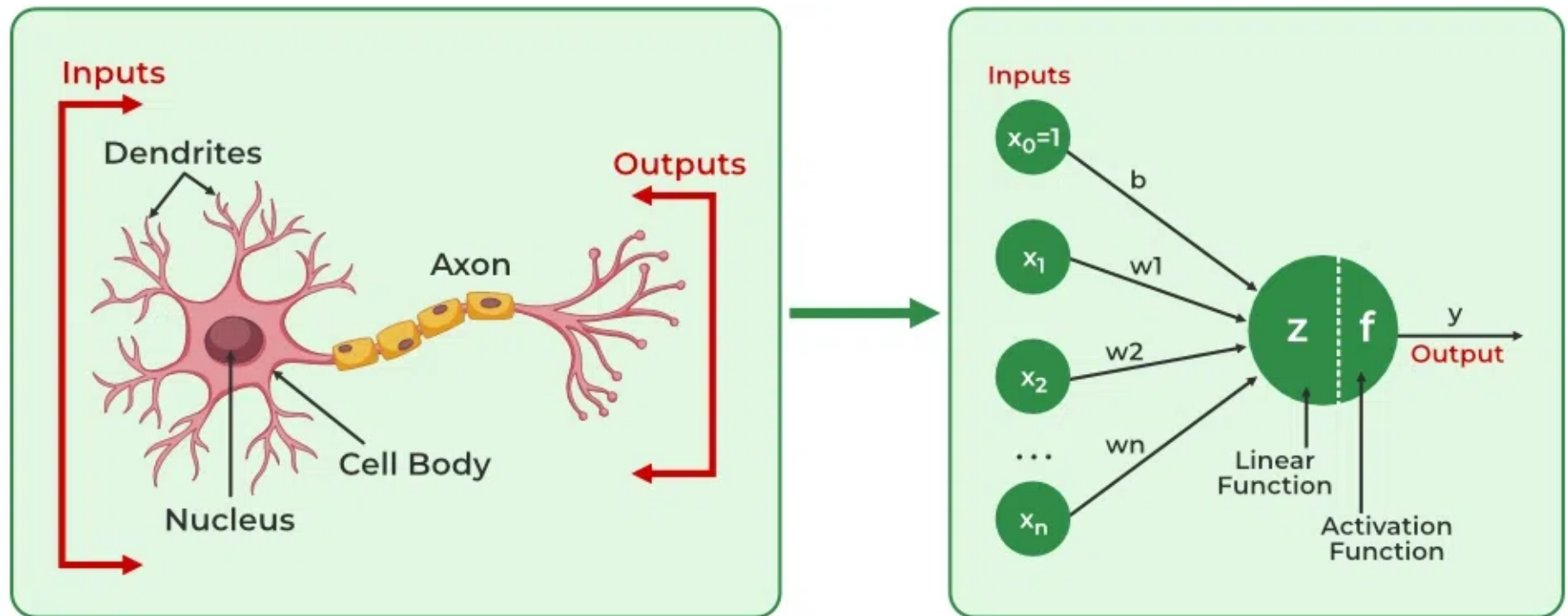
Deep Learning

Definizione

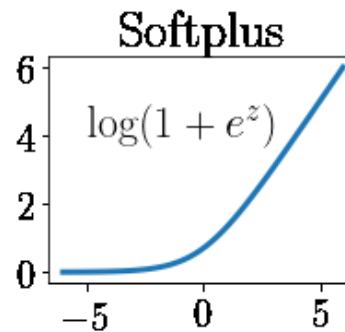
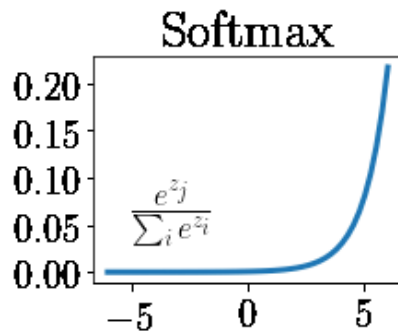
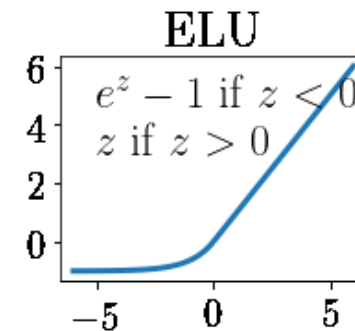
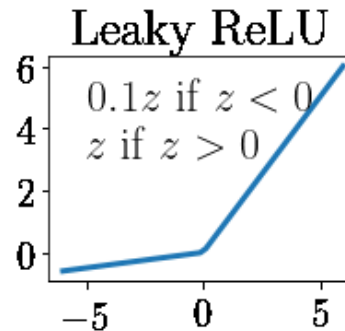
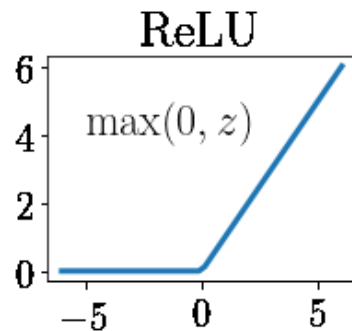
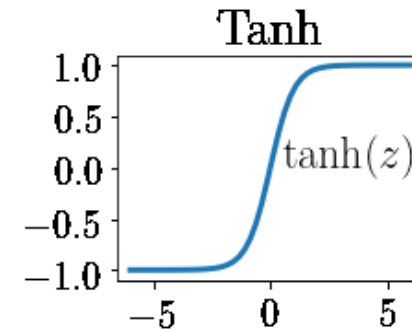
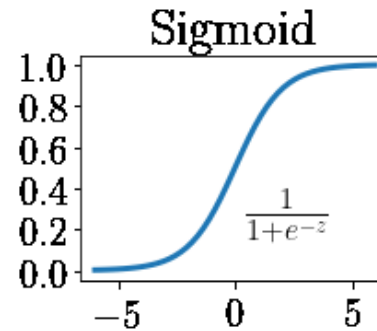
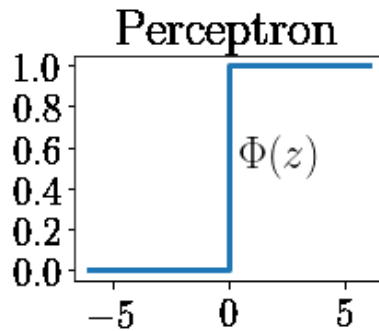
Il **Deep Learning** è un sottoinsieme del machine learning nell'intelligenza artificiale (IA) che *imita il funzionamento del cervello umano* nell'elaborazione dei dati per il rilevamento di oggetti, il riconoscimento vocale, la traduzione di lingue e il processo decisionale.

Reti Neurali Artificiali

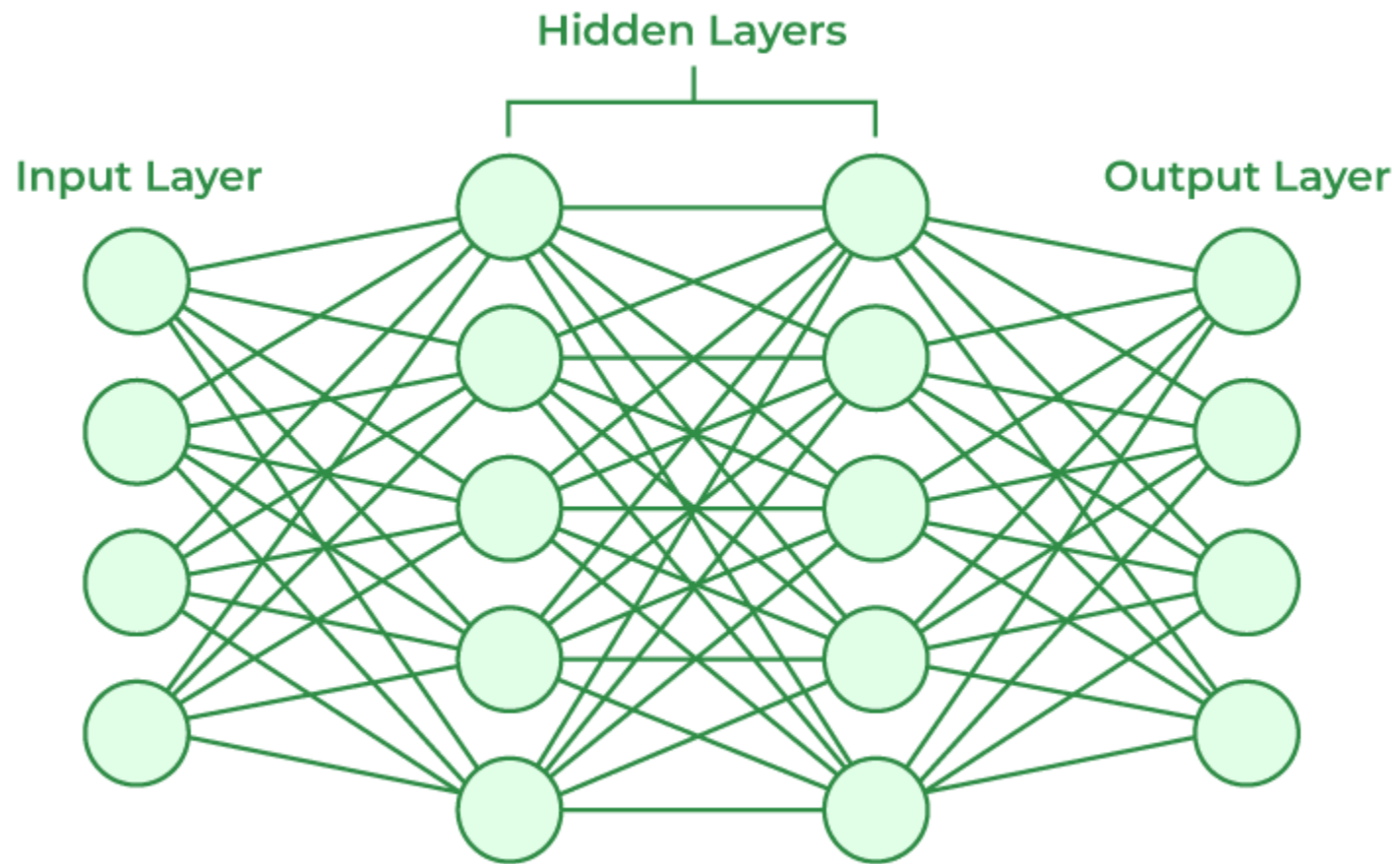
Neuroni



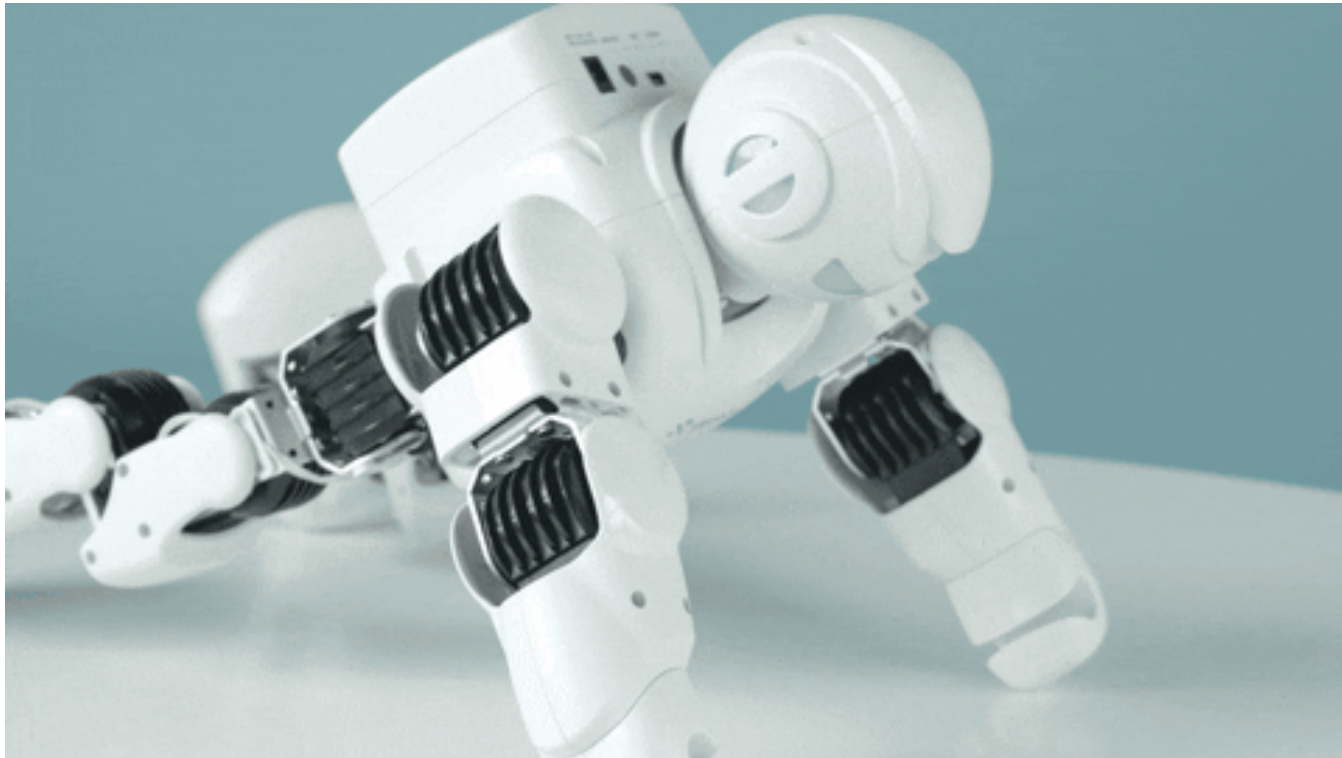
Funzione di Attivazione



Strati



Addestramento delle Reti Neurali

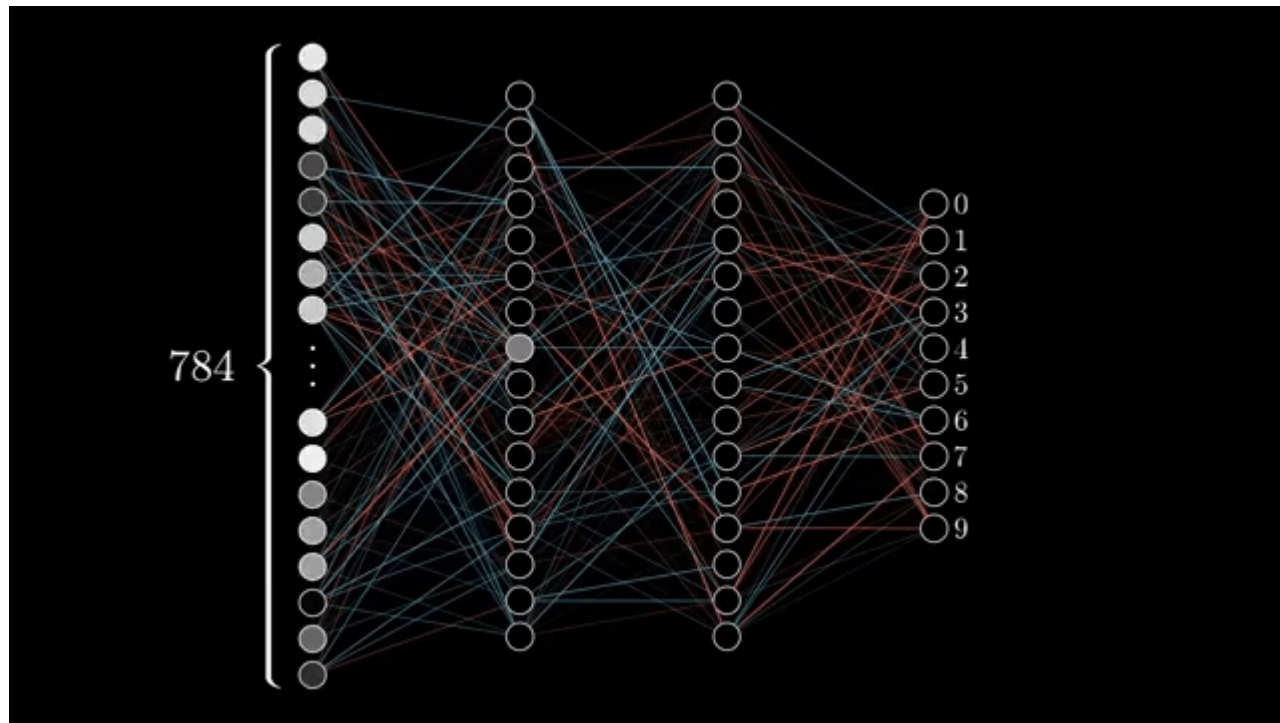


Un Compito di Classificazione



Gradient Descent

Previsioni Errate

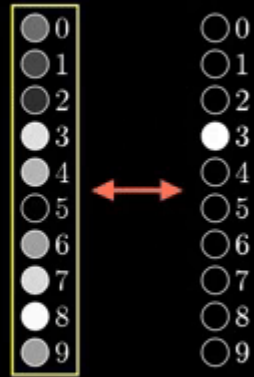


Valutazione della Perdita / Costo

Cost of 3

$\left\{ \begin{array}{l} (0.43 - 0.00)^2 + \\ (0.28 - 0.00)^2 + \\ (0.19 - 0.00)^2 + \\ (0.88 - 1.00)^2 + \\ (0.72 - 0.00)^2 + \\ (0.01 - 0.00)^2 + \\ (0.64 - 0.00)^2 + \\ (0.86 - 0.00)^2 + \\ (0.99 - 0.00)^2 + \\ (0.63 - 0.00)^2 \end{array} \right.$

What's the "cost" of this difference?



