Transformers

Guillem & Roderic, Summer 2025

Transformers

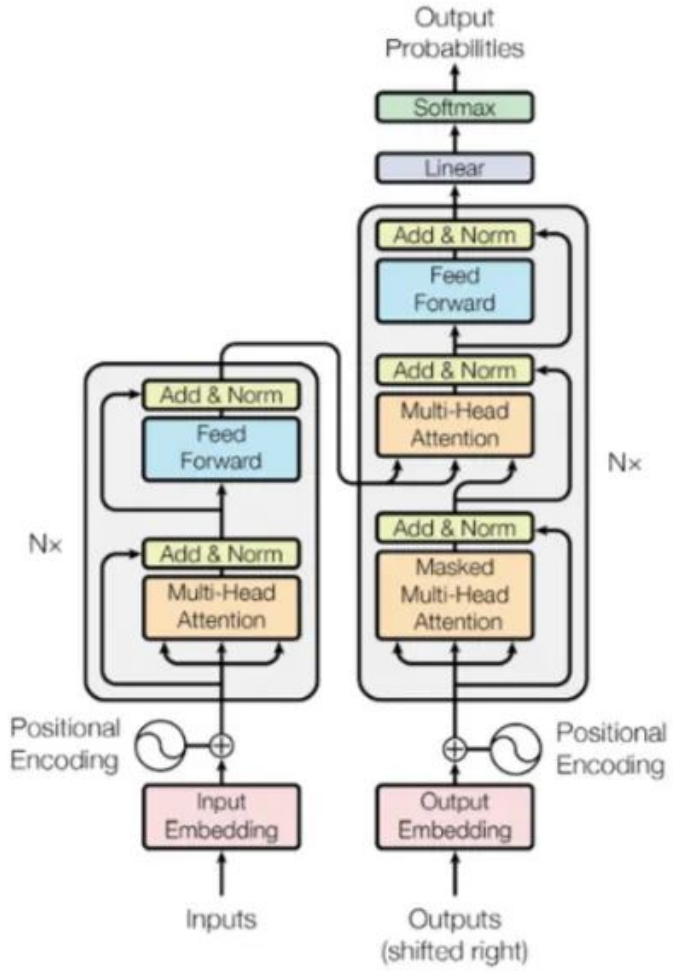
&

ticons

Transformers

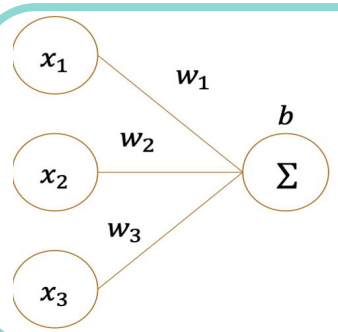
Transformer

Attention Is All You Need

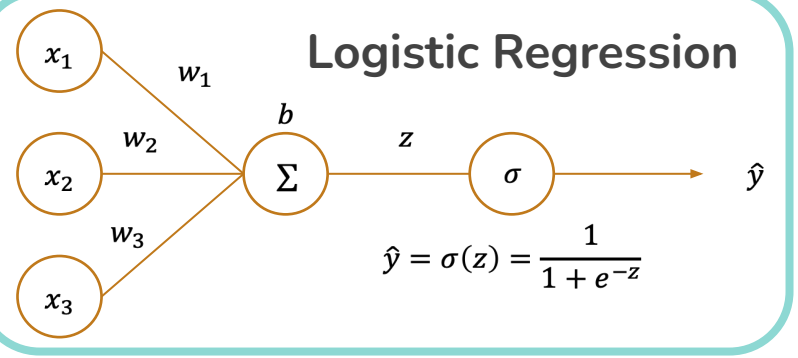


Recap:

Linear Regression

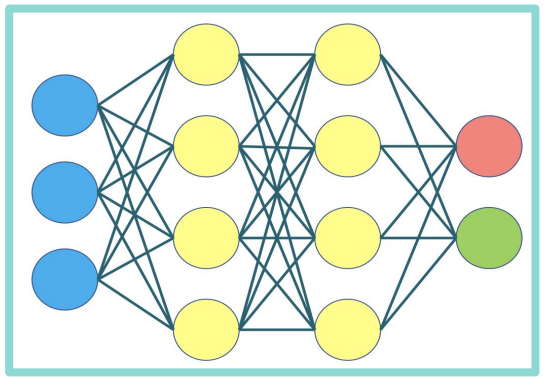


Logistic Regression



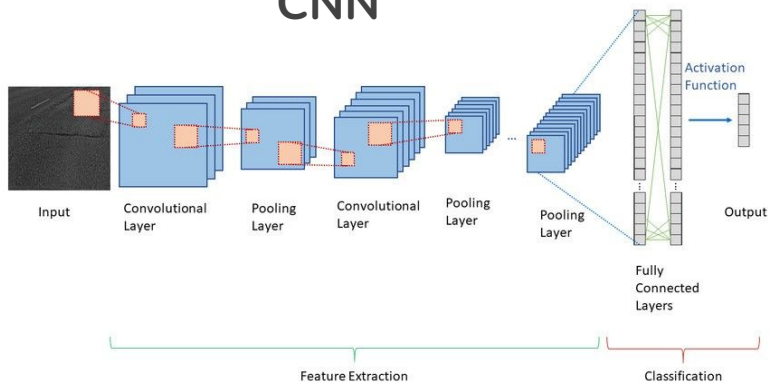
$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Neural Networks