Utter trash

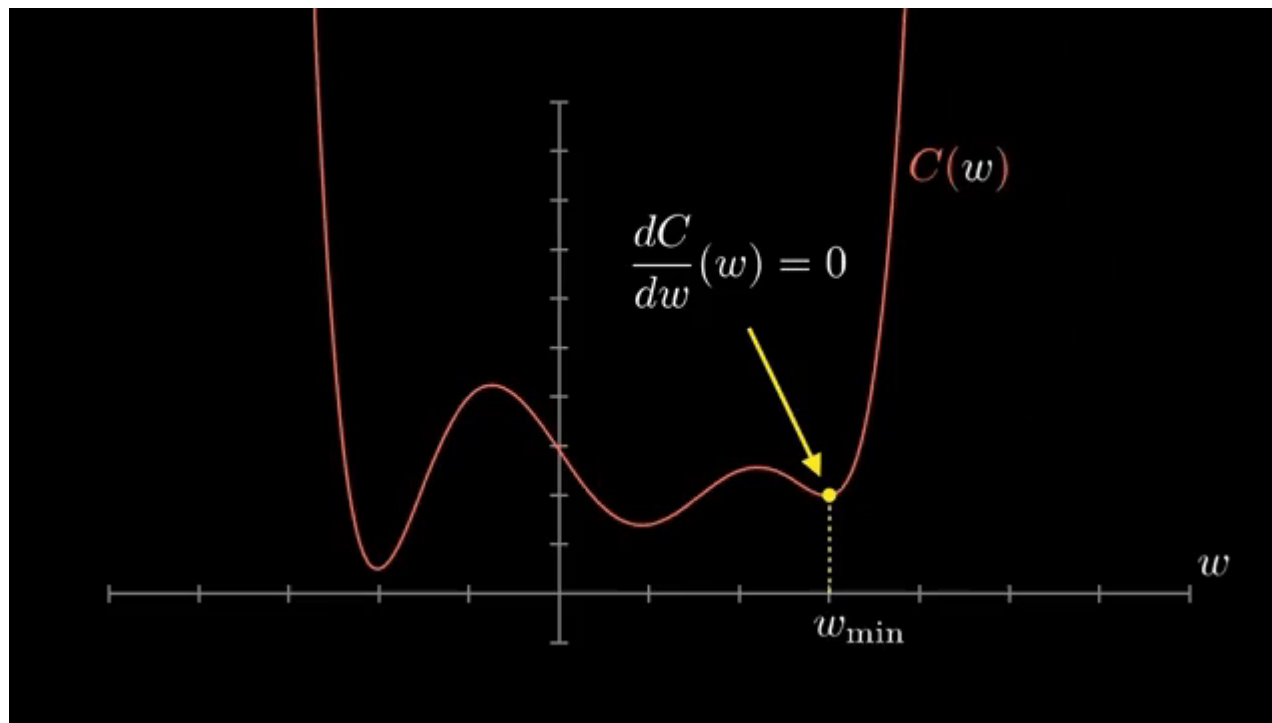
Cost Function

Cost function

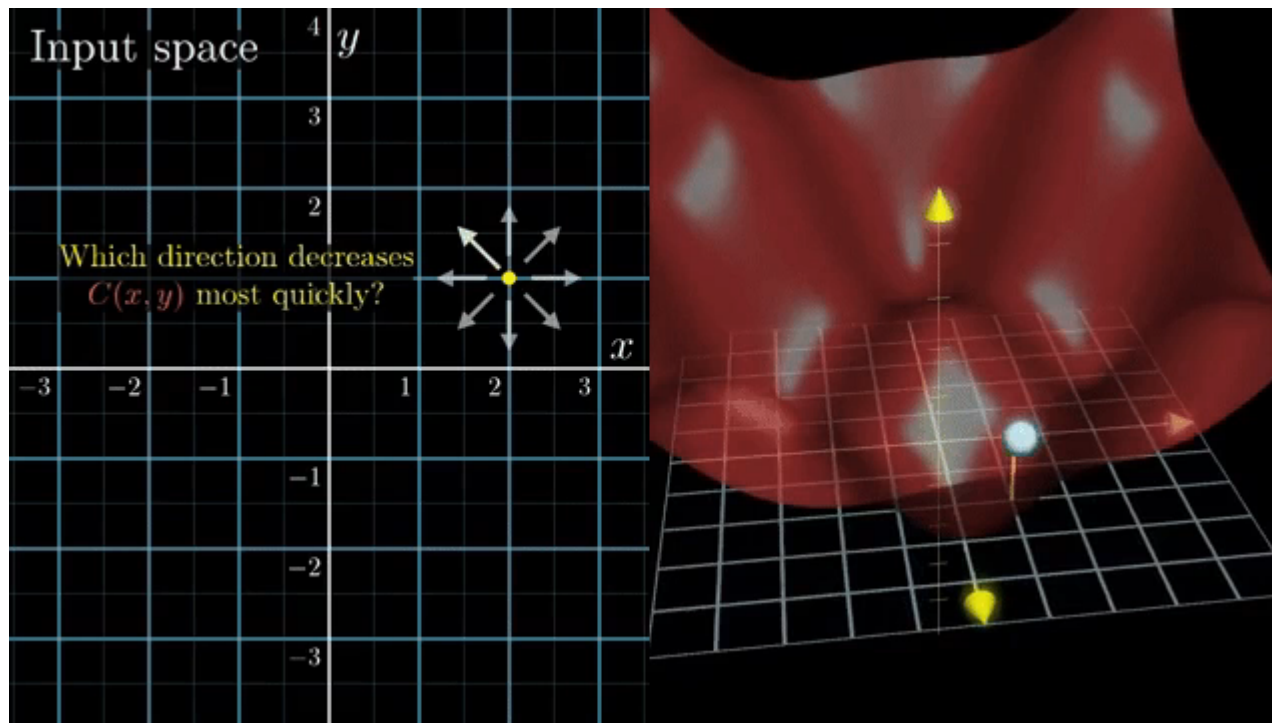
$$C(w_1, w_2, \dots, w_{13,002})$$

Weights and biases

Minimo Locale vs Globale

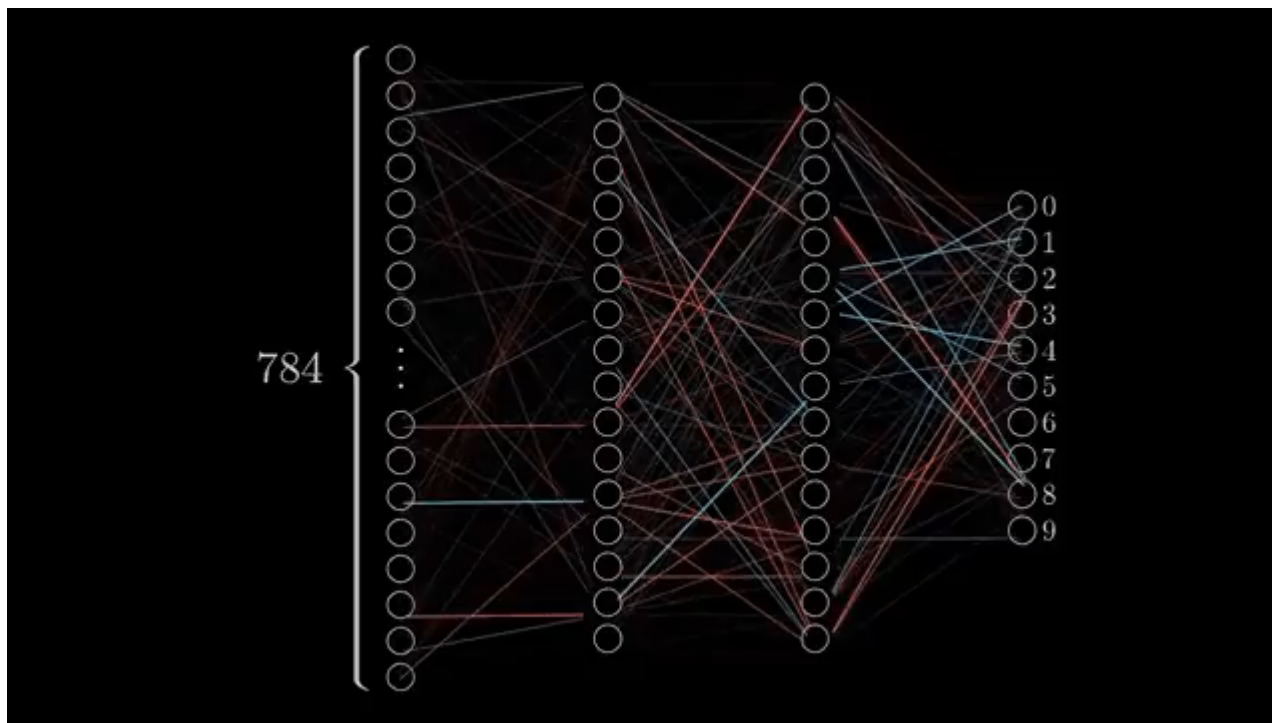


Direzione di Discesa



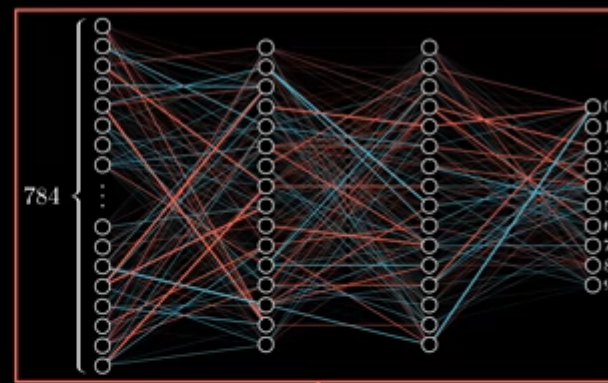
Backpropagation

Aggiornamento dei Parametri (1)



Aggiornamento dei Parametri (2)

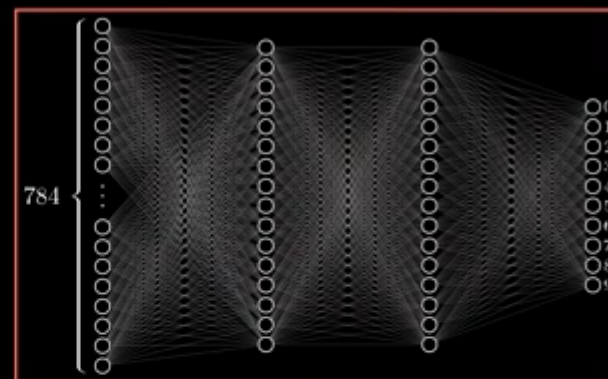
$$-\nabla C(\underbrace{\dots}_{\text{All weights and biases}}) = \begin{bmatrix} 0.20 \\ 0.83 \\ -0.84 \\ \vdots \\ 0.04 \\ 1.57 \\ 1.59 \end{bmatrix}$$



$C(w_0, w_1, \dots, w_{13,001}) = 3.4$

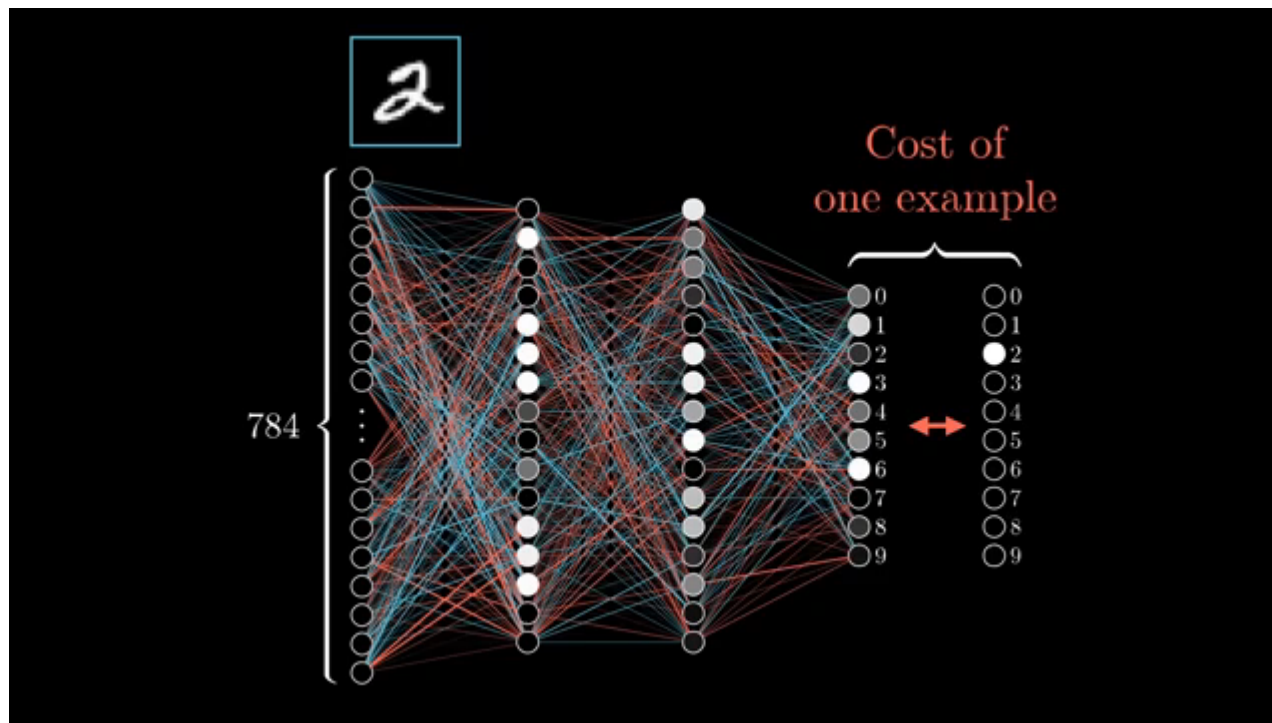
Influenza della Variazione

$$-\nabla C(\underbrace{\dots}_{\text{All weights and biases}}) = \begin{bmatrix} 0.16 \\ \vdots \\ 0.2 \\ \mathbf{3.20} \\ \vdots \\ 0.10 \\ 1.54 \\ 1.52 \end{bmatrix}$$

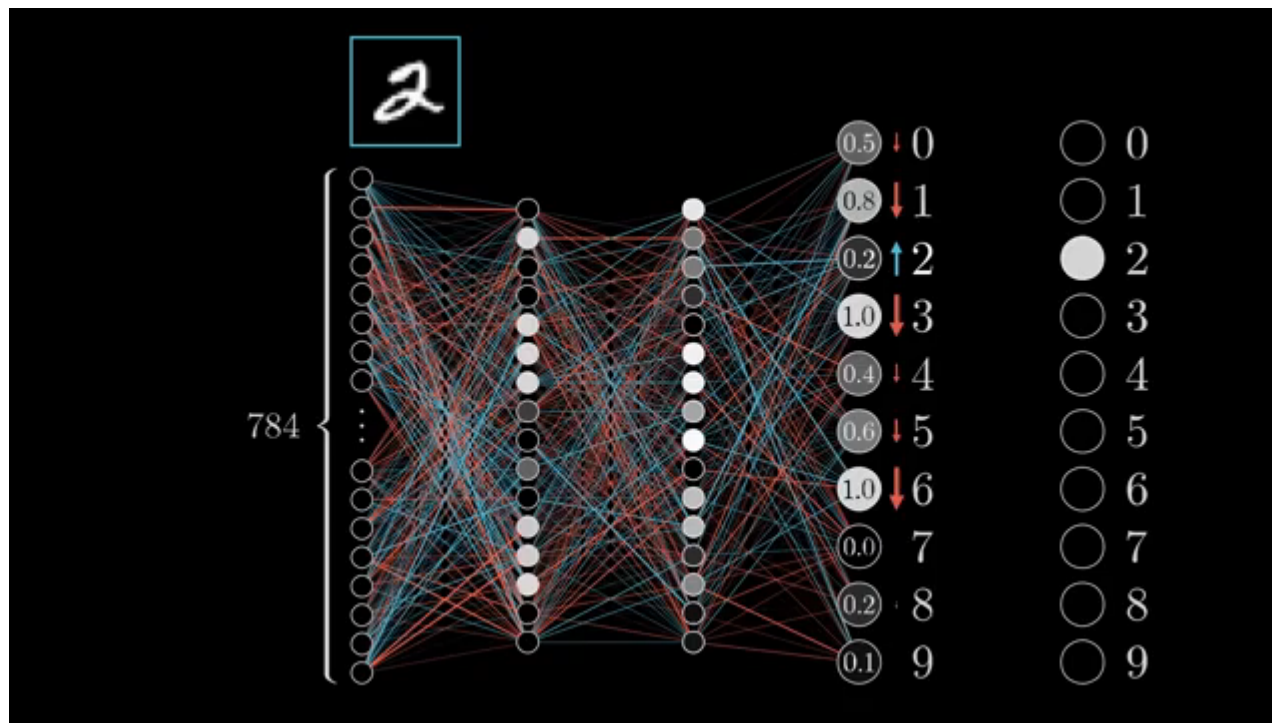


$$C(w_{0..1}, \mathbf{w_n}, \dots, w_{k..10}) = 2.85$$

Corrispondenza con le Aspettative



Regolazione di un Singolo Neurone



Regolazione di Neuroni Multipli



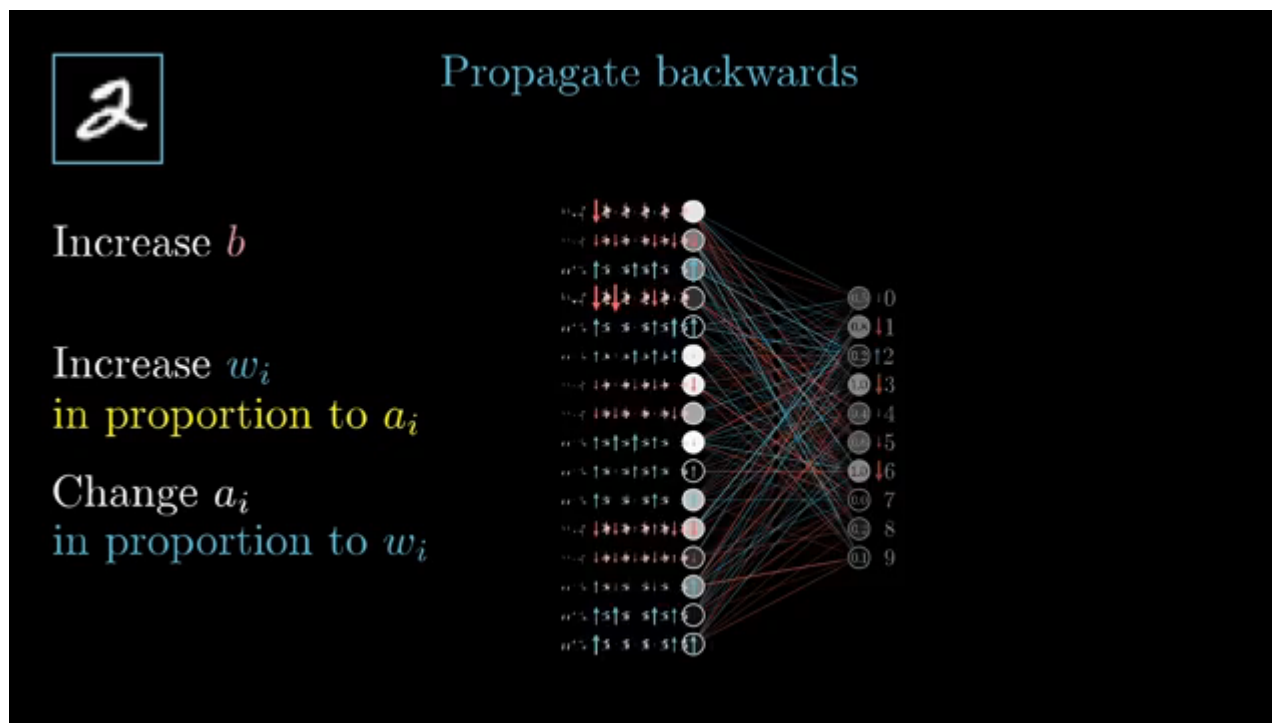
Increase b

Increase w_i
in proportion to a_i

Change a_i
in proportion to w_i



Propagazione tra Strati



Rete Neurale Molto Semplice

$$C(w_1, b_1, w_2, b_2, w_3, b_3)$$

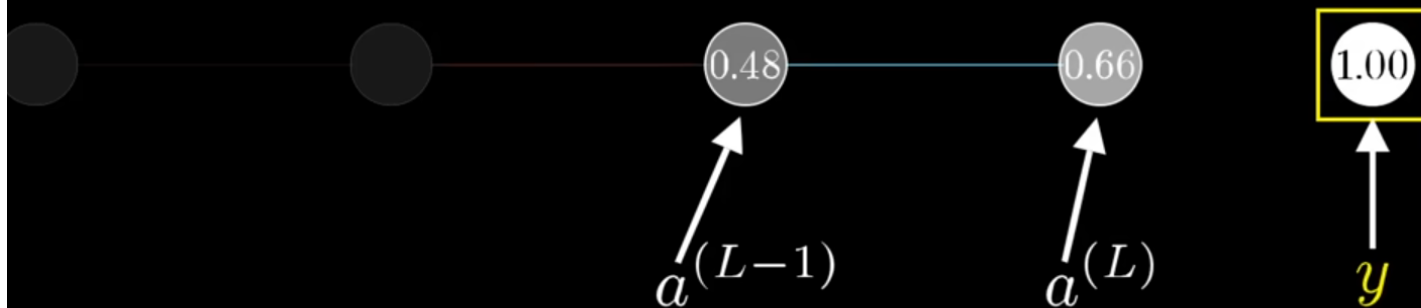


Calcolo della Funzione di Costo

$$C_0(\dots) = (a^{(L)} - y)^2$$

For example: $(0.66 - 1.00)^2$

Desired
output



Derivate e Regola della Catena

$$\frac{\partial C_0}{\partial a^{(L)}} = 2(a^{(L)} - y)$$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)}$$

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}}$$

Chain rule

Large Language Models

Cos'è un LLM (1)

Paris is a city in _____



→ France

Cos'è un LLM (2)

What follows is a conversation between a user and a helpful, very knowledgeable AI assistant.

User: Give me some ideas for what to do when visiting Santiago.

AI Assistant: Sure, there are plenty of things to do in Santiago! One option **could**



is	55%
could	31%
would	12%
might	0%
you	0%
may	0%
for	0%
that	0%
to	0%
I	0%
can	0%
,	0%
⋮	

Alimentazione

In 1912, the Titanic sank after hitting an iceberg on its maiden voyage.

Call me Ishmael.

The recipe said it would take 30 minutes, but three hours and two minor breakdowns later, I proudly served my family something that vaguely resembled food.

The Hubble Space Telescope has provided some of the most detailed images of distant galaxies.

The Galápagos Islands are known for their unique wildlife and were studied by Charles Darwin.

I think, therefore I am.

Large
Language
Model

The past is never dead. It's not even past.

I swear, if I hear one more person say "it's not the heat,

It was a bright cold day in

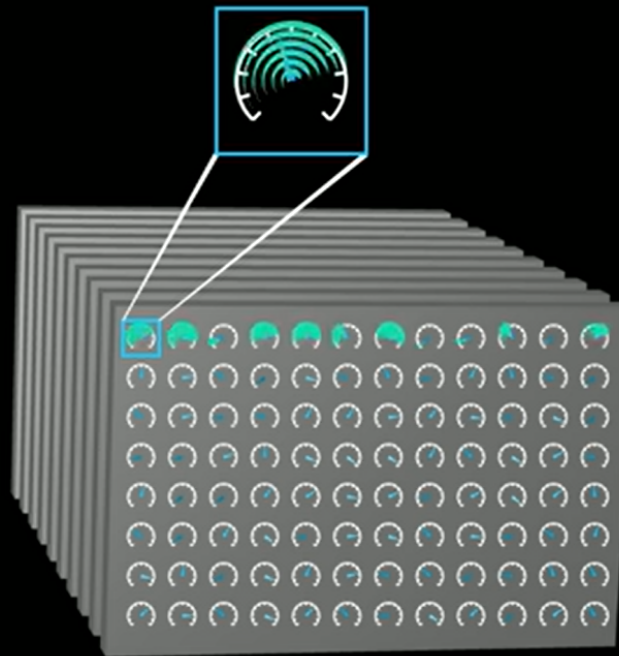
My friend asked me to be honest about her new haircut. I'm now looking for a new friend and a

Et tu, Brute?
The self-checkout machine kept yelling "unexpected item in bagging area" at me, and I've never felt so personally attacked by a

Regolazione

Parameter / Weight

It was the best
of times it was
the _____ →

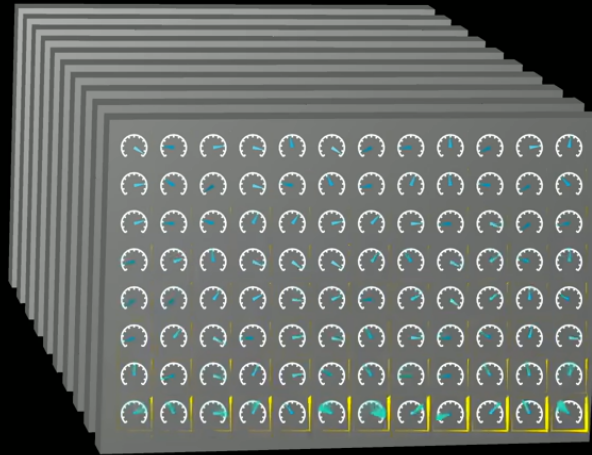


worst	<div></div>	60%
age	<div></div>	21%
worse	<div></div>	6%
best	<div></div>	5%
most	<div></div>	1%
end	<div></div>	1%
very	<div></div>	1%
blur	<div></div>	0%
⋮		

Previsione

It was the best of times it was the **worst**

It was the best
of times it was
the _____



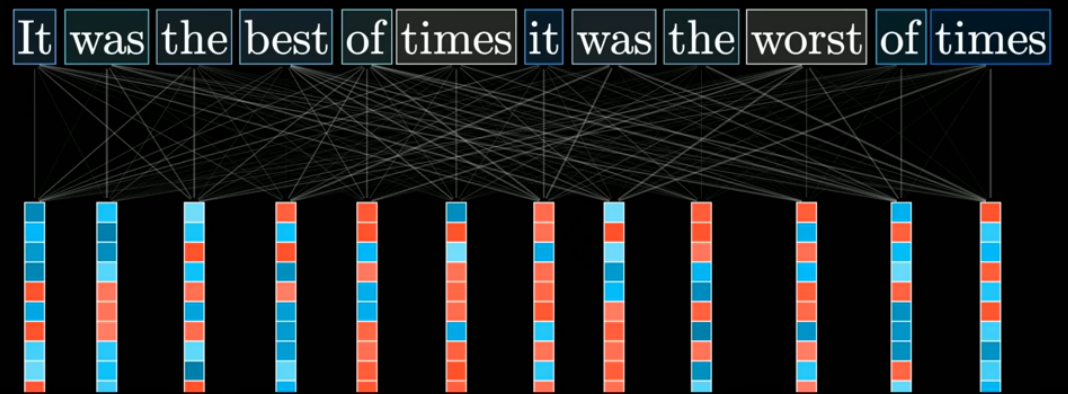
worst	<div></div>	57%
age	<div></div>	33%
worse	<div></div>	6%
best	<div></div>	0%
most	<div></div>	0%
end	<div></div>	0%
very	<div></div>	0%
blur	<div></div>	0%

Transformer (1)

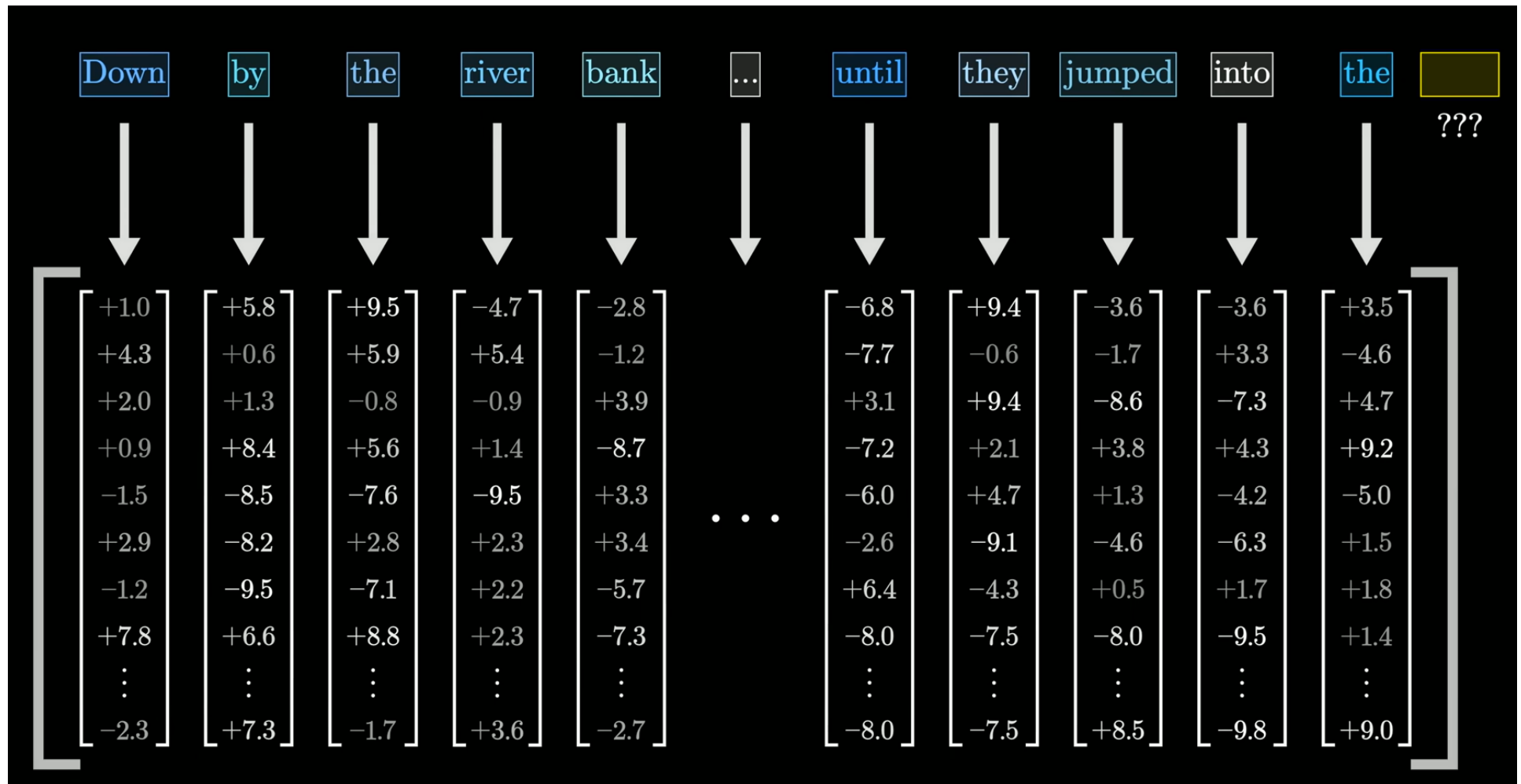
Previous Models



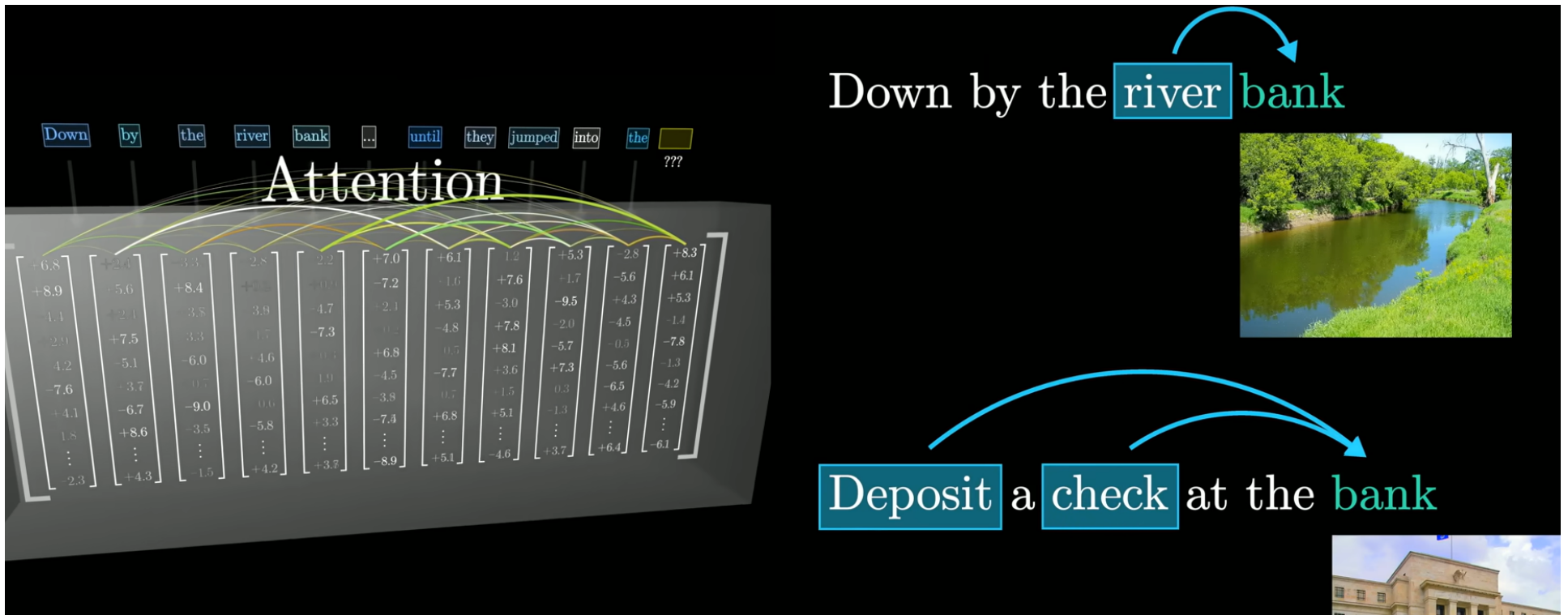
Transformers



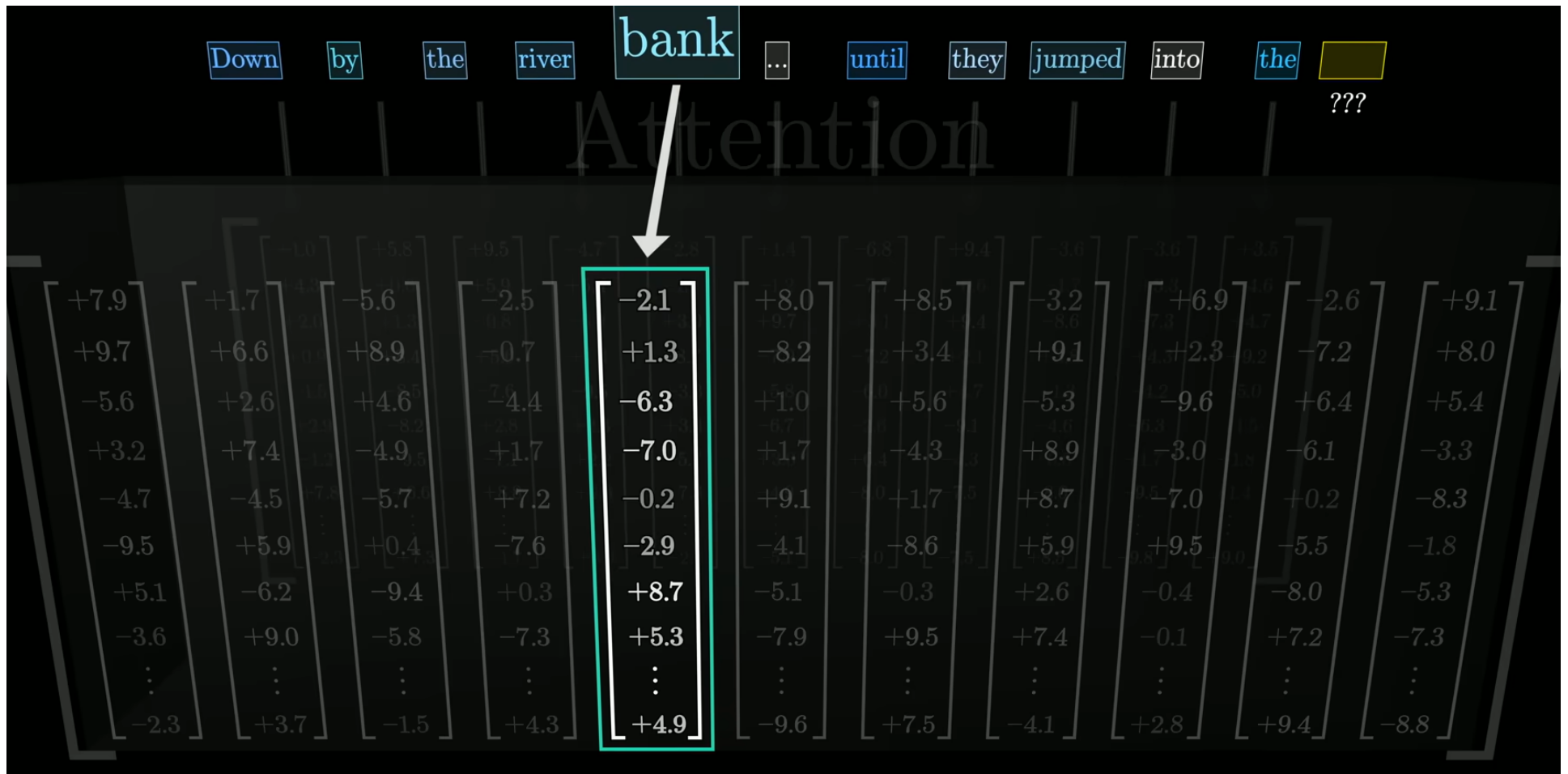
Embedding delle Parole



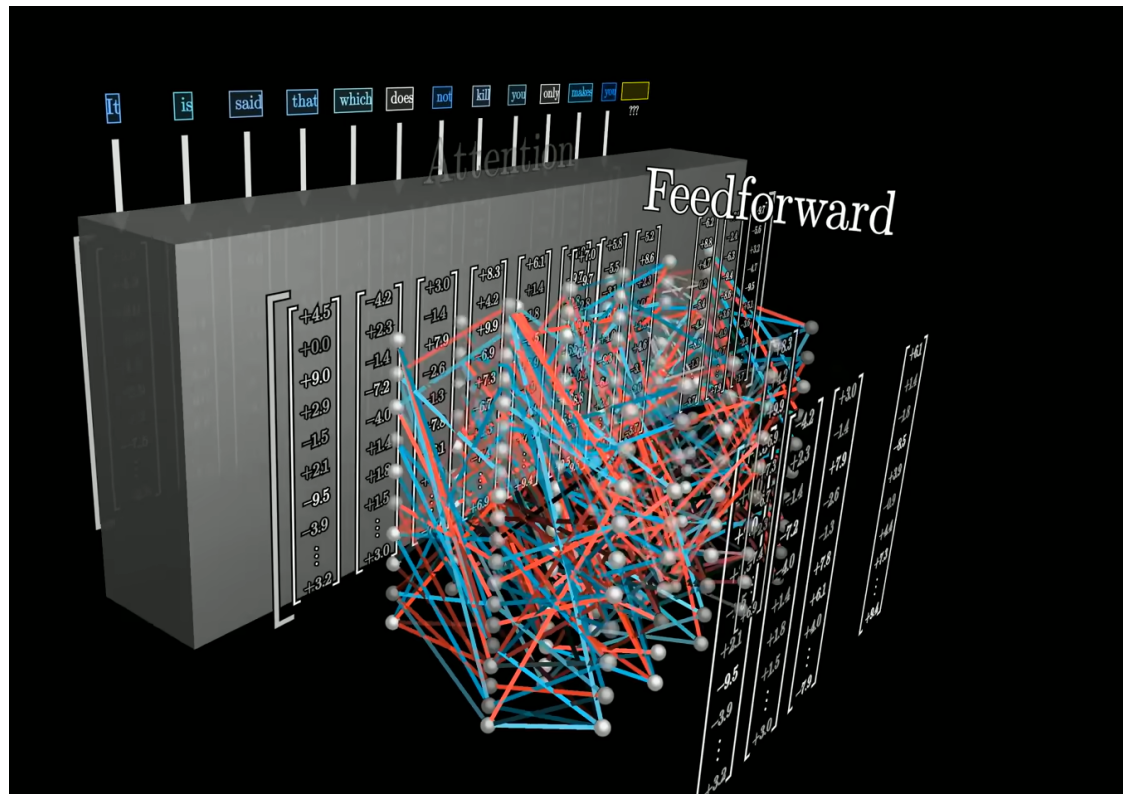
Meccanismo di Attenzione (1)



Meccanismo di Attenzione (2)

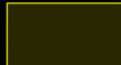


Memorizzazione e Recupero di Fatti



Selezione della Prossima Parola

Down by the river bank ... until they jumped into the



???