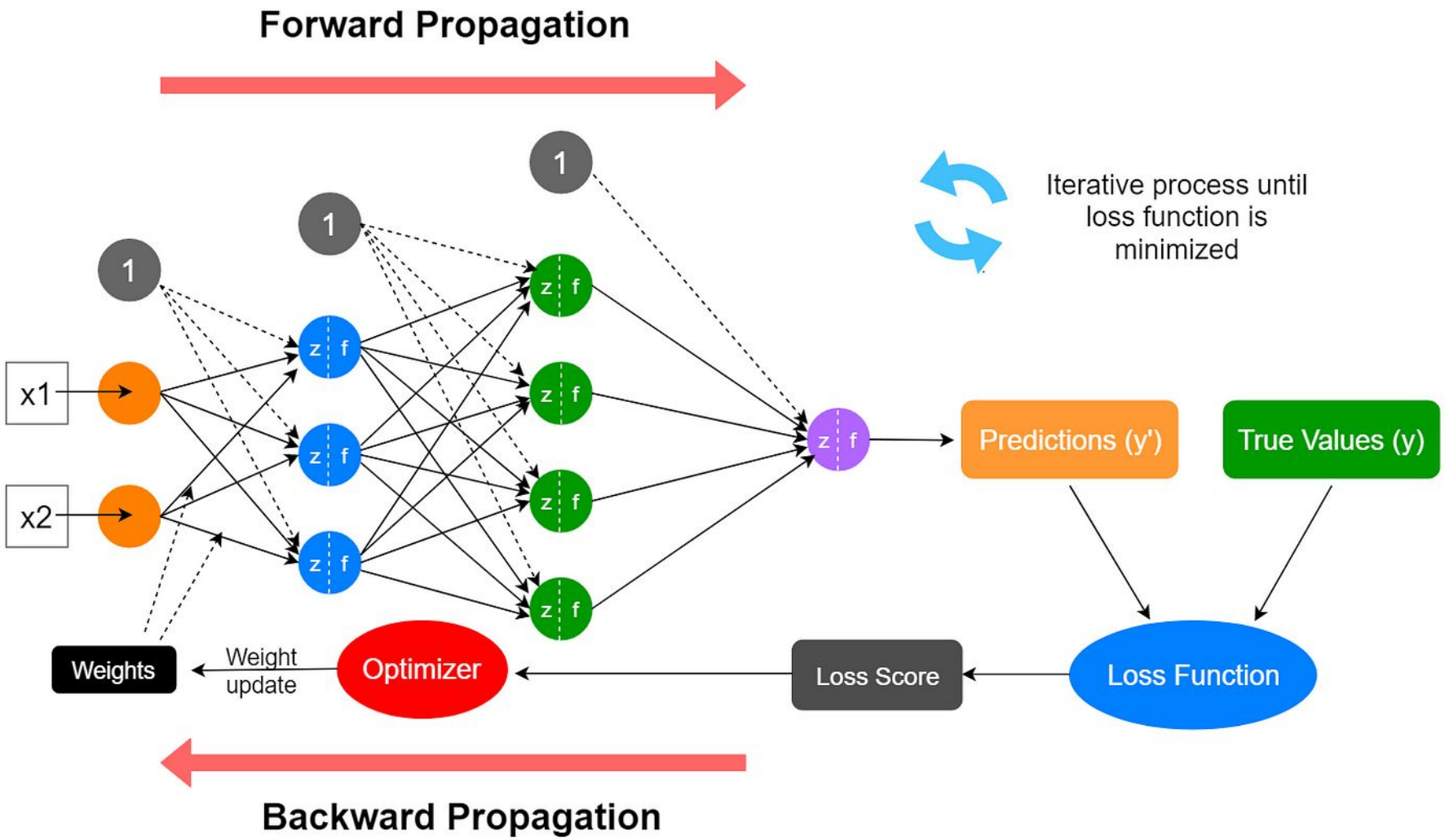


MLP

CNN

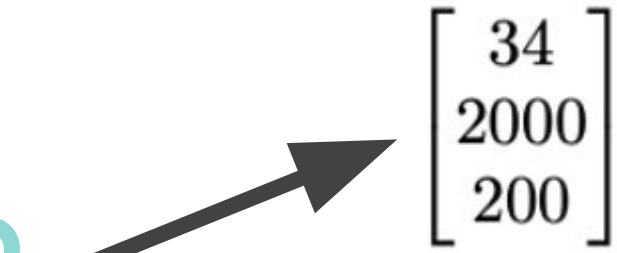


Deep learning training



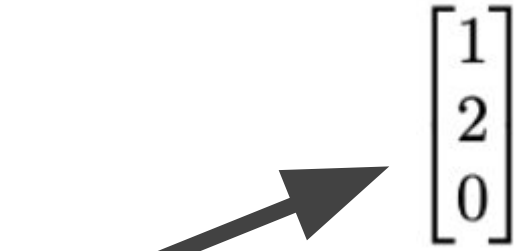
Vectorization: Numerical Data

Age	Monthly Salary	Monthly Savings
23	1200	200
34	2000	200
55	1500	300
⋮	⋮	⋮



Vectorization: Categorical Data

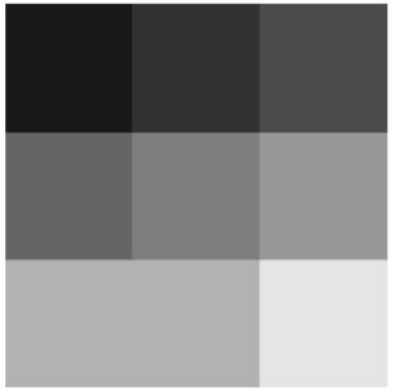
Sex	Hair Color	Smoker
M	Black	Yes
F	Brown	No
F	Black	No
⋮	⋮	⋮



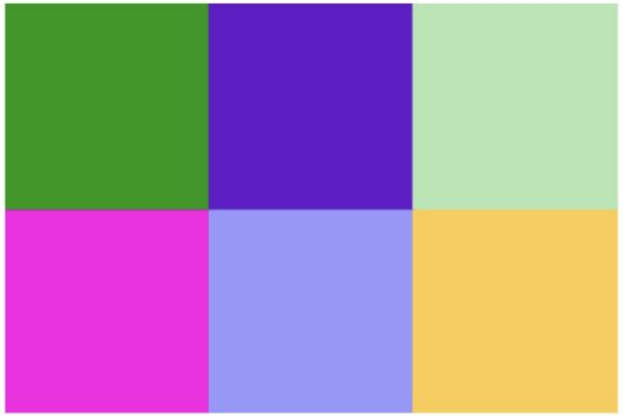
$$\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

Blond:0, Black:1, Brown: 2, Blue: 3, ...

Vectorization: Images

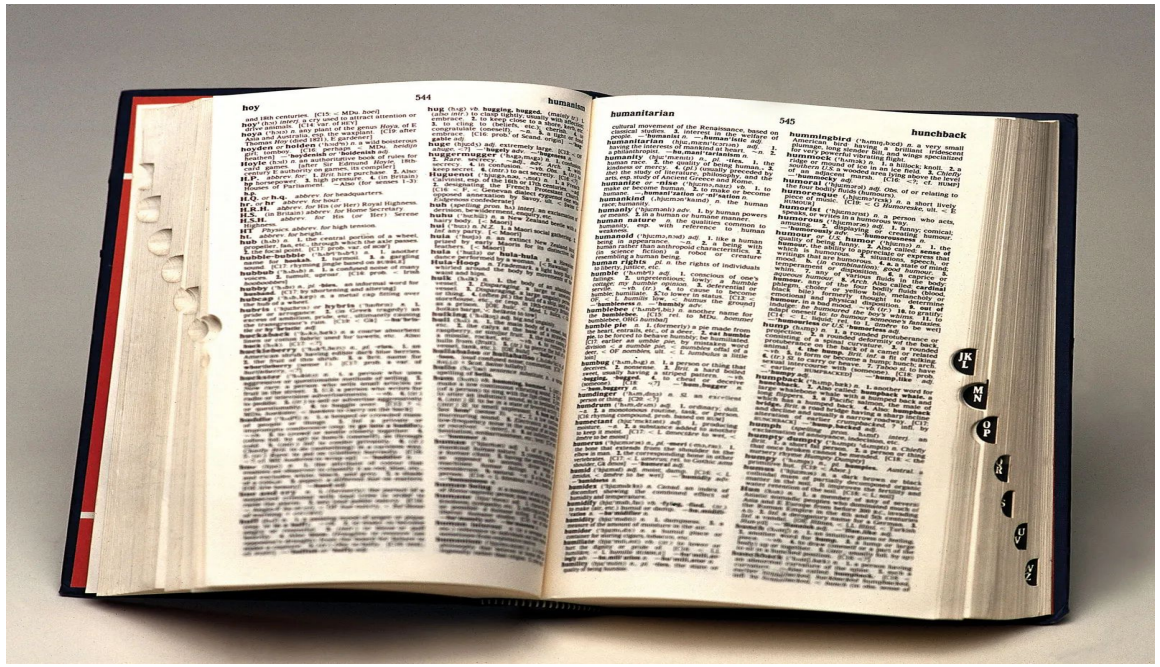


$$\begin{bmatrix} 25 & 51 & 76 \\ 102 & 127 & 153 \\ 178 & 178 & 229 \end{bmatrix}$$



$$\begin{bmatrix} \begin{bmatrix} 0 \\ 153 \\ 0 \end{bmatrix} & \begin{bmatrix} 102 \\ 25 \\ 204 \end{bmatrix} & \begin{bmatrix} 178 \\ 229 \\ 178 \end{bmatrix} \\ \begin{bmatrix} 255 \\ 0 \\ 229 \end{bmatrix} & \begin{bmatrix} 153 \\ 153 \\ 255 \end{bmatrix} & \begin{bmatrix} 255 \\ 204 \\ 76 \end{bmatrix} \end{bmatrix}$$

Vectorization: Words & Text



List of Words in English Dictionary:

A
Aardvark
Abeam
Abacus
Abandon
Abase
...

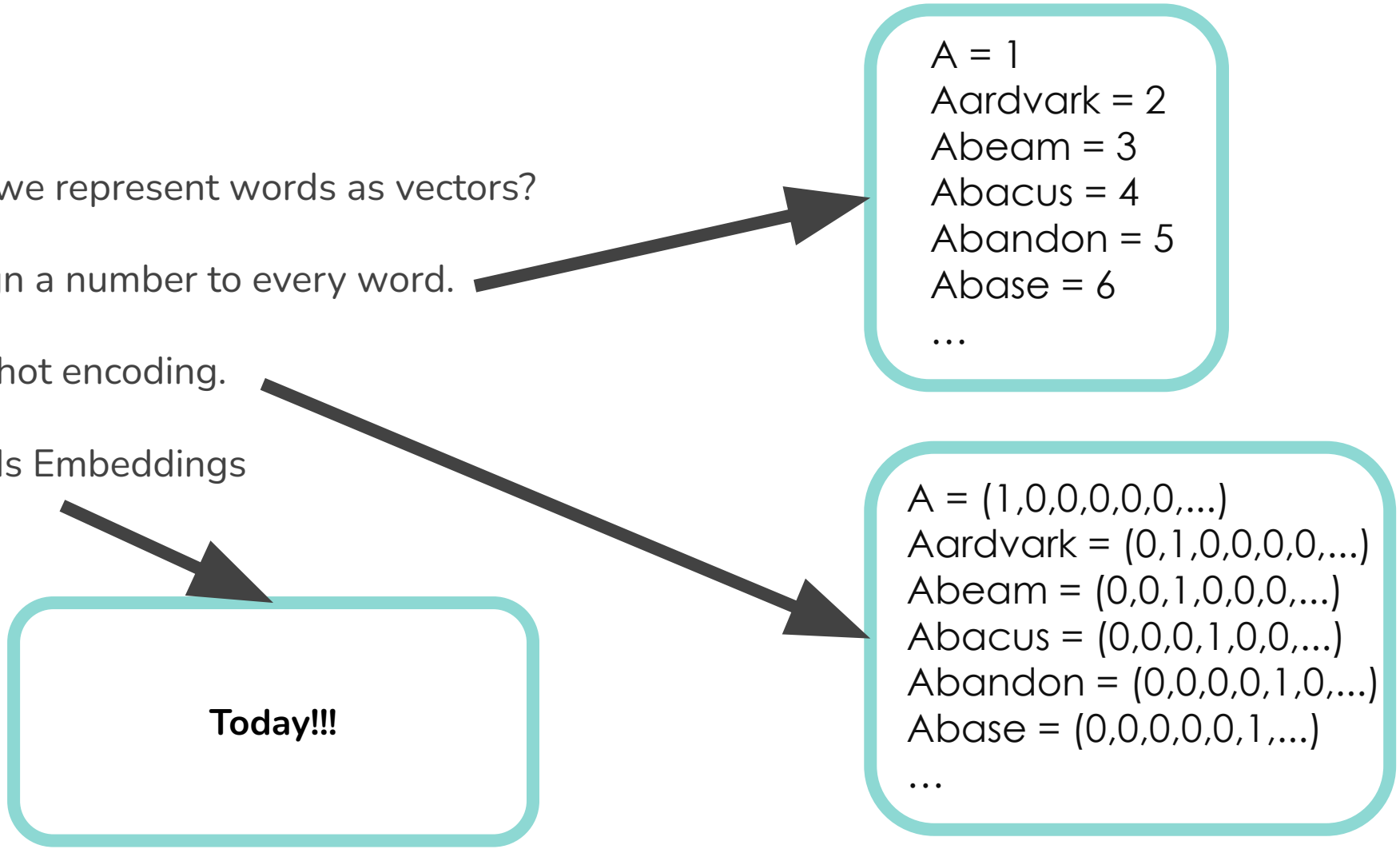
Vectorization: Words & Text

How can we represent words as vectors?

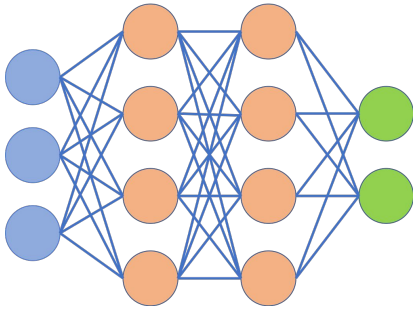
1) Assign a number to every word.

2) One-hot encoding.

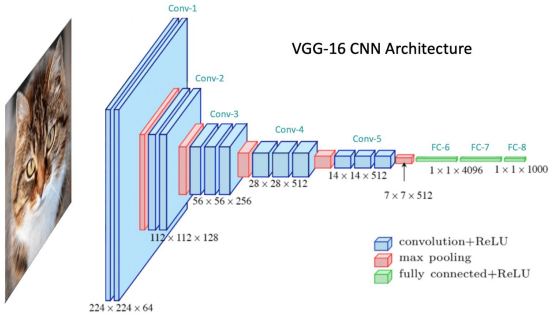
3) Words Embeddings



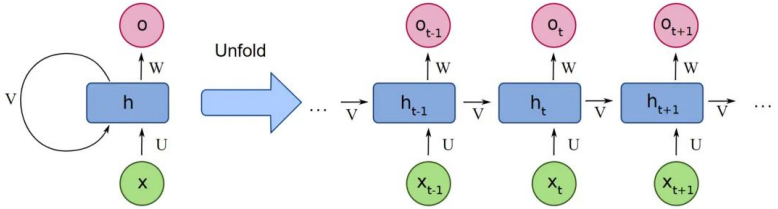
Neural Network architectures



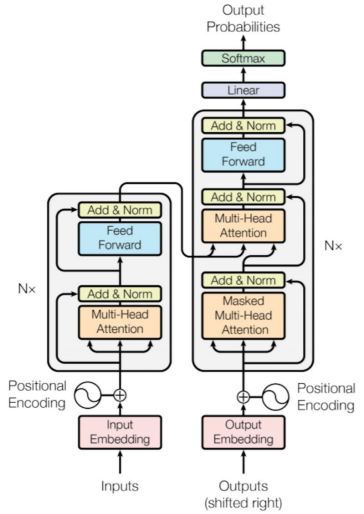
fully connected/linear/dense layers or MLP →
linear map + non-linear activation functions



convolutional layers → CNNs
computer vision, signal processing, etc.