Embedding

Tokenizzazione

The Truth

This process (known fancifully as tokenization) frequently subdivides words



A Convenient Lie

It's nice to sometimes pretend tokens are words

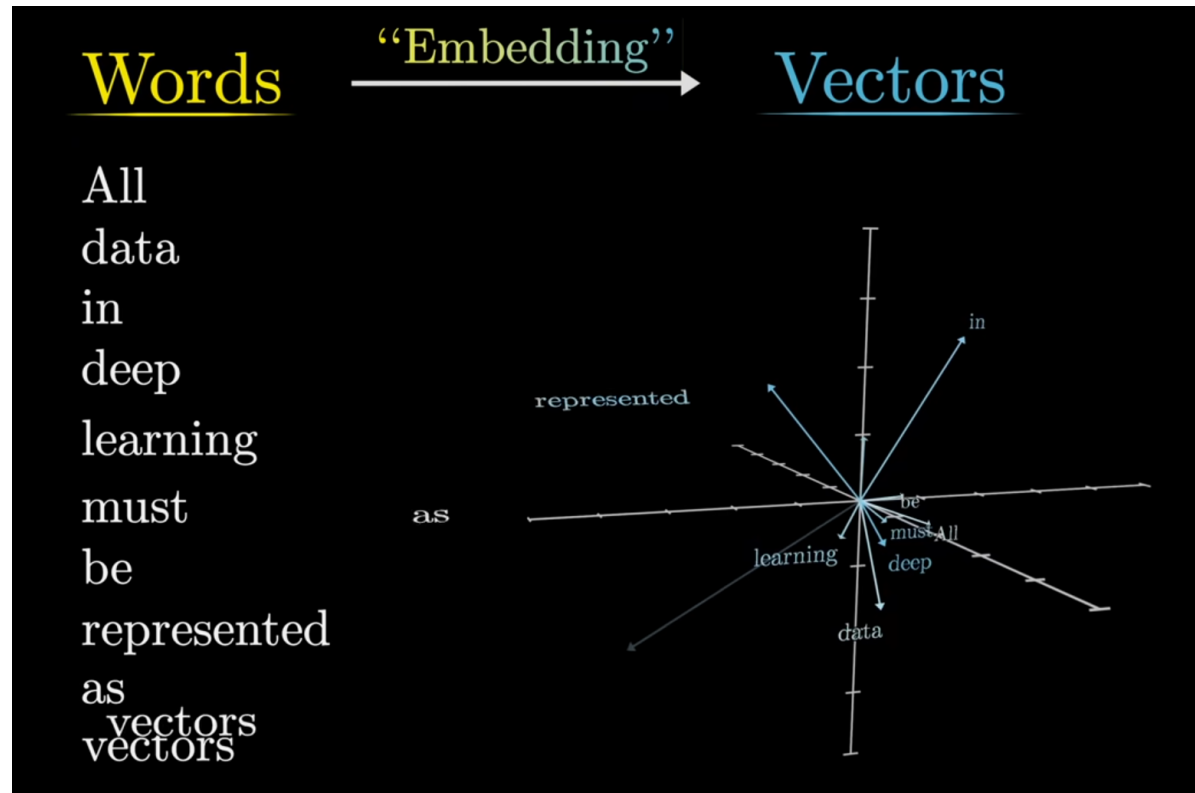
Vocabolario

All words, ~ 50k

aah	aardvark	aardwolf	aargh	ab	aback	abacterial	abacus	abalone	abandon	...	zygoid	zygomatic	zygomorphic	zygosis	zygote	zygotic	zyne	zymogen	zymosis	zzz
+7.0	-2.4	+2.8	-2.7	-6.5	+4.2	-2.6	+2.1	+5.1	+2.4	...	+1.6	+4.3	+2.3	-2.0	-7.3	-1.8	+5.4	+1.3	+7.9	+7.8
+2.4	-2.3	+4.0	-8.1	+0.6	-5.7	+6.2	+0.2	-2.2	-3.5	...	+0.5	+0.0	-8.1	+1.1	+2.5	+1.8	+8.1	+2.7	-3.1	-1.2
+3.5	-7.8	+3.4	+3.4	-5.3	-7.4	-3.5	-2.6	+1.5	-1.3	...	-7.9	-5.8	-6.7	+3.1	-4.9	-0.7	-5.1	-6.7	-7.7	+3.1
-7.2	-6.0	-2.6	+6.4	-8.0	+6.7	-8.0	+9.4	-0.6	+9.4	...	+4.7	-9.1	-4.3	-7.5	-4.0	-7.5	-3.6	-1.7	-8.6	+3.8
+1.3	-4.6	+0.5	-8.0	+1.5	+8.5	-3.6	+3.3	-7.3	+4.3	...	-6.3	+1.7	-9.5	+6.5	-9.8	+3.5	-4.6	+4.7	+9.2	-5.0
+1.5	+1.8	+1.4	-5.5	+9.0	-1.0	+6.9	+3.9	-4.0	+6.2	...	+7.5	+1.6	+7.6	+3.8	+4.5	+0.0	+9.0	+2.9	-1.5	+2.1
-9.5	-3.9	+3.2	-4.2	+2.3	-1.4	-7.2	-4.0	+1.4	+1.8	...	+3.0	+3.0	-1.4	+7.9	-2.6	-1.3	+7.8	+6.1	+4.0	-7.9
+8.3	+4.2	+9.9	-6.9	+7.3	-6.7	+2.3	-7.4	+6.9	+6.1	...	-1.8	-8.5	+3.9	-0.9	+4.4	+7.3	+9.4	+7.0	-9.7	-2.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-3.7	-2.0	-5.7	-6.2	+8.8	+4.7	-0.2	-5.4	-4.9	-8.8	...	-3.7	+3.9	-2.4	-6.3	-9.4	-8.6	+3.6	-0.9	+0.7	+7.9

Embedding matrix

Mappatura



Raggruppamento per Significato Simile

```
In [2]: import gensim.downloader # You need to pip install gensim
```

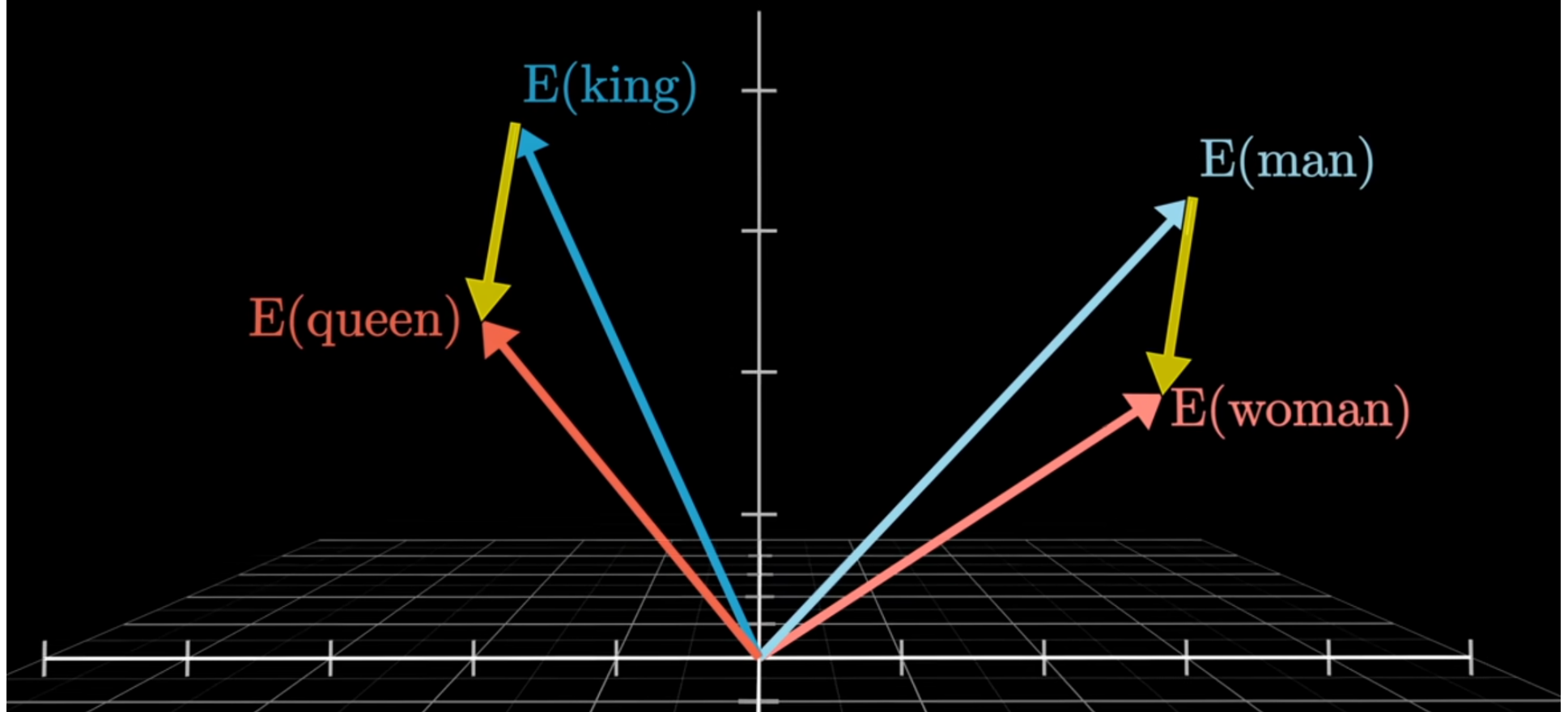
```
In [3]: model = gensim.downloader.load("glove-wiki-gigaword-50")
```

```
In [4]: model["tower"]
```



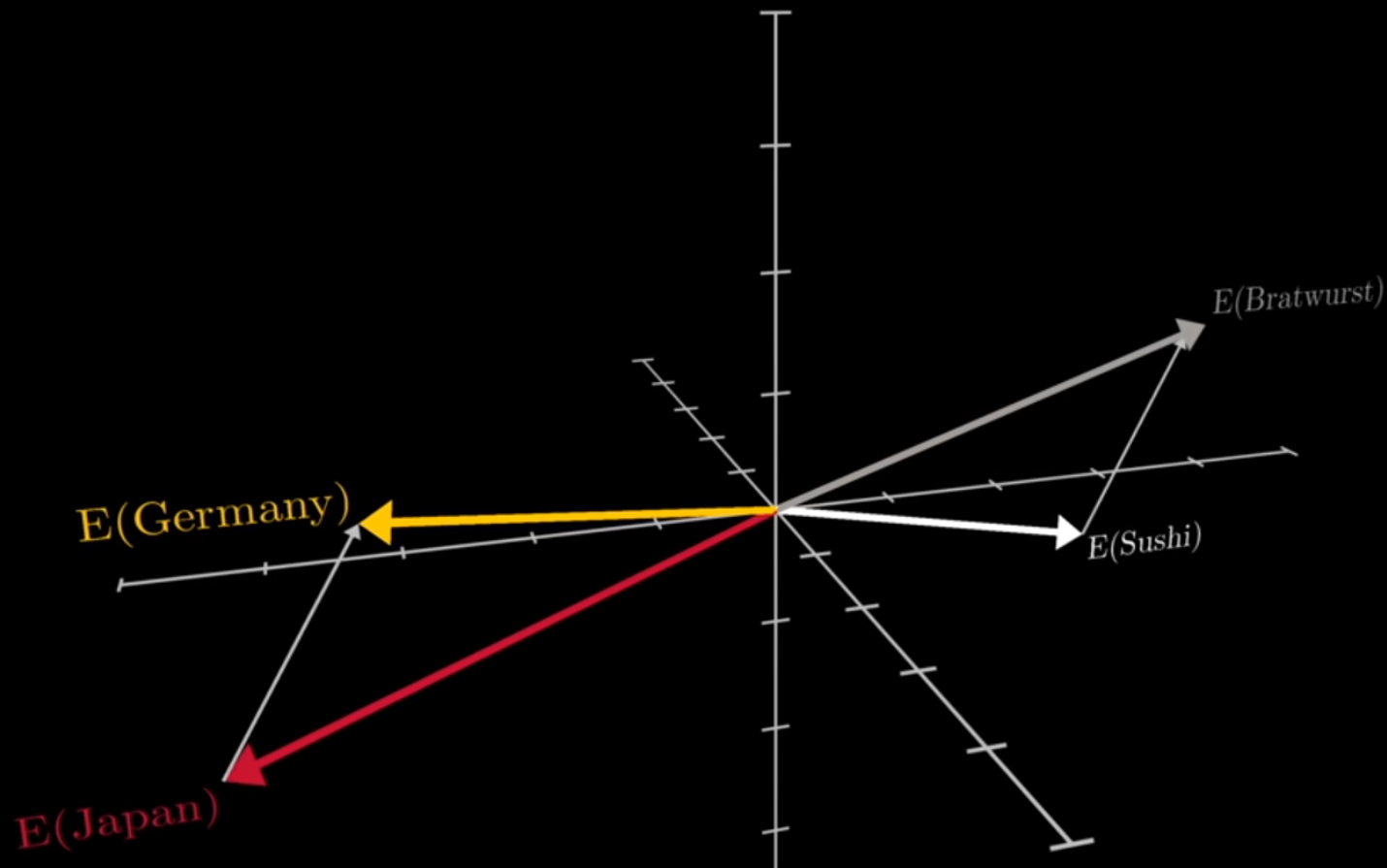
Significato della Direzione

$$E(\text{queen}) - E(\text{king}) \approx E(\text{woman}) - E(\text{man})$$

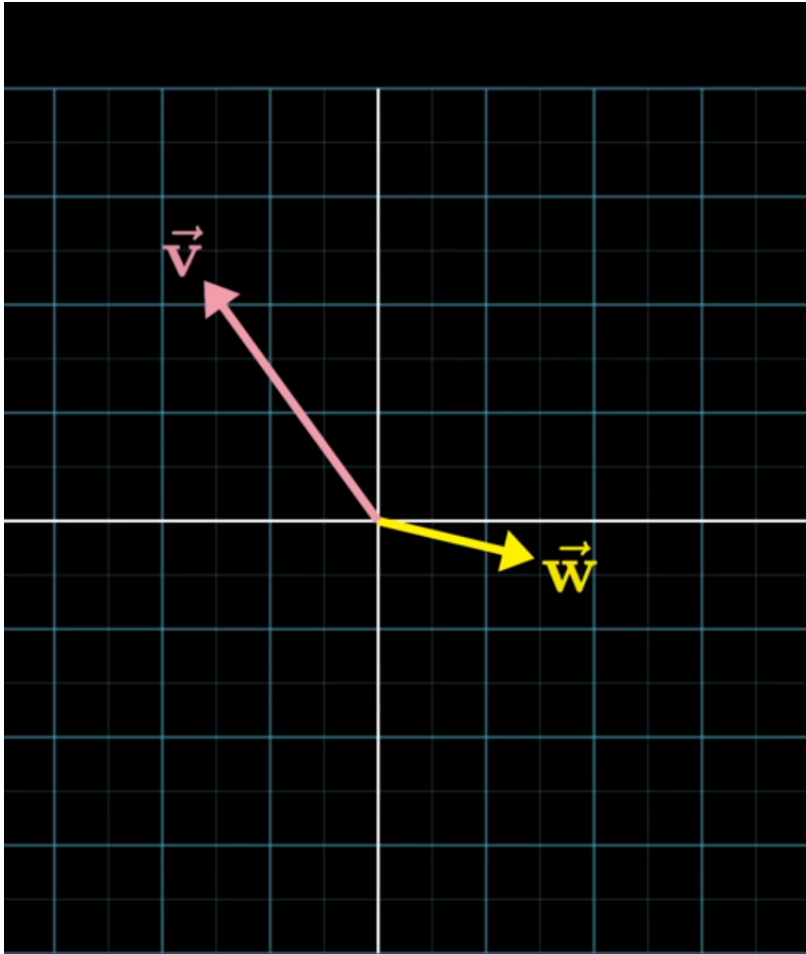


Aritmetica delle Parole

$$E(\text{Sushi}) + E(\text{Germany}) - E(\text{Japan}) \approx E(\text{Bratwurst})$$



Similarità tra Vettori



$$\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix}}_{\text{Dot product}} = \begin{aligned} &v_1 w_1 \\ &+ \\ &v_2 w_2 \\ &+ \\ &v_3 w_3 \\ &+ \\ &\vdots \\ &+ \\ &v_n w_n \end{aligned} \quad || \quad -3.11$$