recurrent layers → RNNs, LSTMs, GRUs
time series, natural language processing

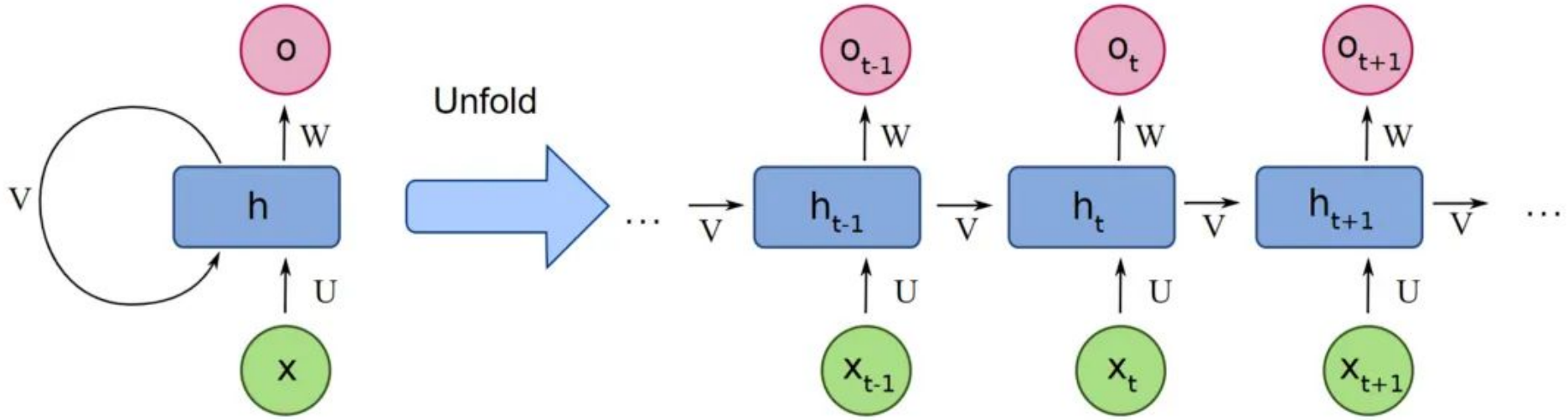


Transformers
Seq to seq models
Multimodal models
Large language models
Generative models
Etc.

NNs in Biology

Type of Data	Example Architecture
Gene expression, single-cell RNA-seq	Fully Connected NN
Biological/medical imaging, histopathology	Convolutional NN
Biological sequences (DNA, RNA, protein), time series	Recurrent Neural Networks, LSTMs
Genomics, proteomics, biomedical language, and structure prediction	Transformers

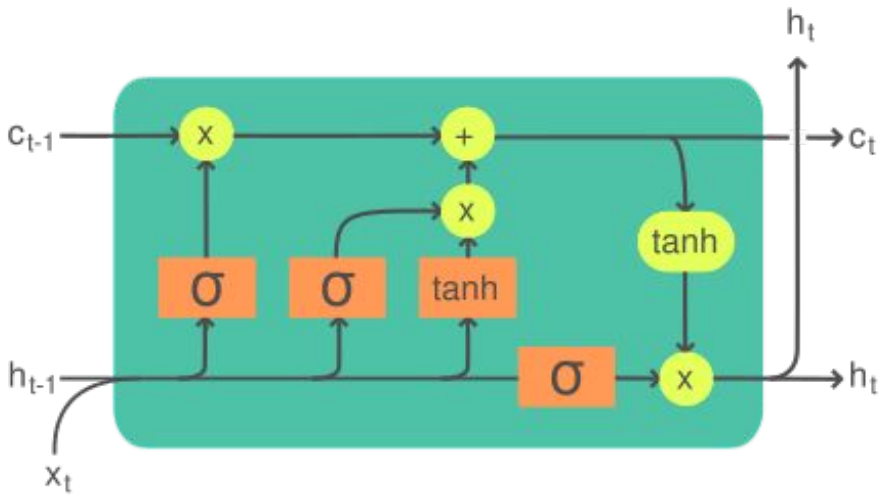
Recurrent Neural Networks (RNNs)



$x \rightarrow$ feature vector
$h \rightarrow$ hidden State
$U, V, W \rightarrow$ weight matrices
$o \rightarrow$ output

problem: vanishing/exploding gradients (backpropagation through time)
RNNs lose contextual information rapidly

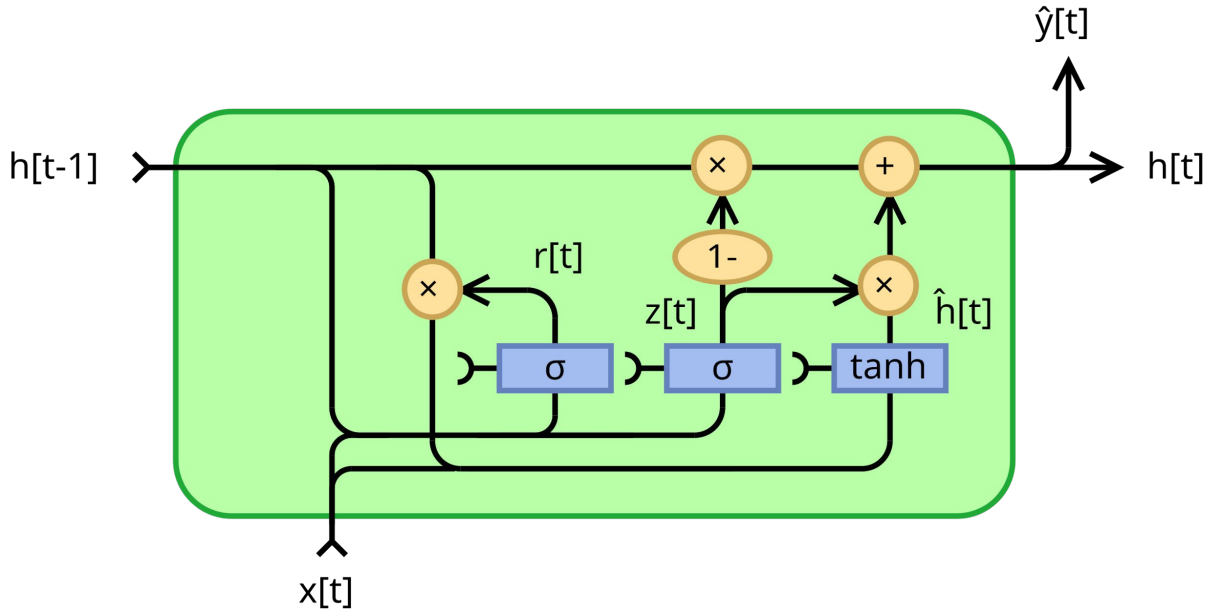
Long Short-Term Memory (LSTMs)



$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \sigma_h(c_t)
\end{aligned}
$$

By adding forget and update gates we can control the information that is retained through time

Gated Recurrent Units (GRUs)

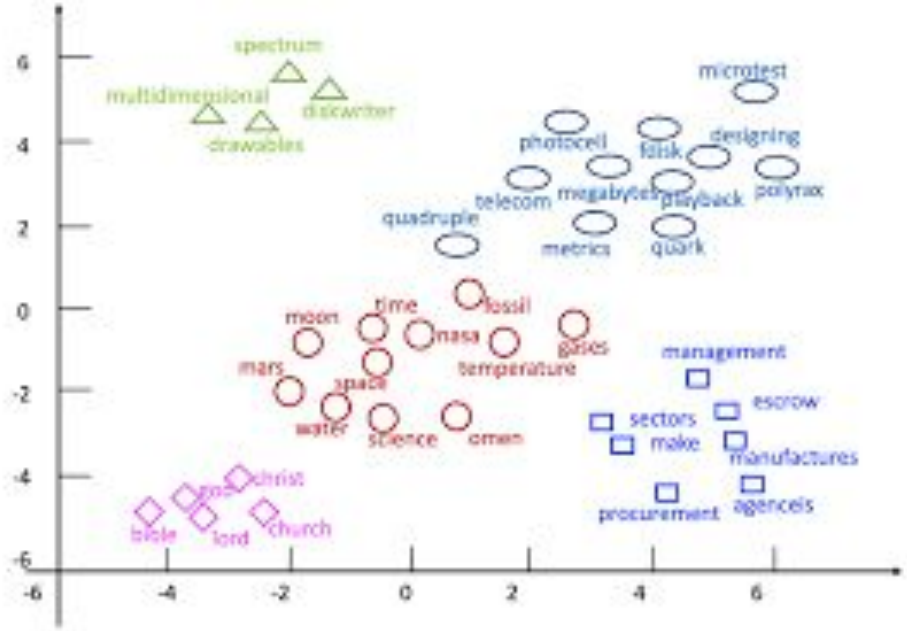


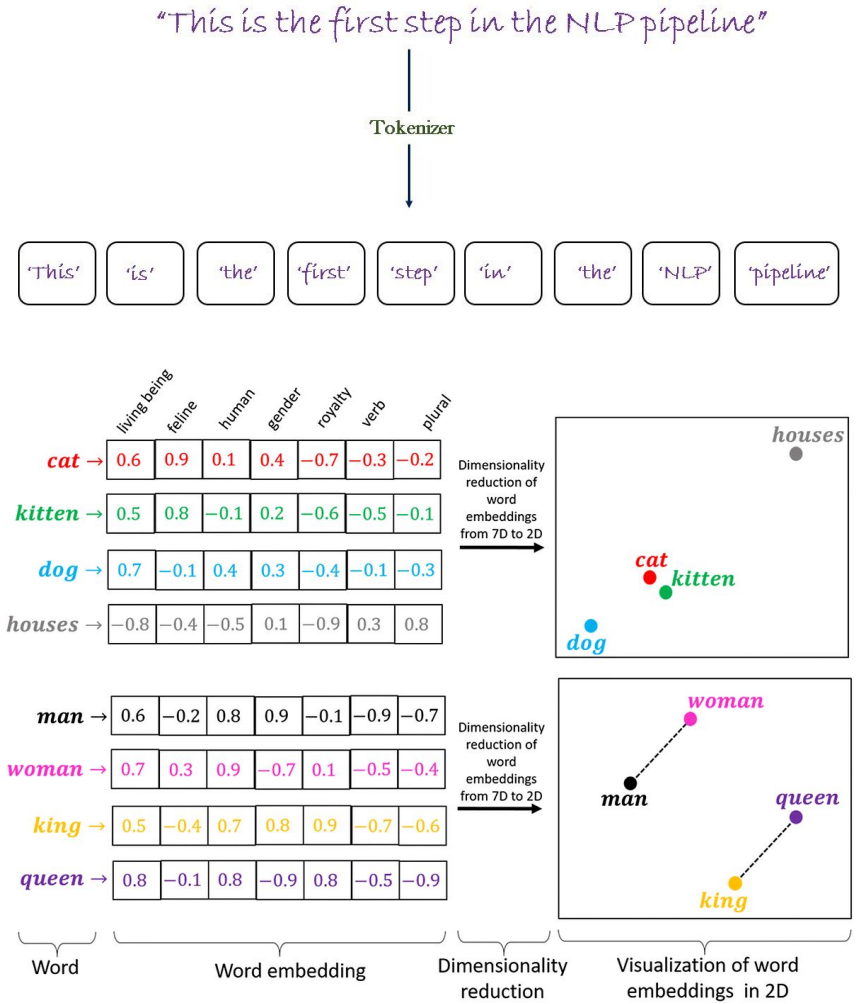
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$\hat{h}_t = \phi(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t$$

Like LSTMs but with less parameters

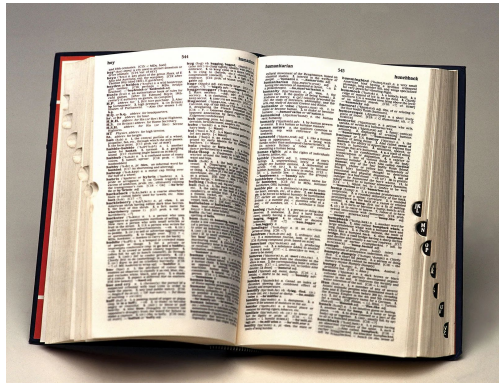
Task: next word prediction

- Tokenization: break text into tokens
- Create word embeddings → words with similar meaning should have similar representations
- Embedding dimension





Task: next word prediction



A
Aardvark
Abeam
Abacus
Abandon
...

Embedding dimension

Vocab size

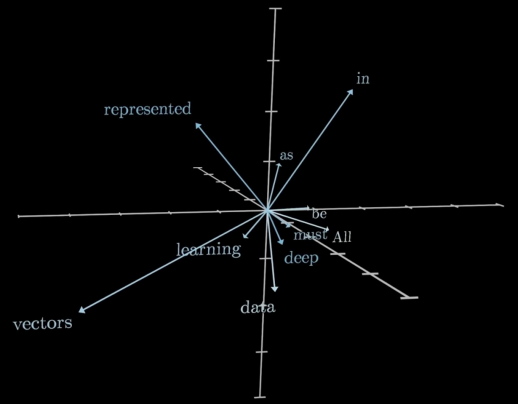
All words, ~ 50k

aah	aardvark	aardwolf	aargh	ab	aback	abacterial	abacus	abalone	abandon	…	zygoid	zygomatic	zygomorphic	zygosis	zygote	zygotic	zyme	zymogen	zymosis	zzz
+1.0	+4.3	+2.0	+0.9	−1.5	+2.9	−1.2	+7.8	+9.2	−2.3	⋯	+0.6	+1.3	+8.4	−8.5	−8.2	−9.5	+6.6	+5.5	+7.3	+9.5
+5.9	−0.8	+5.6	−7.6	+2.8	−7.1	+8.8	+0.4	−1.7	−4.7	⋯	−0.9	+1.4	−9.5	+2.3	+2.2	+2.3	+8.8	+3.6	−2.8	−1.2
+3.9	−8.7	+3.3	+3.4	−5.7	−7.3	−3.7	−2.7	+1.4	−1.2	⋯	−7.9	−5.8	−6.7	+3.0	−4.9	−0.7	−5.1	−6.8	−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