Attention

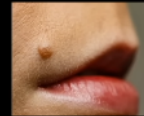
Significati Multipli



American shrew mole

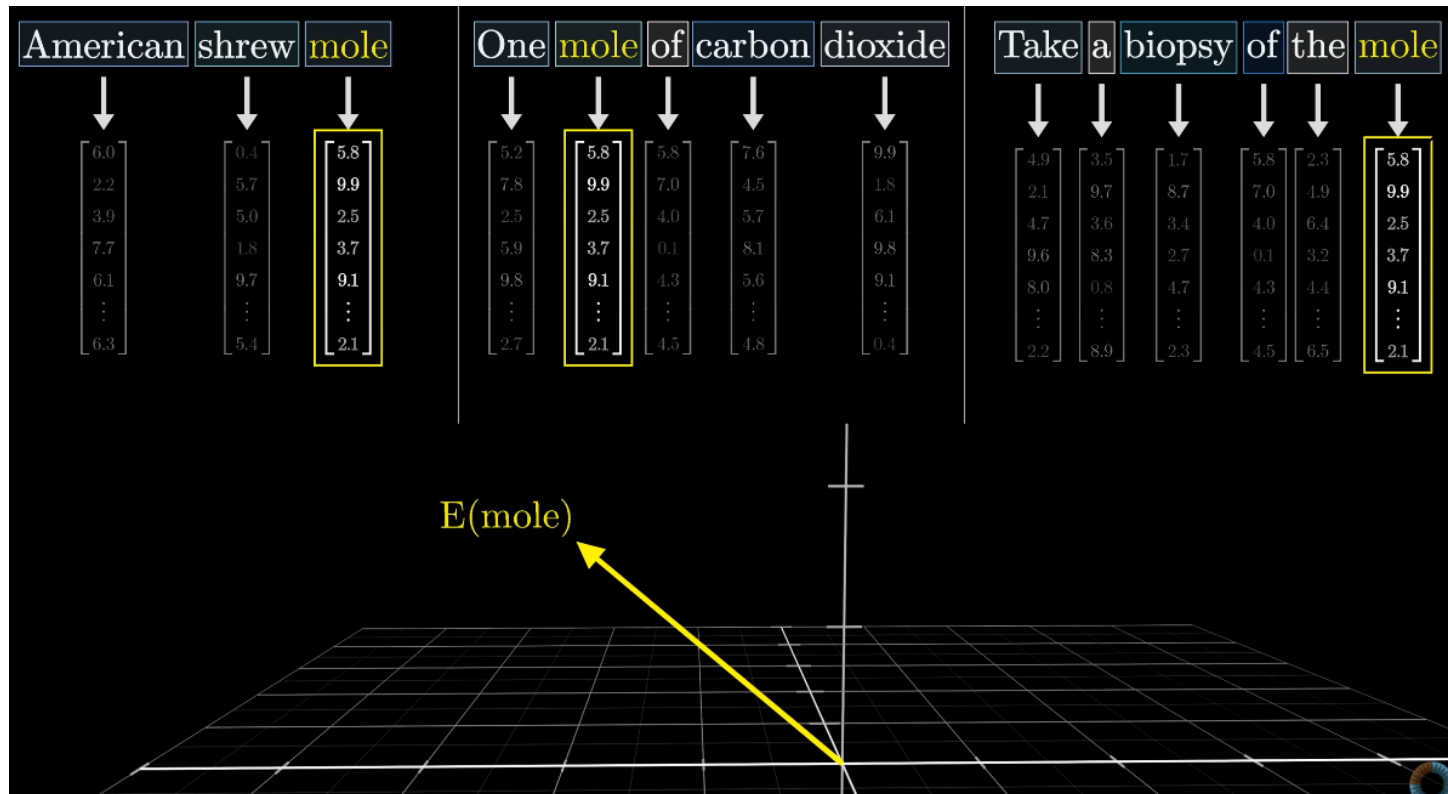
6.02×10^{23}

One mole of carbon dioxide

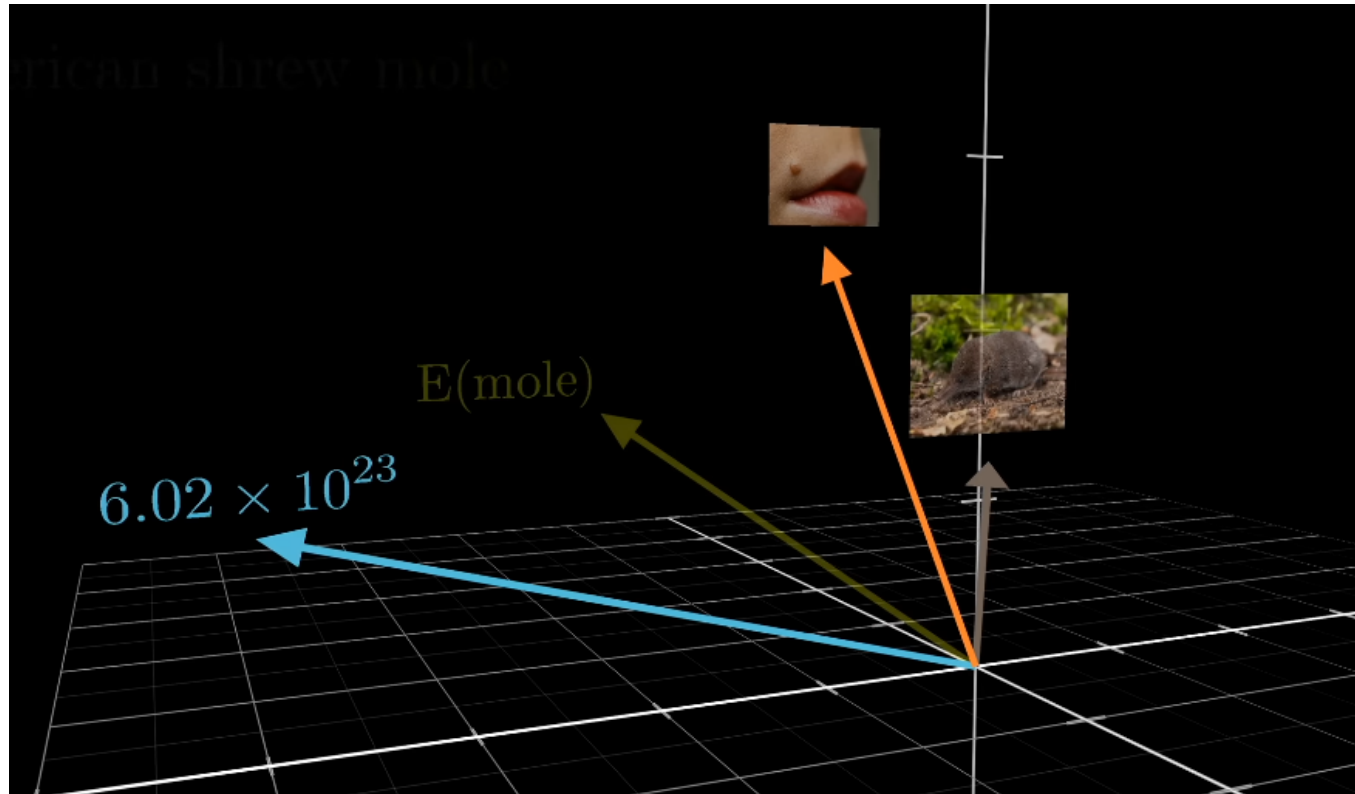


Take a biopsy of the mole

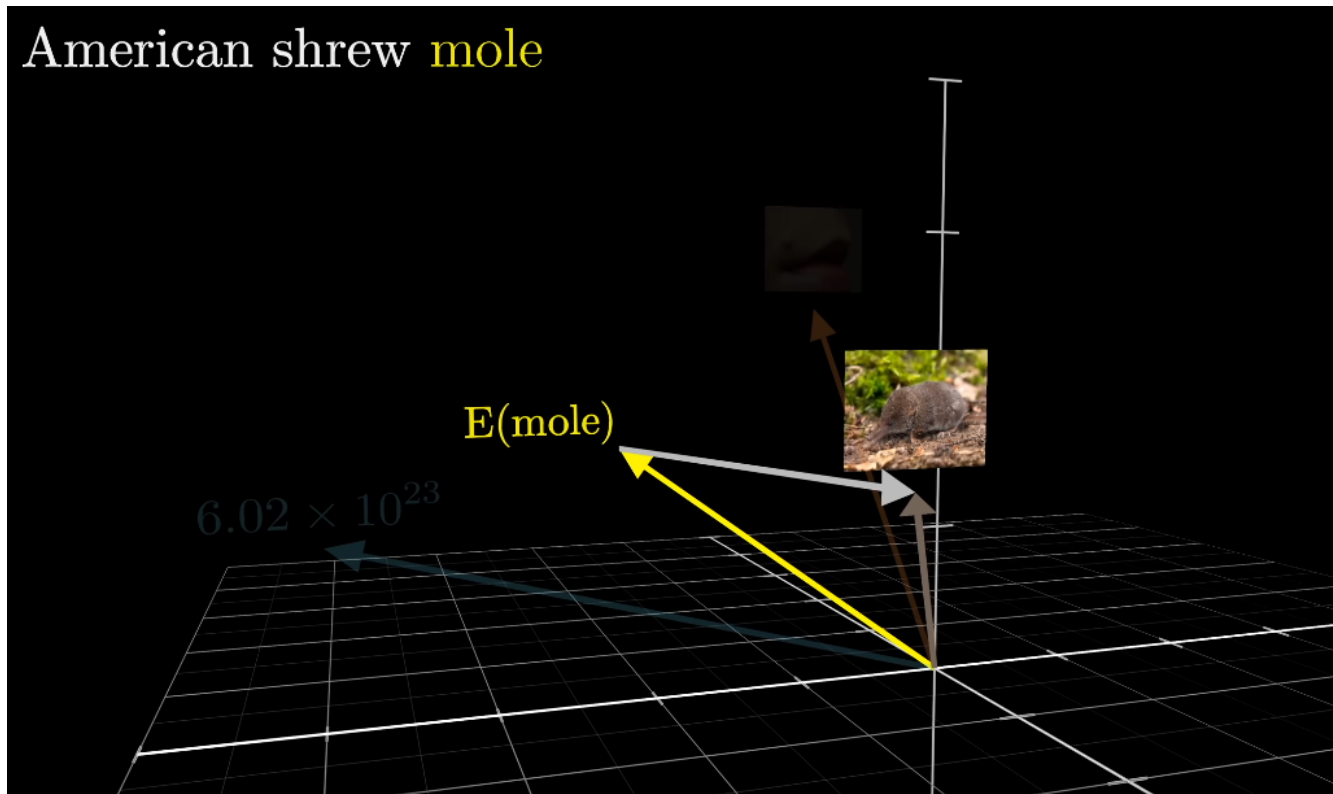
Embedding dei Token



Direzioni di Significato



Raffinamento dell'Attention

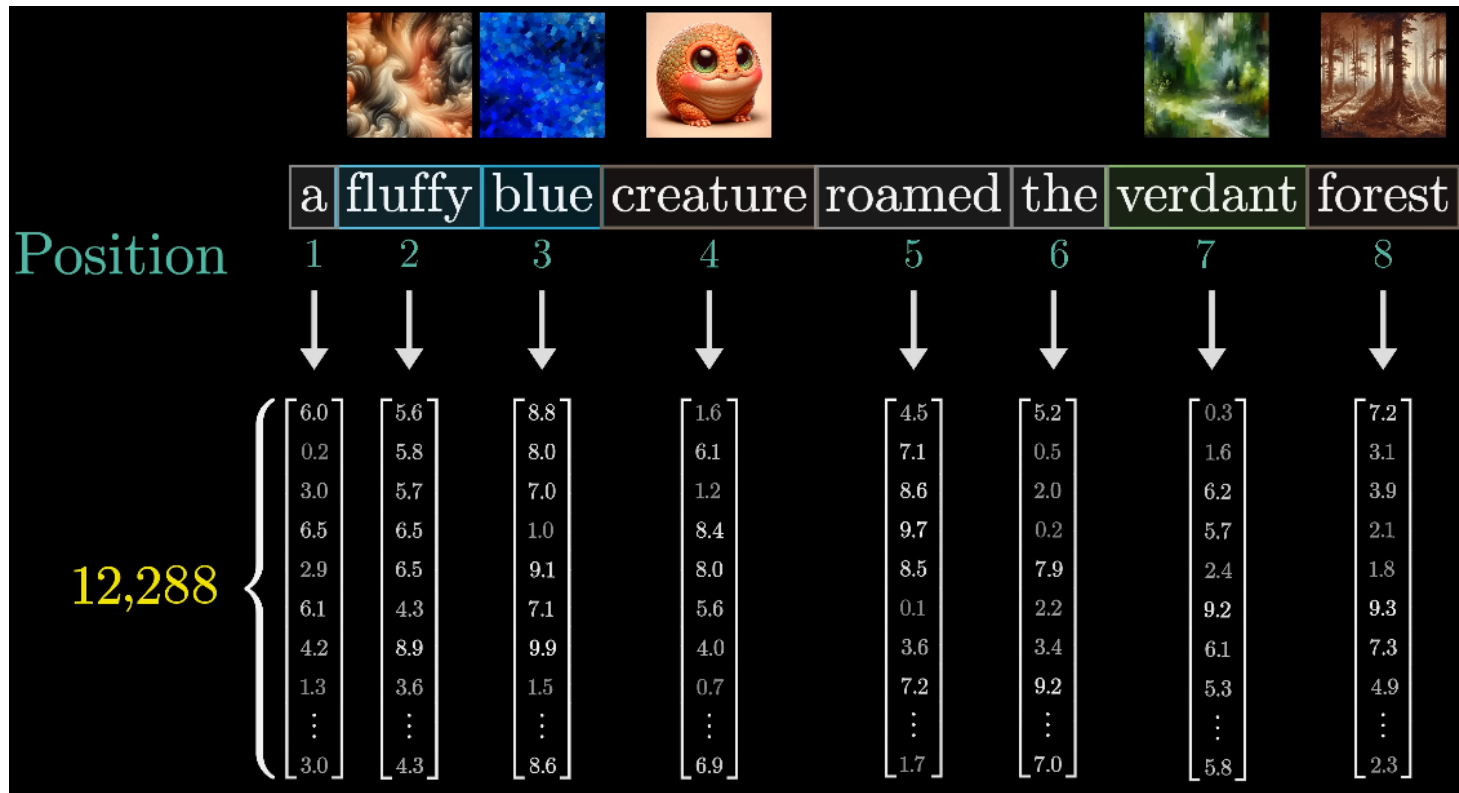


Influenza del Contesto

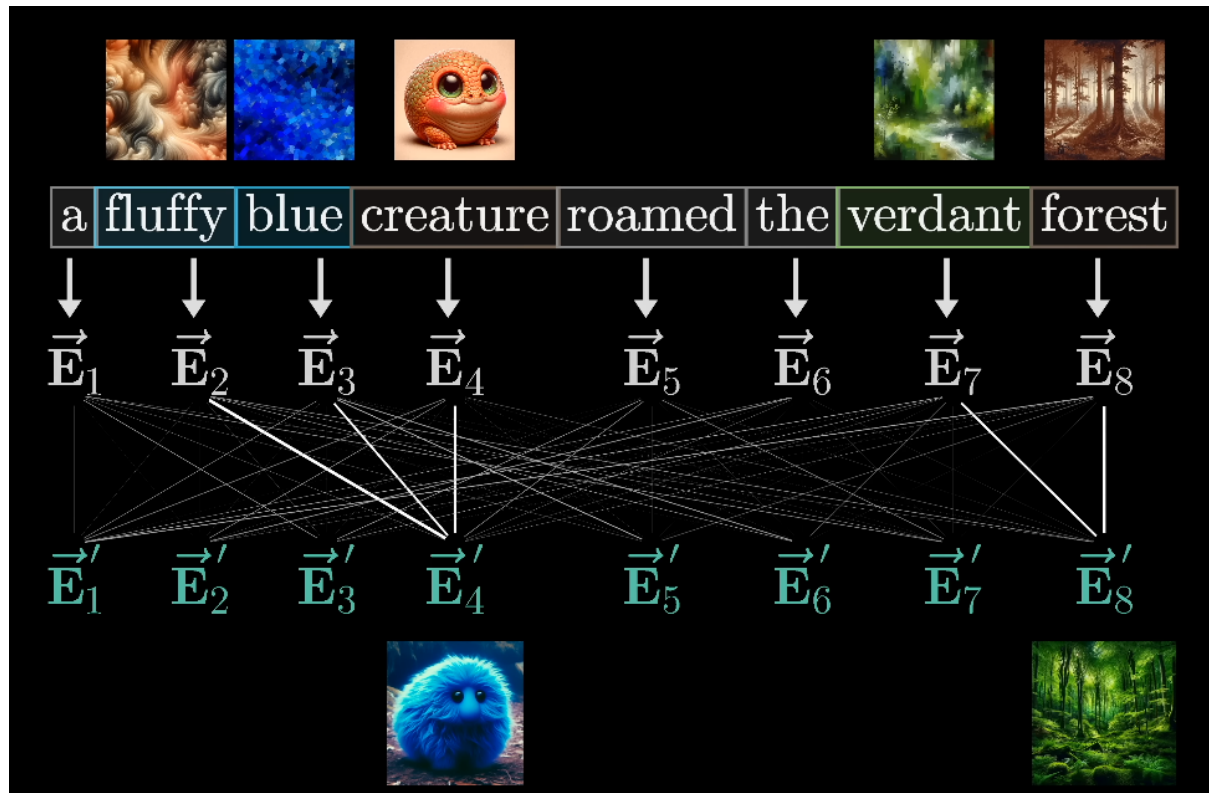
a fluffy blue creature roamed the verdant forest