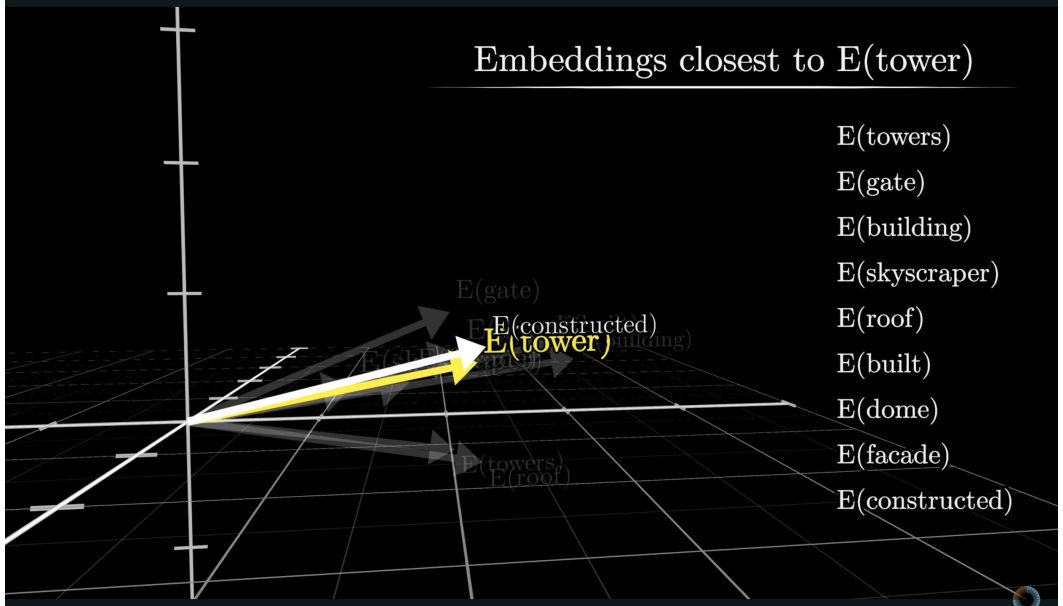
Words "Embedding" Vectors

All
data
in
deep
learning
must
be
represented
as
vectors

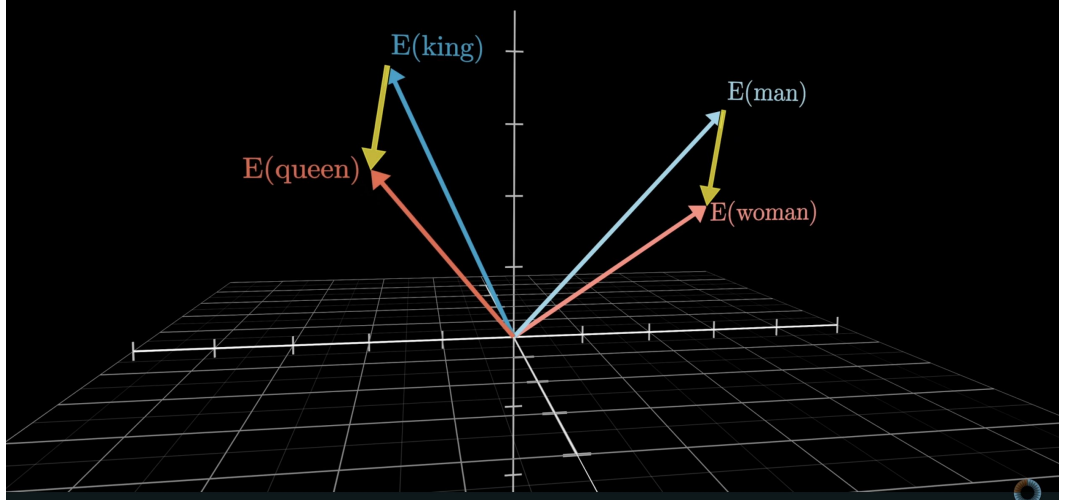


Embeddings closest to E(tower)

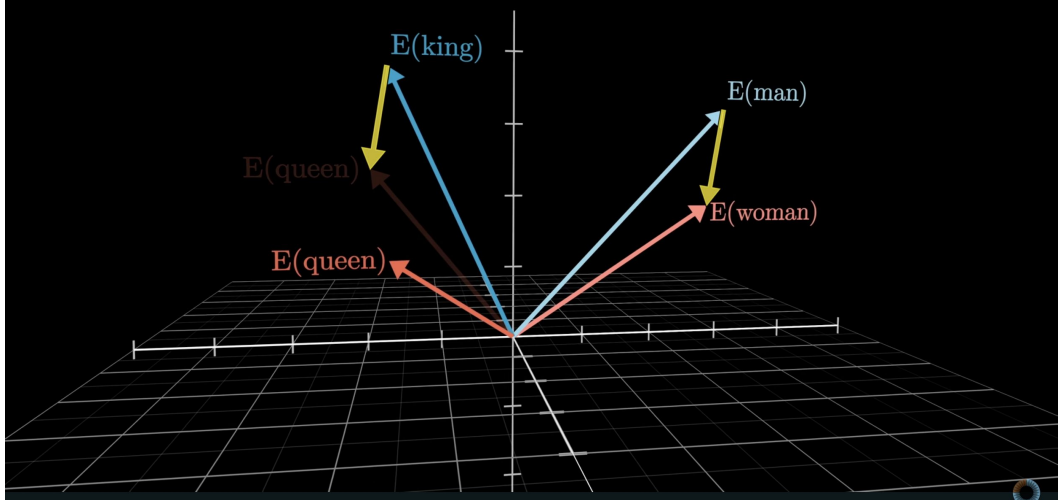
E(towers)
E(gate)
E(building)
E(skyscraper)
E(roof)
E(built)
E(dome)
E(facade)
E(constructed)



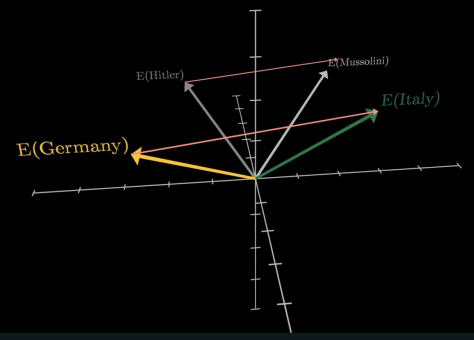
E(queen) ≈ E(king)+E(woman) - E(man)



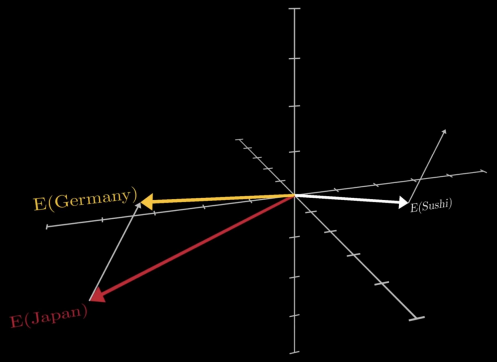
E(queen) ≈ E(king)+E(woman) - E(man)



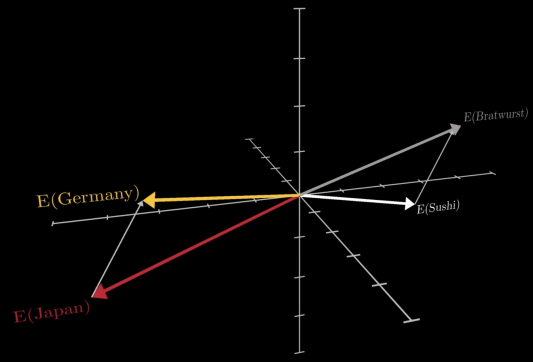
E(Hitler) + E(Italy) - E(Germany) ≈ E(Mussolini)



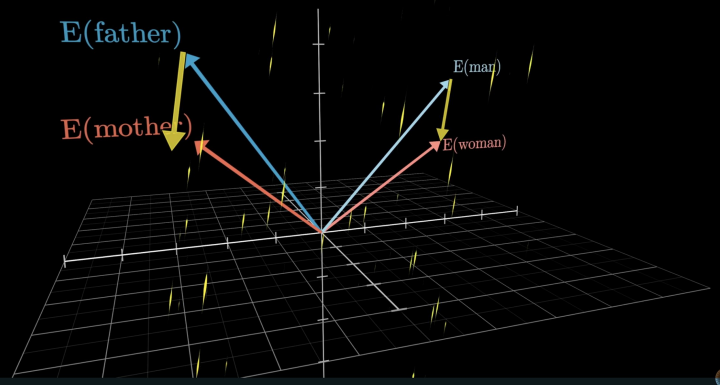
E(Sushi) + E(Germany) - E(Japan) ≈



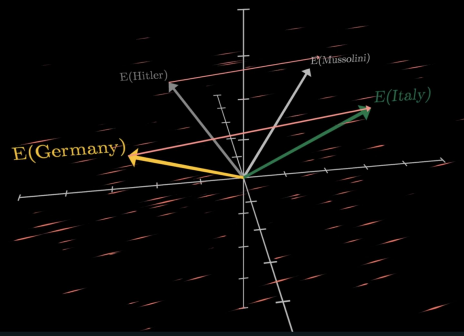
E(Sushi) + E(Germany) - E(Japan) ≈ E(Bratwurst)



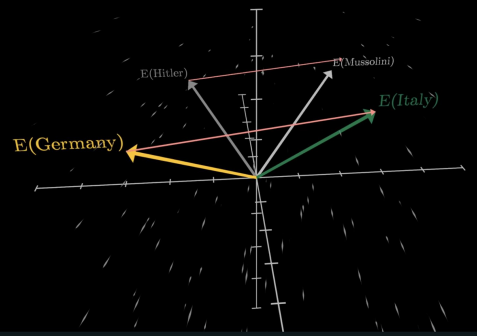
E(mother) - E(father) ≈ E(woman) - E(man)



E(Hitler) + E(Italy) - E(Germany) ≈ E(Mussolini)



E(Hitler) + E(Italy) - E(Germany) ≈ E(Mussolini)