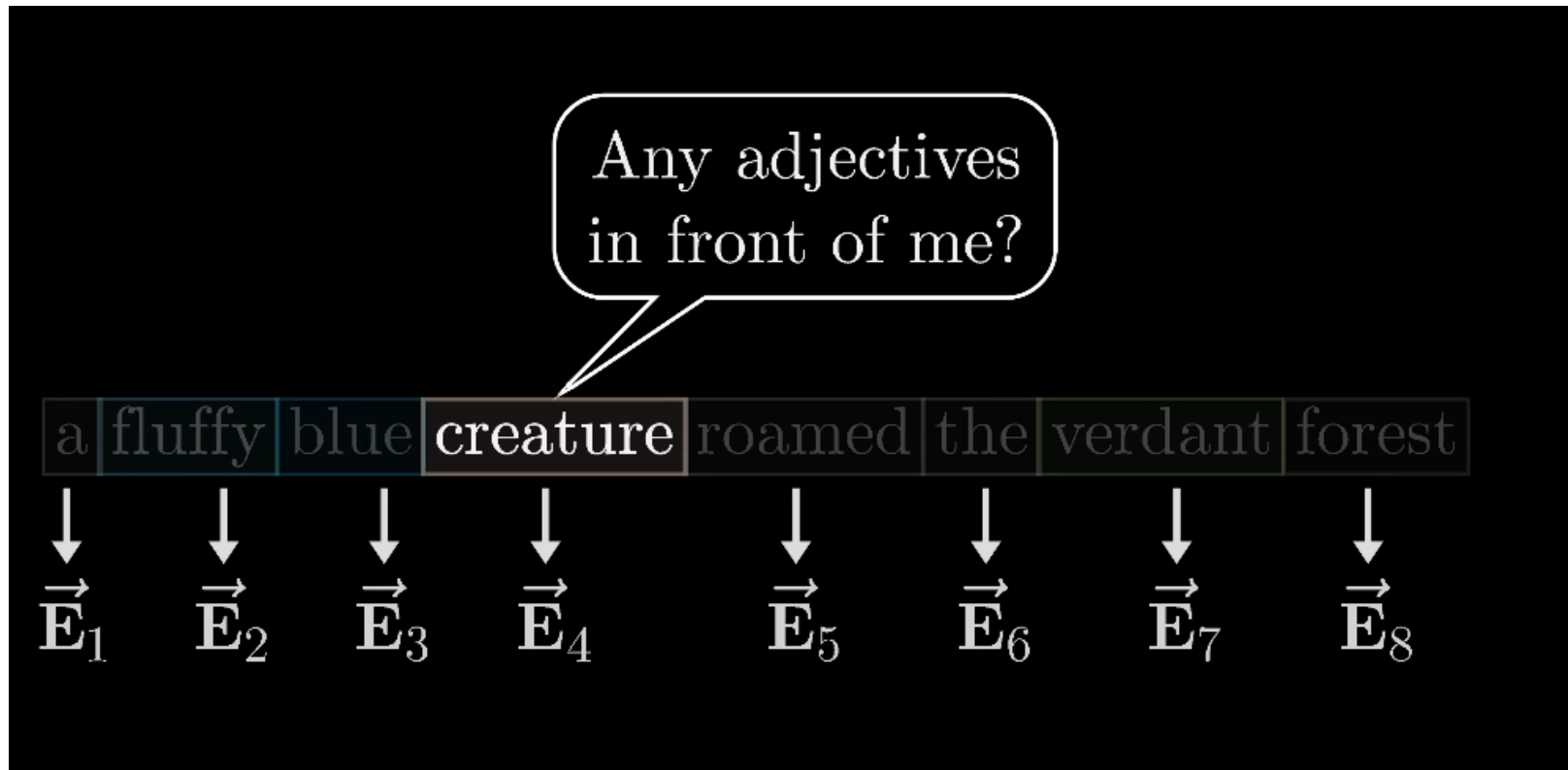
Codifica della Posizione



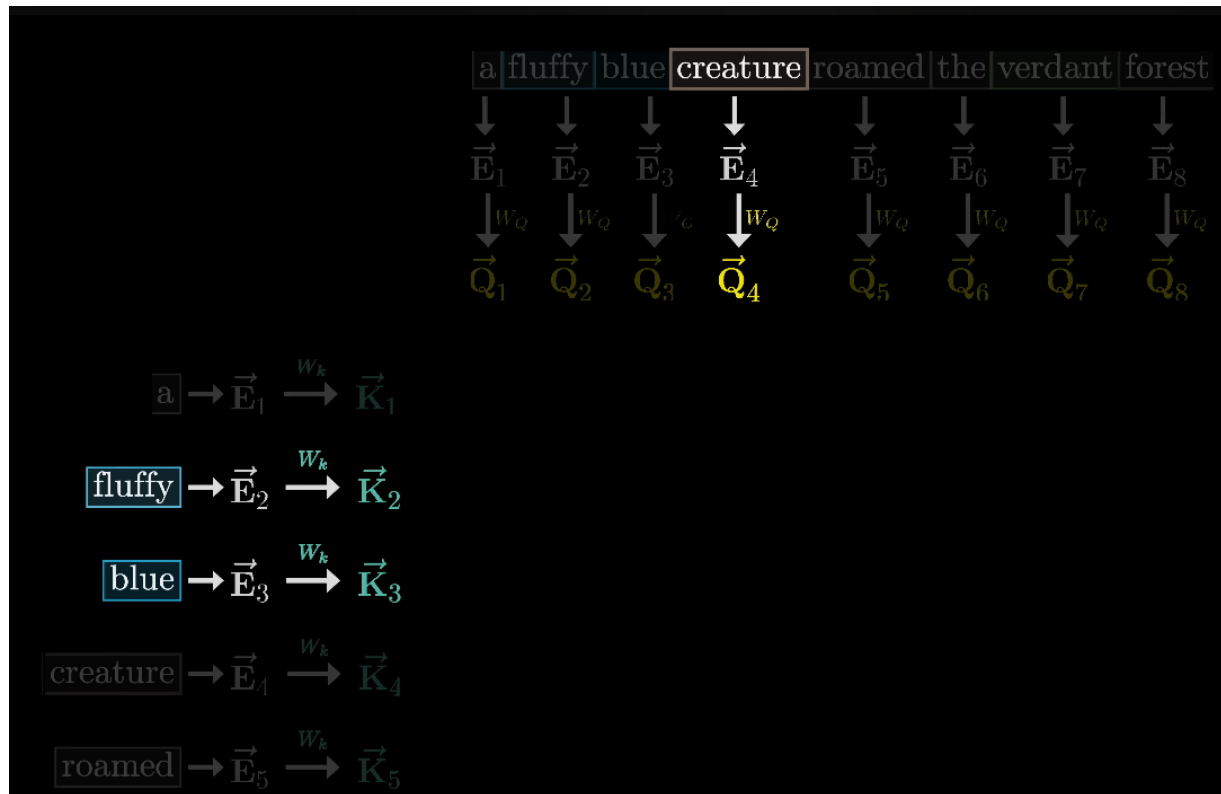
Raffinamento dell'Embedding



Query



Matrice delle Chiavi



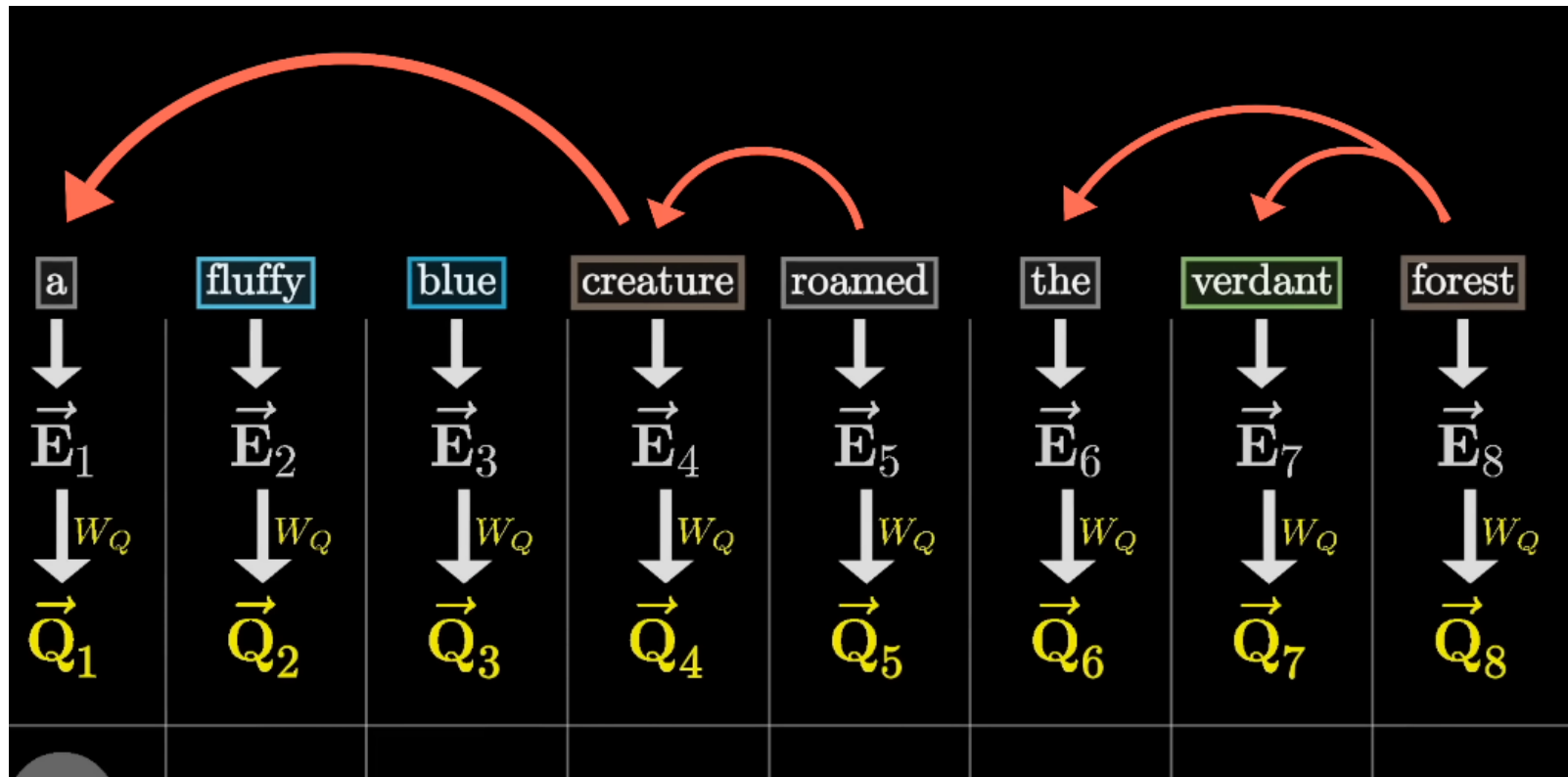
Prodotto Scalare

	a	fluffy	blue	creature	roamed	the	verdant	forest	
	\downarrow \vec{E}_1 \downarrow_{W_Q} \vec{Q}_1	\downarrow \vec{E}_2 \downarrow_{W_Q} \vec{Q}_2	\downarrow \vec{E}_3 \downarrow_{W_Q} \vec{Q}_3	\downarrow \vec{E}_4 \downarrow_{W_Q} \vec{Q}_4	\downarrow \vec{E}_5 \downarrow_{W_Q} \vec{Q}_5	\downarrow \vec{E}_6 \downarrow_{W_Q} \vec{Q}_6	\downarrow \vec{E}_7 \downarrow_{W_Q} \vec{Q}_7	\downarrow \vec{E}_8 \downarrow_{W_Q} \vec{Q}_8	
$\boxed{\text{a}} \rightarrow \vec{E}_1 \xrightarrow{W_k} \vec{K}_1$	$\vec{K}_1 \cdot \vec{Q}_1$	$\vec{K}_1 \cdot \vec{Q}_2$	$\vec{K}_1 \cdot \vec{Q}_3$	$\vec{K}_1 \cdot \vec{Q}_4$	$\vec{K}_1 \cdot \vec{Q}_5$	$\vec{K}_1 \cdot \vec{Q}_6$	$\vec{K}_1 \cdot \vec{Q}_7$	$\vec{K}_1 \cdot \vec{Q}_8$	
$\boxed{\text{fluffy}} \rightarrow \vec{E}_2 \xrightarrow{W_k} \vec{K}_2$	$\vec{K}_2 \cdot \vec{Q}_1$	$\vec{K}_2 \cdot \vec{Q}_2$	$\vec{K}_2 \cdot \vec{Q}_3$	$\vec{K}_2 \cdot \vec{Q}_4$	$\vec{K}_2 \cdot \vec{Q}_5$	$\vec{K}_2 \cdot \vec{Q}_6$	$\vec{K}_2 \cdot \vec{Q}_7$	$\vec{K}_2 \cdot \vec{Q}_8$	
$\boxed{\text{blue}} \rightarrow \vec{E}_3 \xrightarrow{W_k} \vec{K}_3$	$\vec{K}_3 \cdot \vec{Q}_1$	$\vec{K}_3 \cdot \vec{Q}_2$	$\vec{K}_3 \cdot \vec{Q}_3$	$\vec{K}_3 \cdot \vec{Q}_4$	$\vec{K}_3 \cdot \vec{Q}_5$	$\vec{K}_3 \cdot \vec{Q}_6$	$\vec{K}_3 \cdot \vec{Q}_7$	$\vec{K}_3 \cdot \vec{Q}_8$	
$\boxed{\text{creature}} \rightarrow \vec{E}_4 \xrightarrow{W_k} \vec{K}_4$	$\vec{K}_4 \cdot \vec{Q}_1$	$\vec{K}_4 \cdot \vec{Q}_2$	$\vec{K}_4 \cdot \vec{Q}_3$	$\vec{K}_4 \cdot \vec{Q}_4$	$\vec{K}_4 \cdot \vec{Q}_5$	$\vec{K}_4 \cdot \vec{Q}_6$	$\vec{K}_4 \cdot \vec{Q}_7$	$\vec{K}_4 \cdot \vec{Q}_8$	
$\boxed{\text{roamed}} \rightarrow \vec{E}_5 \xrightarrow{W_k} \vec{K}_5$	$\vec{K}_5 \cdot \vec{Q}_1$	$\vec{K}_5 \cdot \vec{Q}_2$	$\vec{K}_5 \cdot \vec{Q}_3$	$\vec{K}_5 \cdot \vec{Q}_4$	$\vec{K}_5 \cdot \vec{Q}_5$	$\vec{K}_5 \cdot \vec{Q}_6$	$\vec{K}_5 \cdot \vec{Q}_7$	$\vec{K}_5 \cdot \vec{Q}_8$	
$\boxed{\text{the}} \rightarrow \vec{E}_6 \xrightarrow{W_k} \vec{K}_6$	$\vec{K}_6 \cdot \vec{Q}_1$	$\vec{K}_6 \cdot \vec{Q}_2$	$\vec{K}_6 \cdot \vec{Q}_3$	$\vec{K}_6 \cdot \vec{Q}_4$	$\vec{K}_6 \cdot \vec{Q}_5$	$\vec{K}_6 \cdot \vec{Q}_6$	$\vec{K}_6 \cdot \vec{Q}_7$	$\vec{K}_6 \cdot \vec{Q}_8$	
$\boxed{\text{verdant}} \rightarrow \vec{E}_7 \xrightarrow{W_k} \vec{K}_7$	$\vec{K}_7 \cdot \vec{Q}_1$	$\vec{K}_7 \cdot \vec{Q}_2$	$\vec{K}_7 \cdot \vec{Q}_3$	$\vec{K}_7 \cdot \vec{Q}_4$	$\vec{K}_7 \cdot \vec{Q}_5$	$\vec{K}_7 \cdot \vec{Q}_6$	$\vec{K}_7 \cdot \vec{Q}_7$	$\vec{K}_7 \cdot \vec{Q}_8$	
$\boxed{\text{forest}} \rightarrow \vec{E}_8 \xrightarrow{W_k} \vec{K}_8$	$\vec{K}_8 \cdot \vec{Q}_1$	$\vec{K}_8 \cdot \vec{Q}_2$	$\vec{K}_8 \cdot \vec{Q}_3$	$\vec{K}_8 \cdot \vec{Q}_4$	$\vec{K}_8 \cdot \vec{Q}_5$	$\vec{K}_8 \cdot \vec{Q}_6$	$\vec{K}_8 \cdot \vec{Q}_7$	$\vec{K}_8 \cdot \vec{Q}_8$	

Applicazione del Softmax



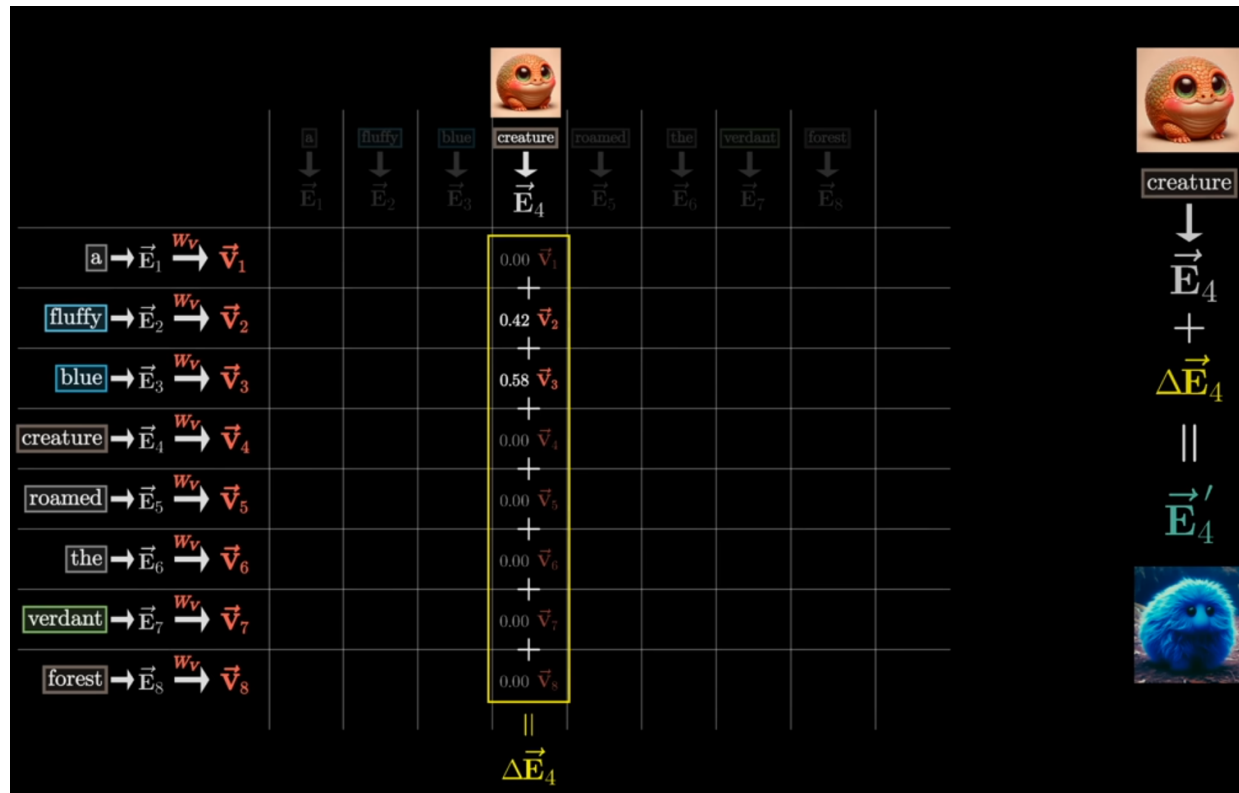
Mascheramento Causale



Vettori di Valore

Value matrix W_V		<div>a</div> ↓ \vec{E}_1	<div>fluffy</div> ↓ \vec{E}_2	<div>blue</div> ↓ \vec{E}_3	<div>creature</div> ↓ \vec{E}_4	<div>roamed</div> ↓ \vec{E}_5	<div>the</div> ↓ \vec{E}_6	<div>verdant</div> ↓ \vec{E}_7	<div>forest</div> ↓ \vec{E}_8	
<div>a</div> → \vec{E}_1 → W_V → \vec{V}_1		●	●	●	●	●	●	●	●	
<div>fluffy</div> → \vec{E}_2 → W_V → \vec{V}_2			●	●	●	●	●	●	●	
<div>blue</div> → \vec{E}_3 → W_V → \vec{V}_3				●	●	●	●	●	●	
<div>creature</div> → \vec{E}_4 → W_V → \vec{V}_4					●	●	●	●	●	
<div>roamed</div> → \vec{E}_5 → W_V → \vec{V}_5						●	●	●	●	
<div>the</div> → \vec{E}_6 → W_V → \vec{V}_6							●	●	●	
<div>verdant</div> → \vec{E}_7 → W_V → \vec{V}_7								●	●	
<div>forest</div> → \vec{E}_8 → W_V → \vec{V}_8									●	

Aggiornamento Delta



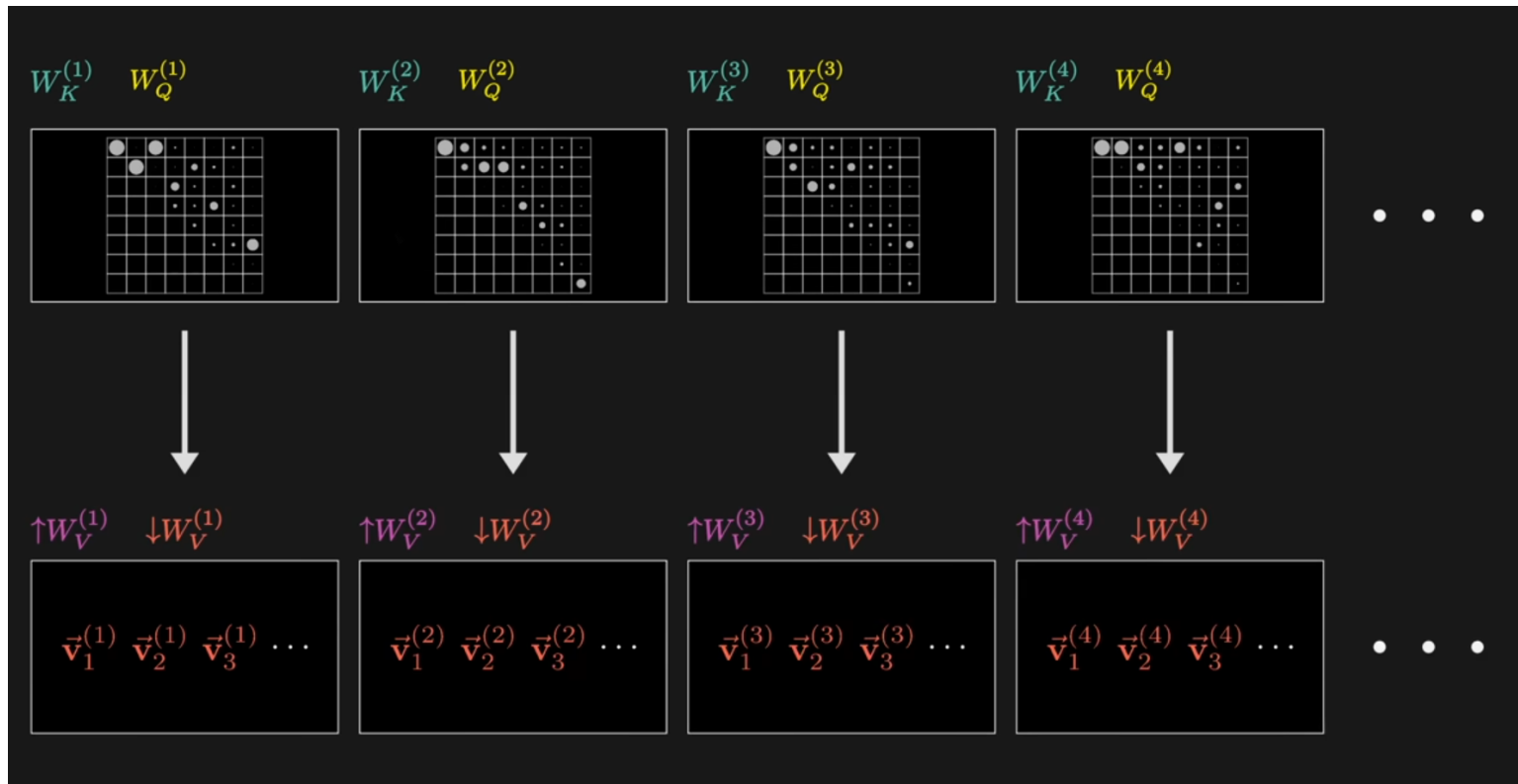
Attention Head

One head of attention

	a	fluffy	blue	creature	roamed	the	verdant	forest
	\vec{E}_1	\vec{E}_2	\vec{E}_3	\vec{E}_4	\vec{E}_5	\vec{E}_6	\vec{E}_7	\vec{E}_8
$\vec{E}_1 \xrightarrow{w_v} \vec{v}_1$	1.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1	0.00 \vec{v}_1
$\vec{E}_2 \xrightarrow{w_v} \vec{v}_2$	0.00 \vec{v}_2	1.00 \vec{v}_2	0.00 \vec{v}_2	0.42 \vec{v}_2	0.00 \vec{v}_2	0.00 \vec{v}_2	0.00 \vec{v}_2	0.00 \vec{v}_2
$\vec{E}_3 \xrightarrow{w_v} \vec{v}_3$	0.00 \vec{v}_3	0.00 \vec{v}_3	1.00 \vec{v}_3	0.58 \vec{v}_3	0.00 \vec{v}_3	0.00 \vec{v}_3	0.00 \vec{v}_3	0.00 \vec{v}_3
$\vec{E}_4 \xrightarrow{w_v} \vec{v}_4$	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4	0.00 \vec{v}_4
$\vec{E}_5 \xrightarrow{w_v} \vec{v}_5$	0.00 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5	0.01 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5	0.00 \vec{v}_5
$\vec{E}_6 \xrightarrow{w_v} \vec{v}_6$	0.00 \vec{v}_6	0.00 \vec{v}_6	0.00 \vec{v}_6	0.00 \vec{v}_6	0.99 \vec{v}_6	1.00 \vec{v}_6	0.00 \vec{v}_6	0.00 \vec{v}_6
$\vec{E}_7 \xrightarrow{w_v} \vec{v}_7$	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	0.00 \vec{v}_7	1.00 \vec{v}_7	1.00 \vec{v}_7
$\vec{E}_8 \xrightarrow{w_v} \vec{v}_8$	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8	0.00 \vec{v}_8
	$\Delta \vec{E}_1$	$\Delta \vec{E}_2$	$\Delta \vec{E}_3$	$\Delta \vec{E}_4$	$\Delta \vec{E}_5$	$\Delta \vec{E}_6$	$\Delta \vec{E}_7$	$\Delta \vec{E}_8$

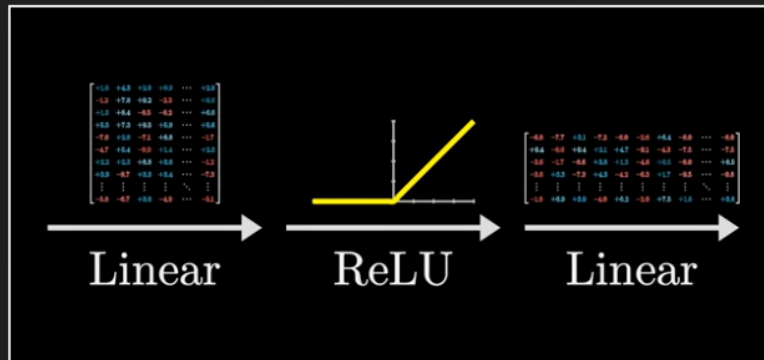
\vec{E}_1	\vec{E}_2	\vec{E}_3	\vec{E}_4	\vec{E}_5	\vec{E}_6	\vec{E}_7	\vec{E}_8
+	+	+	+	+	+	+	+
$\Delta \vec{E}_1$	$\Delta \vec{E}_2$	$\Delta \vec{E}_3$	$\Delta \vec{E}_4$	$\Delta \vec{E}_5$	$\Delta \vec{E}_6$	$\Delta \vec{E}_7$	$\Delta \vec{E}_8$
\vec{E}'_1	\vec{E}'_2	\vec{E}'_3	\vec{E}'_4	\vec{E}'_5	\vec{E}'_6	\vec{E}'_7	\vec{E}'_8

Multi-Head Attention

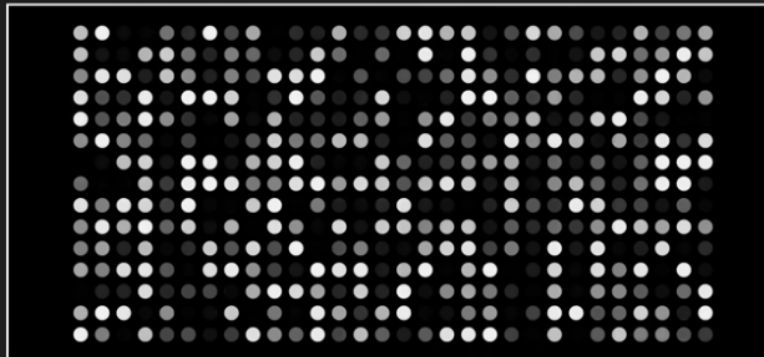


Multi-Layer Perceptron

Panoramica MLP



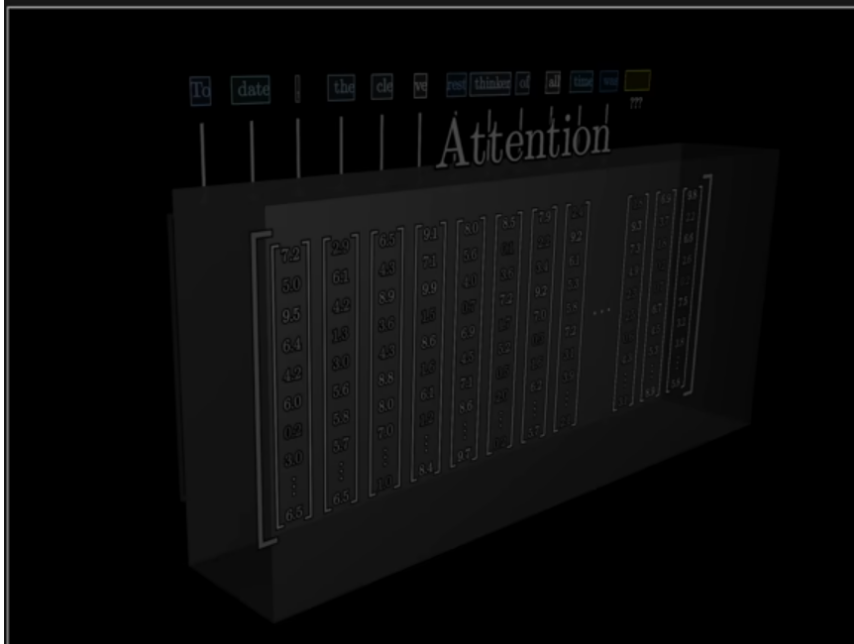
Structure:
Easy



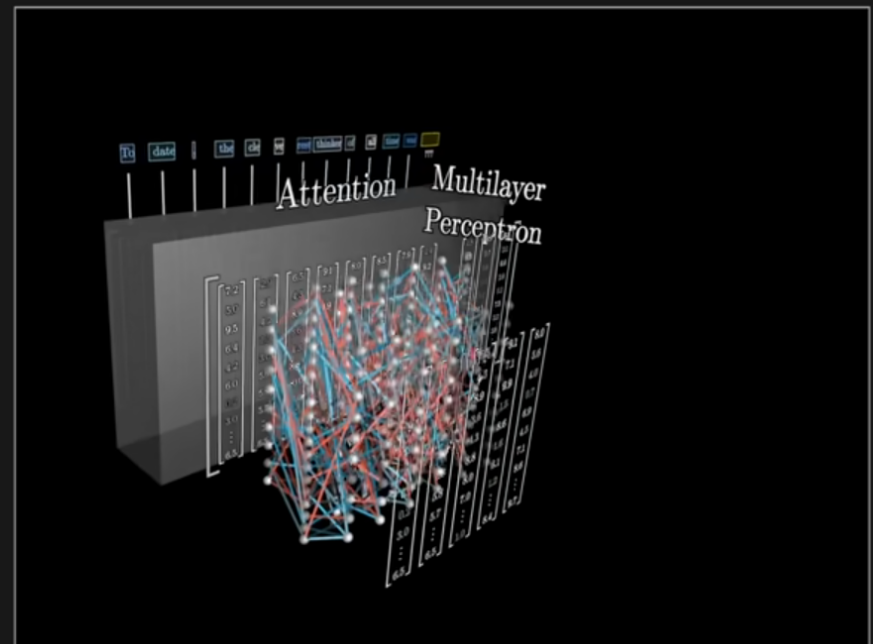
Emergent behavior:
Exceedingly challenging

Struttura MLP

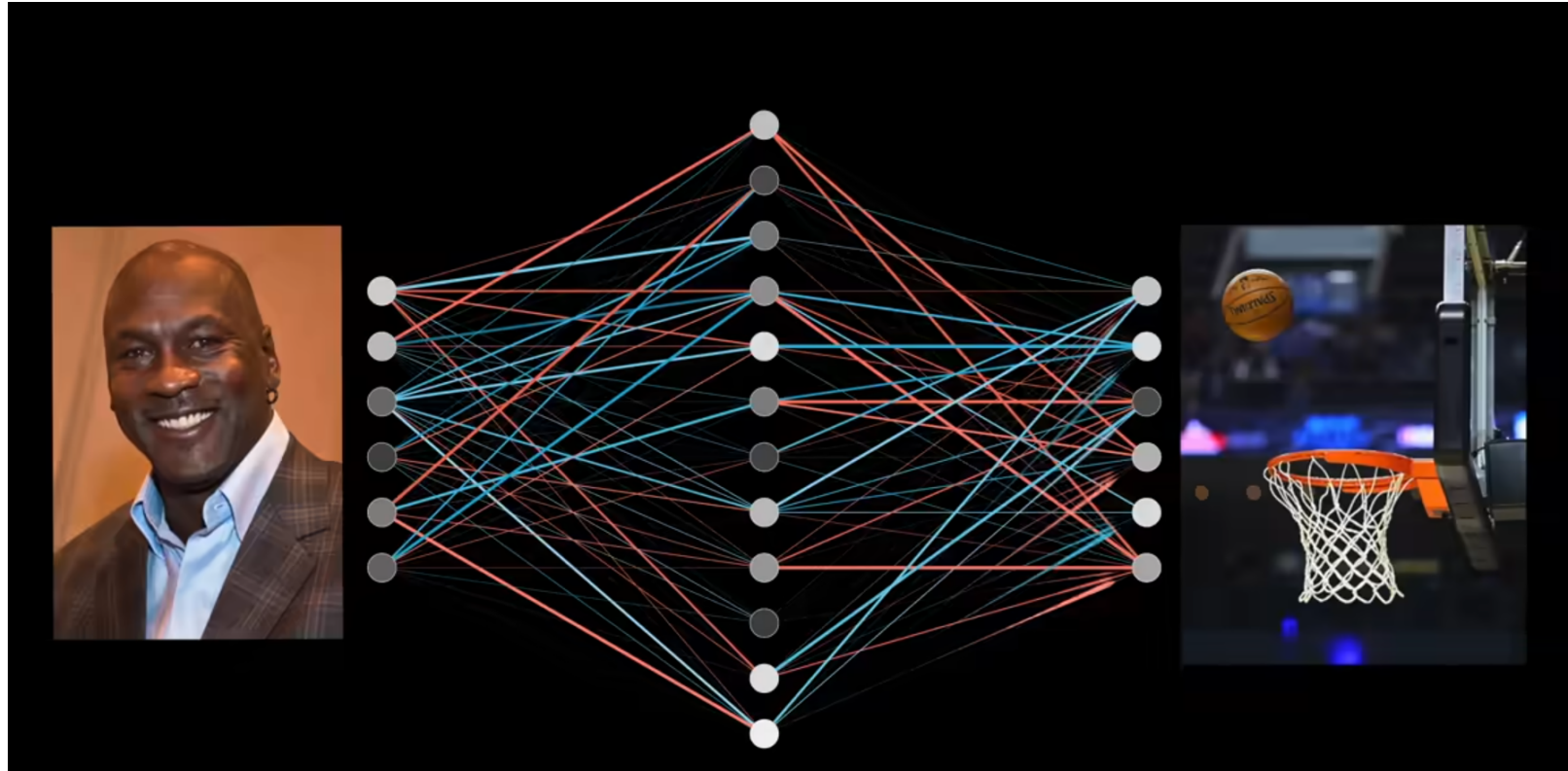
≈ 1/3 of the Parameters



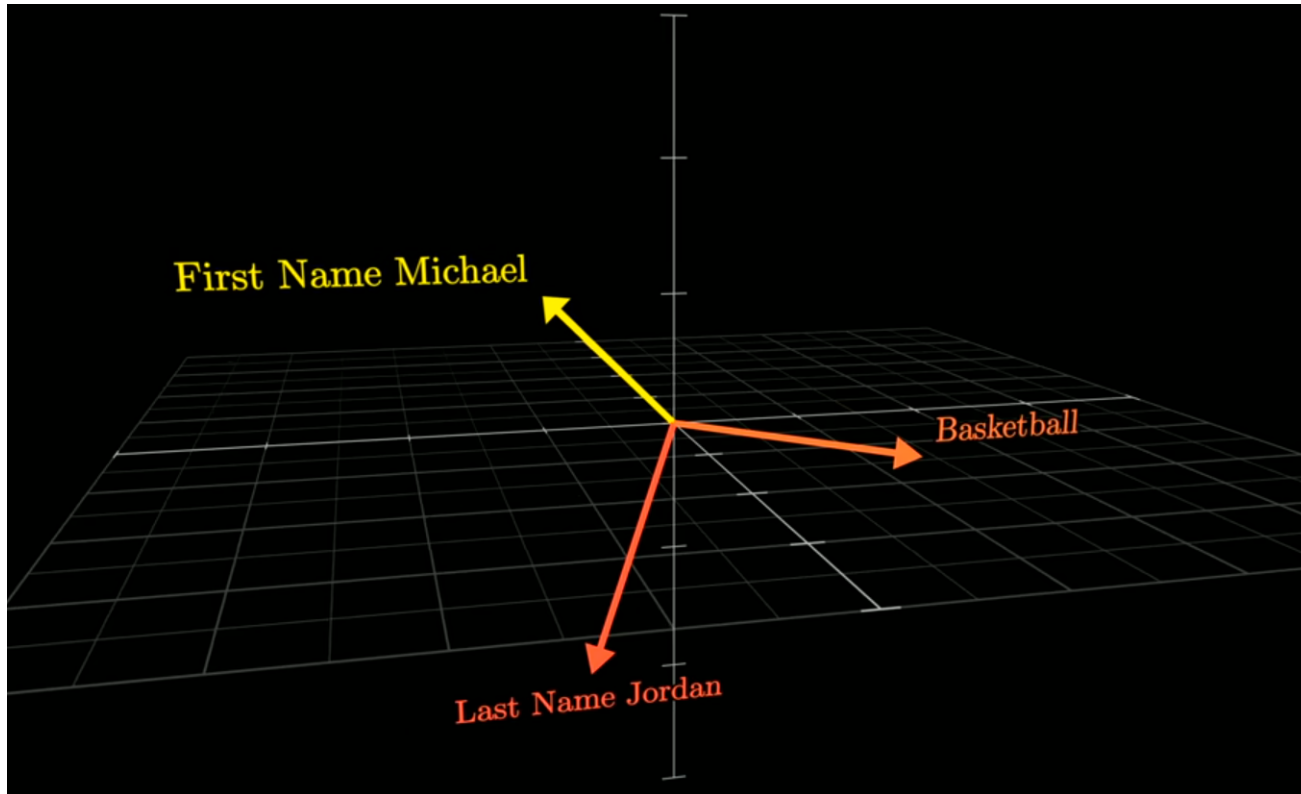
$\approx 2/3$ of the Parameters



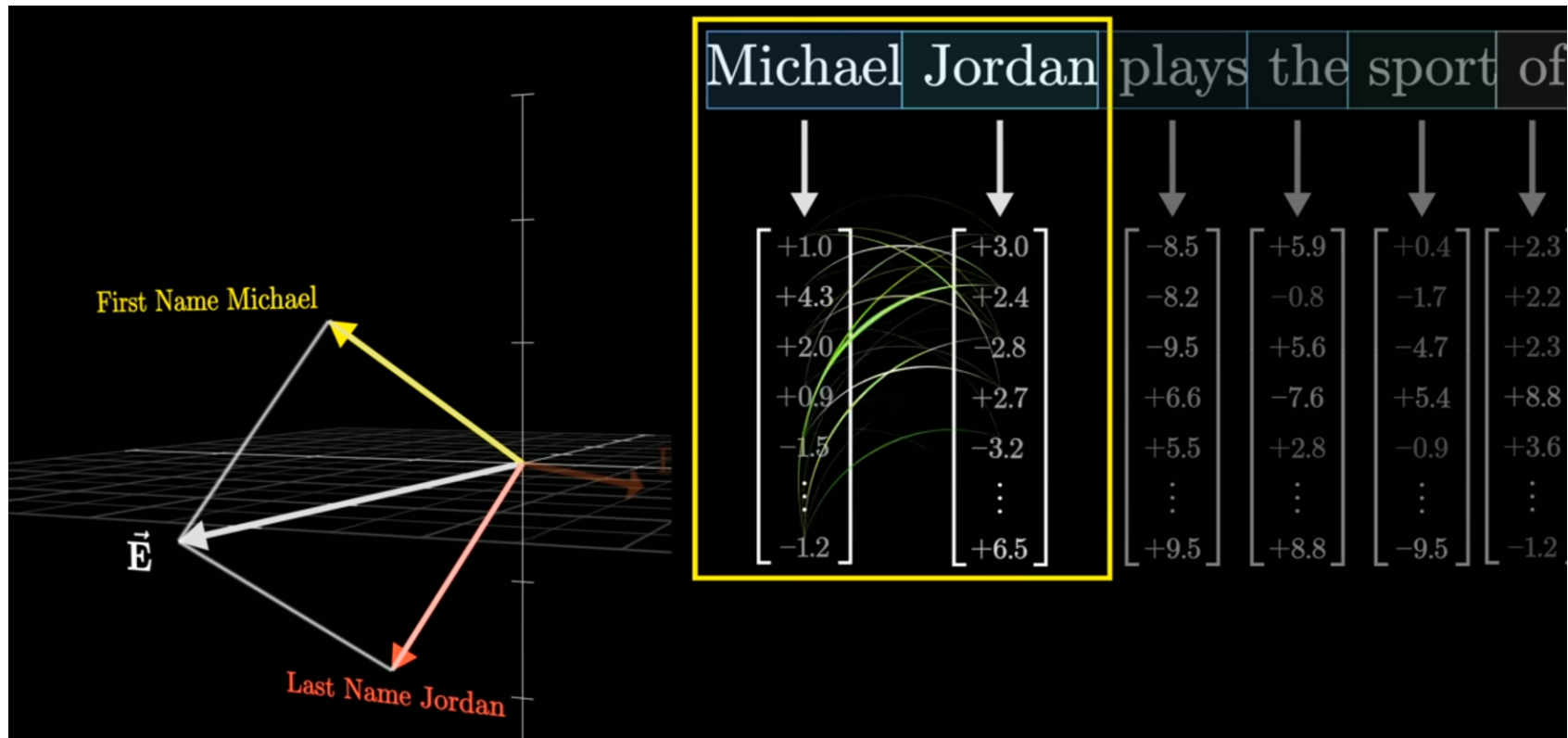
Fatti Memorizzati



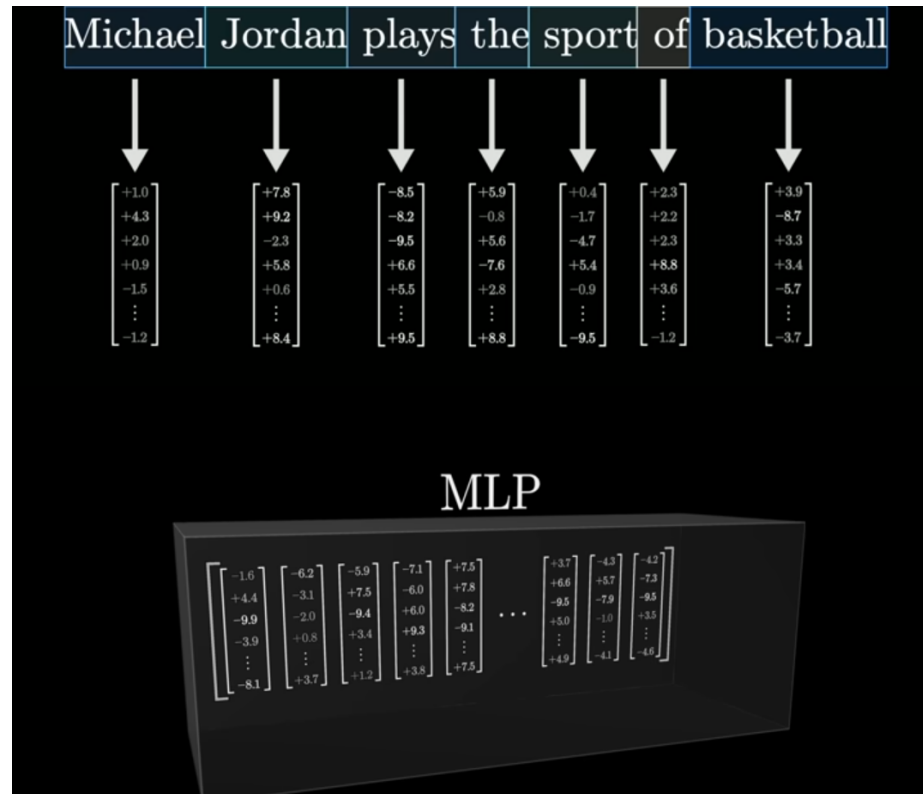
Direzioni dei Nomi



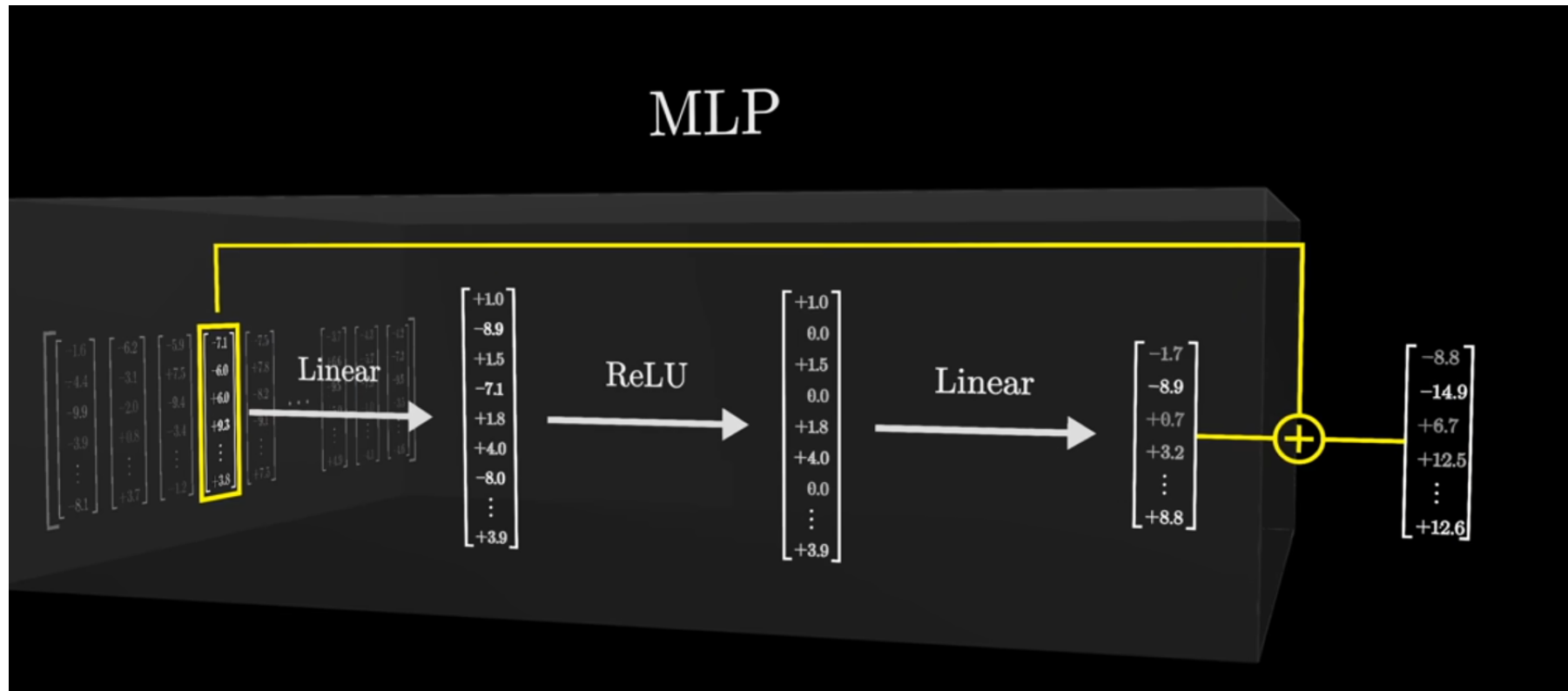
Nome Completo



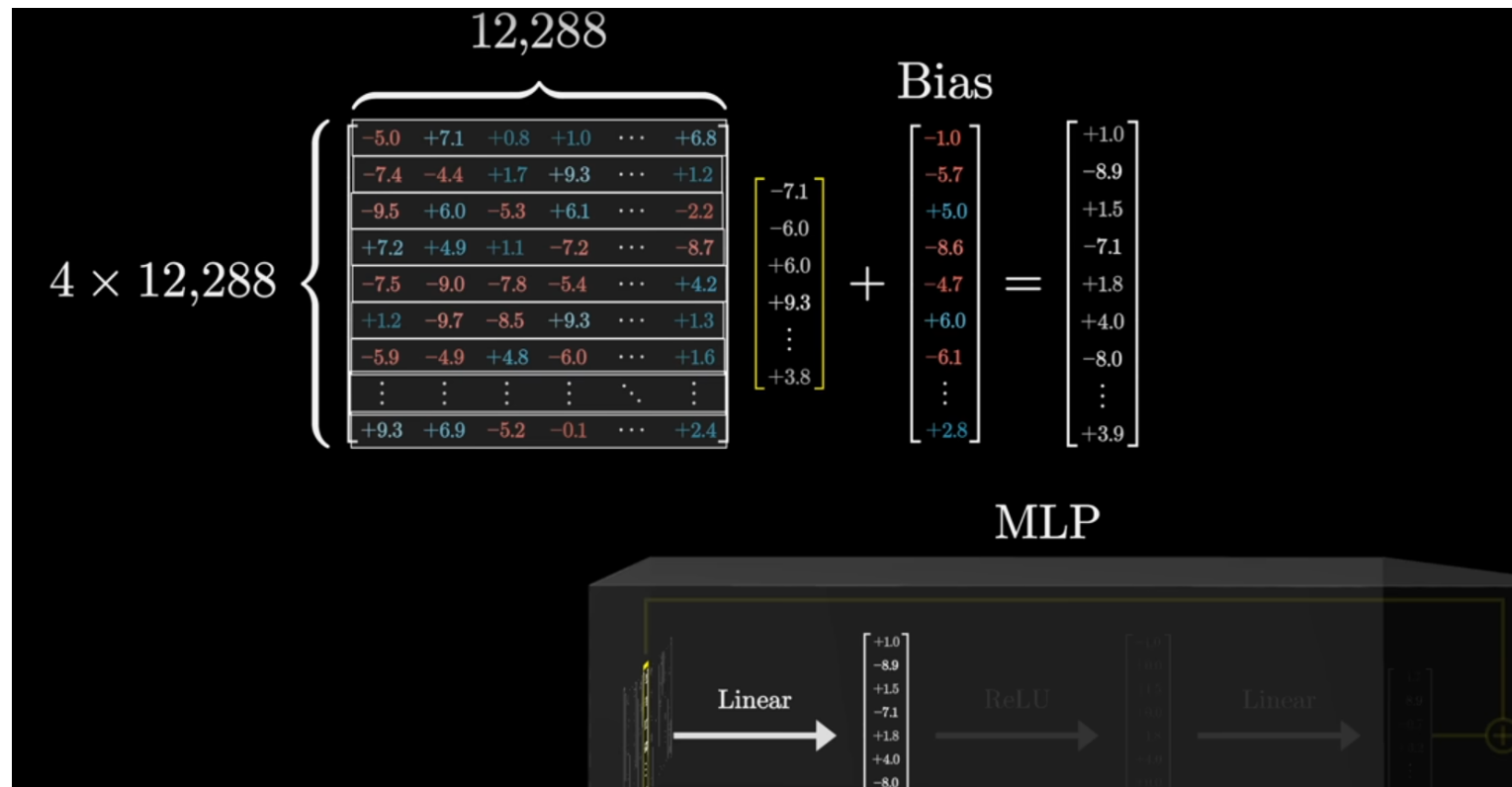
Operazioni MLP



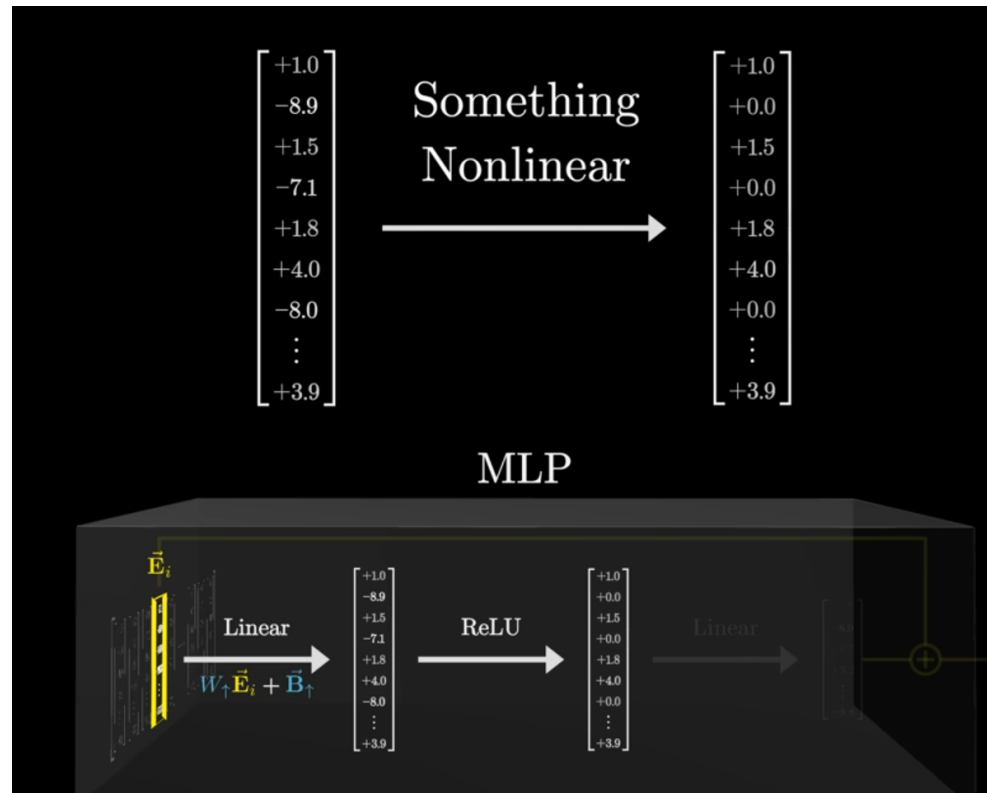
Addizione Residua



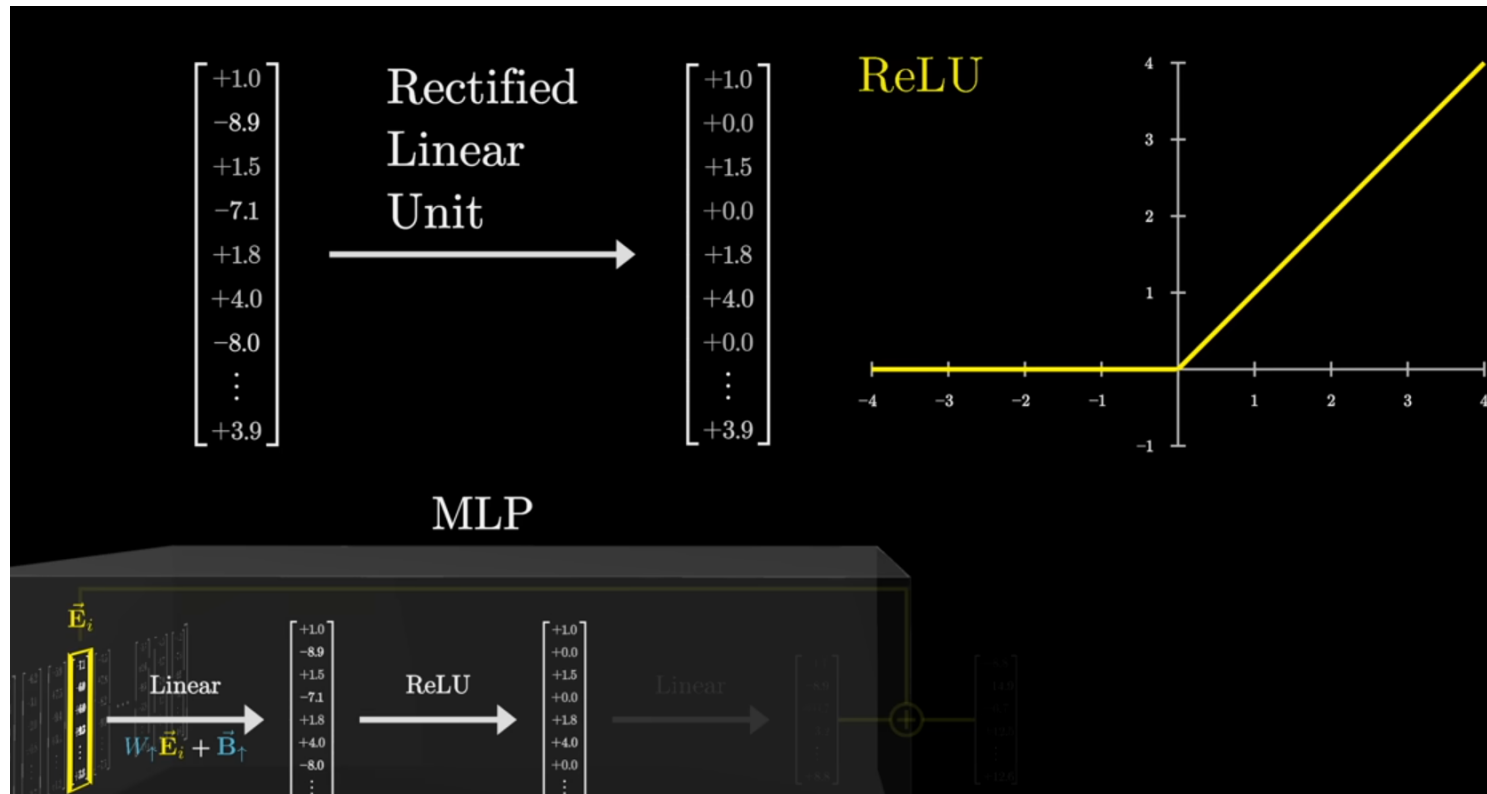
Proiezione Verso l'Alto



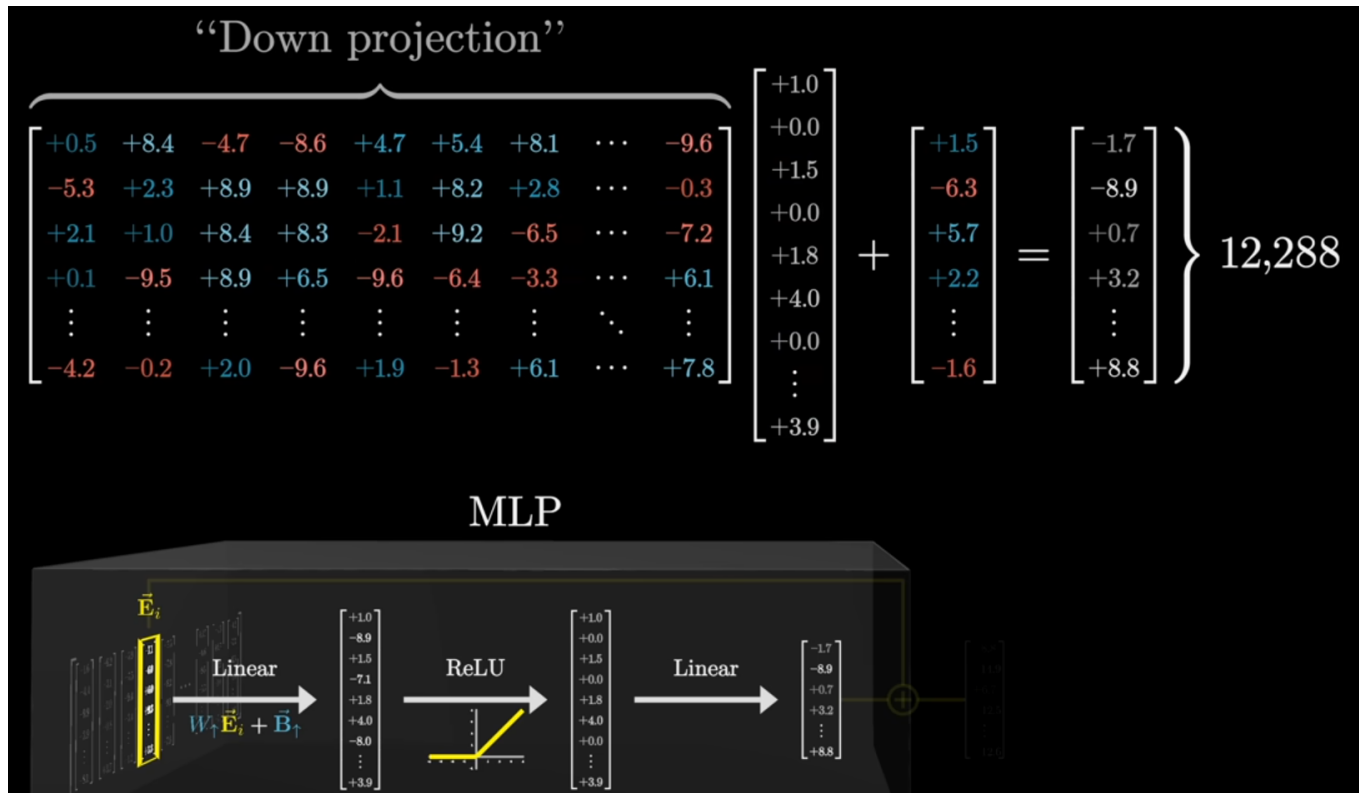
Attivazione Non Lineare



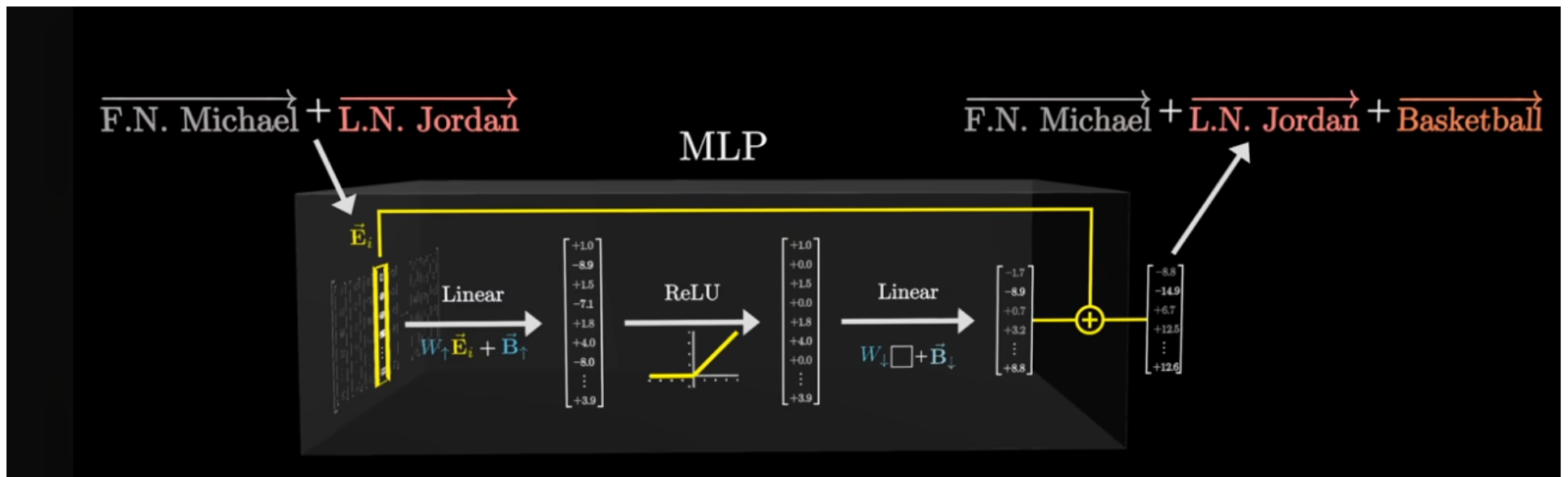
Funzione ReLU



Proiezione Verso il Basso



Output Finale



Riepilogo Finale

- 1 Il **testo** viene scomposto in **token** e convertito in **vettori numerici** (embedding)
- 2 I blocchi di **attenzione** permettono ai token di scambiarsi informazioni contestuali
- 3 I blocchi **MLP** memorizzano e recuperano fatti appresi durante l'addestramento
- 4 Questi blocchi si ripetono decine di volte, raffinando progressivamente la rappresentazione
- 5 Alla fine, il modello produce una **distribuzione di probabilità** sulla prossima parola

Sviluppi Recenti

La Corsa ai Modelli

Anno	OpenAI	Anthropic	Google	Meta
2023	GPT-4	Claude 2	Gemini 1.0	Llama 2
2024	GPT-4o, o1	Claude 3/3.5	Gemini 1.5	Llama 3 (405B)
2025	o3, GPT-5	Claude Opus 4	Gemini 2.0	Llama 4

Modelli di Ragionamento

I modelli di ragionamento “pensano prima di rispondere”, generando una catena di pensiero interna prima di produrre l’output finale.

LLM Open Source

- **Llama 3** (Meta, 2024): fino a 405 miliardi di parametri, liberamente disponibile
- **Mistral / Mixtral** (Mistral AI): modelli europei ad alte prestazioni
- **DeepSeek V3/R1** (DeepSeek, 2025): modelli cinesi open source competitivi con GPT-4

Benchmark 2026

I modelli attuali superano regolarmente:

- **Esame di avvocato** (Bar Exam): top 10%
- **Esame medico** (USMLE): punteggi da specialista
- **Programmazione competitiva**: livello esperto su Codeforces
- **Matematica olimpica**: medaglia d'oro alle IMO (o3)

Limiti Attuali

- **Allucinazioni:** generano informazioni false con apparente sicurezza
- **Finestra di contesto:** anche con 1M+ token, la qualità degrada su testi lunghi
- **Knowledge cutoff:** conoscenza limitata alla data di addestramento
- **Errori di ragionamento:** possono fallire su problemi logici che sembrano banali

Il Futuro: Agenti e Strumenti

Un agente IA è un sistema che utilizza un LLM per **ragionare**, **pianificare** e **agire** autonomamente, interagendo con strumenti esterni e ambienti digitali.

Crediti

Molte delle visualizzazioni e spiegazioni in questa presentazione sono ispirate alla serie di video di **3Blue1Brown** (Grant Sanderson):

- Neural Networks
- Gradient Descent
- Backpropagation
- GPT: What is a Transformer?
- Attention in Transformers