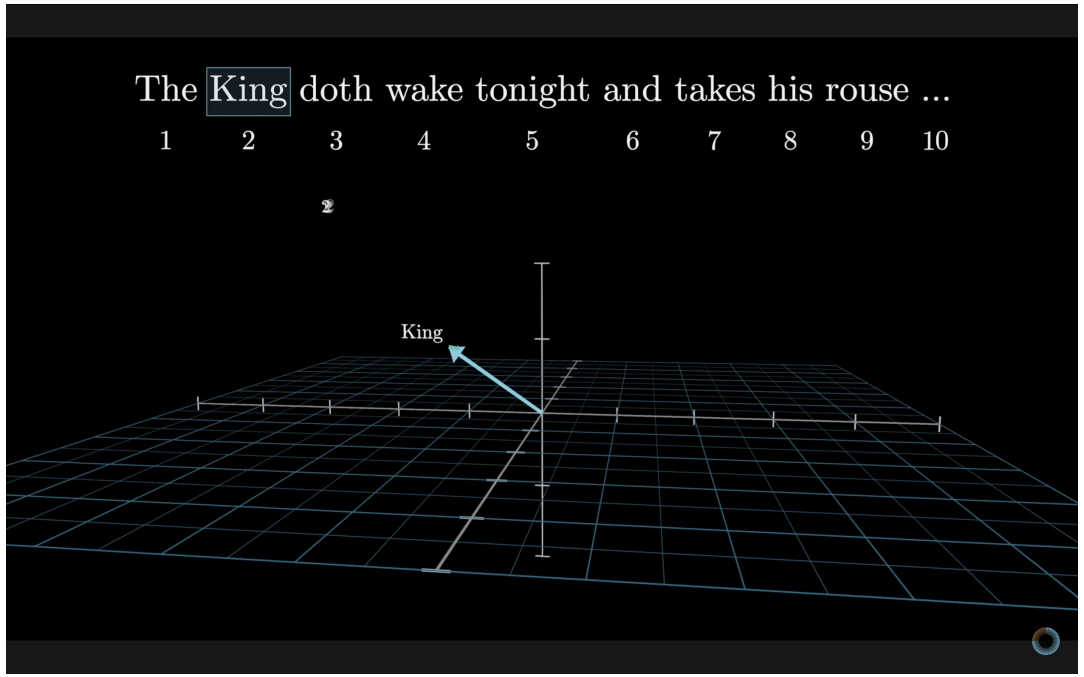


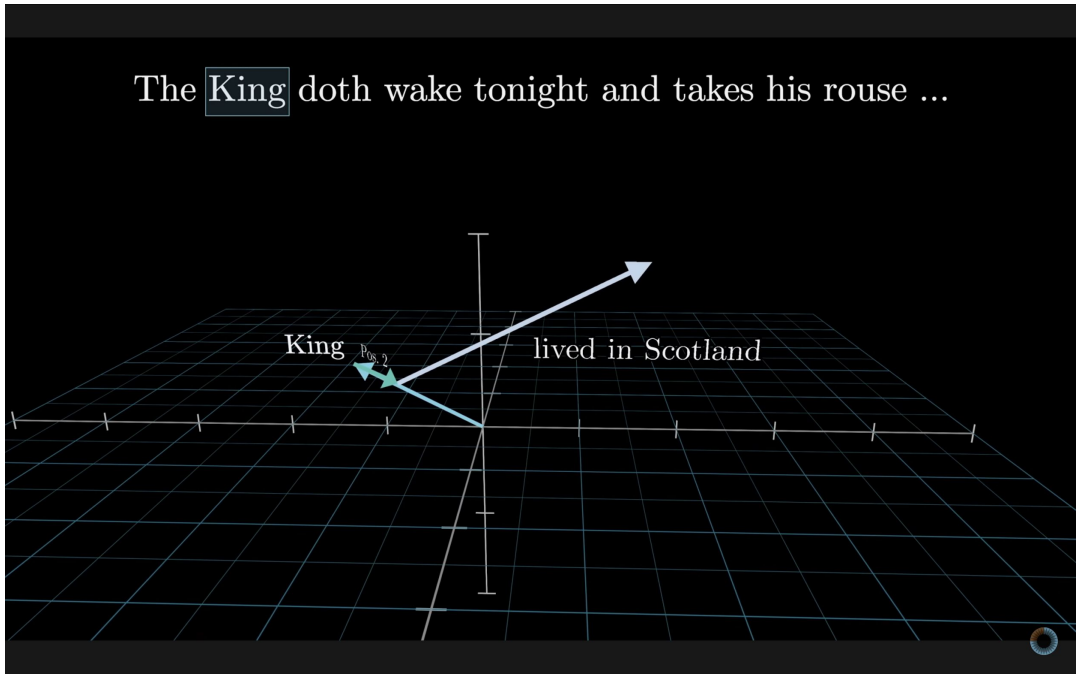
Task: next word prediction

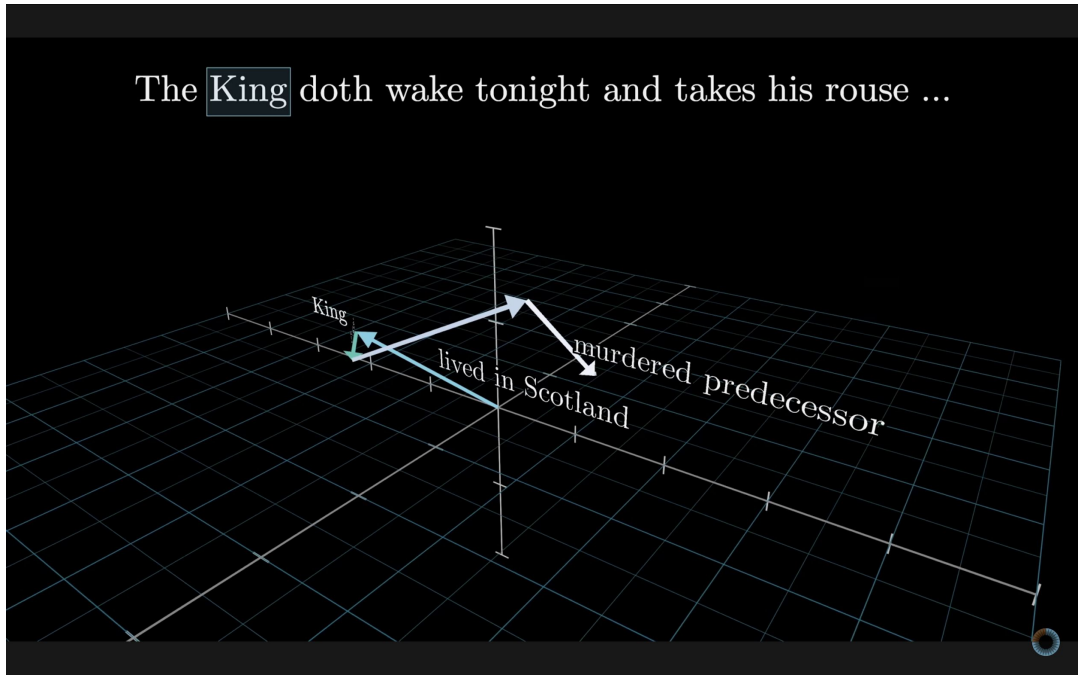
- We can use each word in a sentence to predict the following word; however, we would be missing the context that gives meaning to it
- Context: how the meaning of a word is affected by the meaning of the other words in a given text

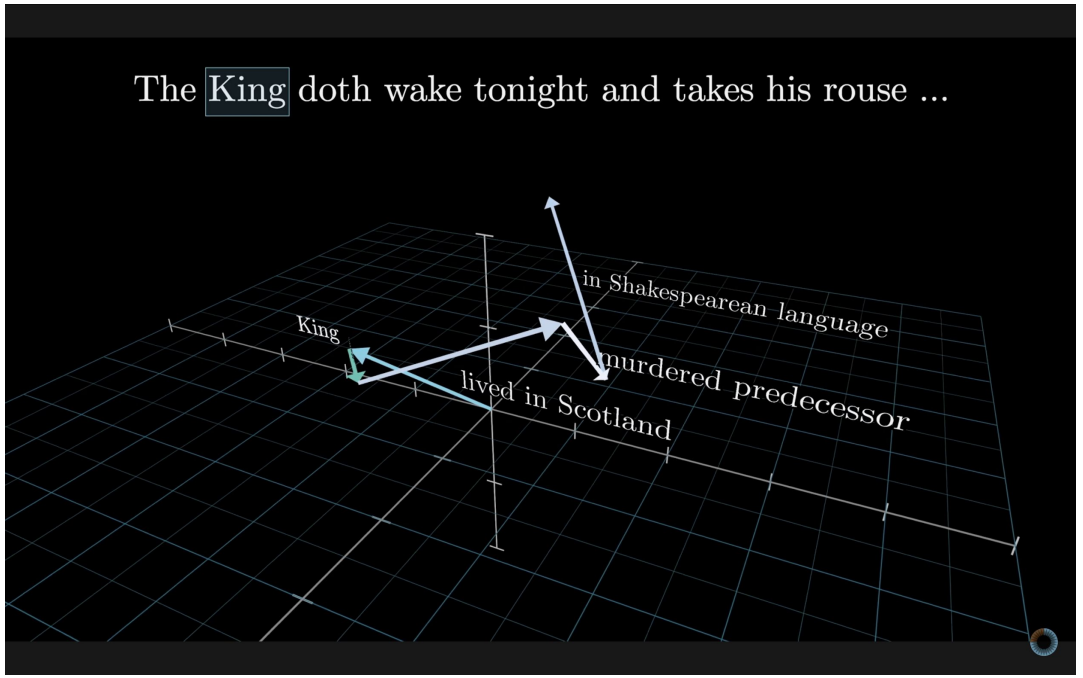
1. Machine learning model
2. Fashion model

- Language model: from an initial embedding or snippet we want to obtain a single final embedding, which we can then run through an MLP and predict the probability of each word in the vocabulary appearing next


The King doth wake tonight and takes his rouse ...
1 2 3 4 5 6 7 8 9 10
King

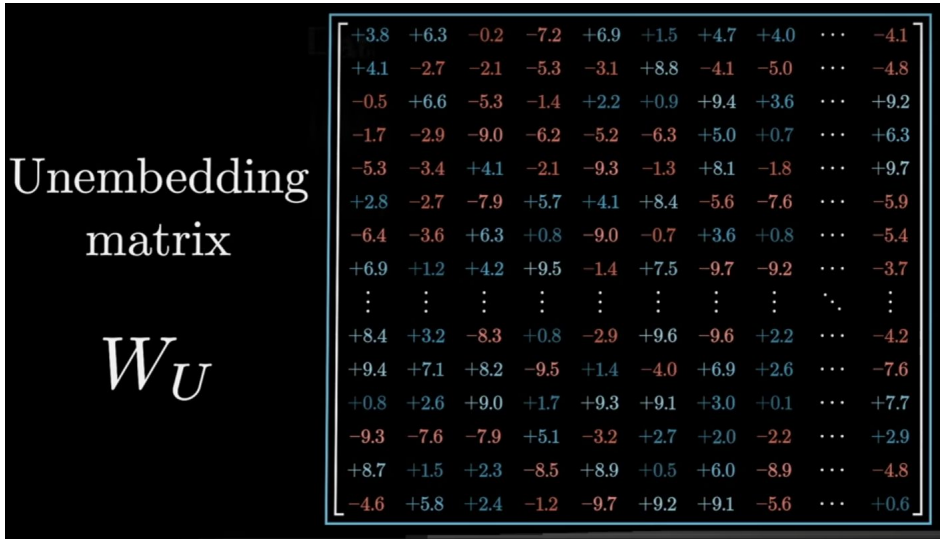

The King doth wake tonight and takes his rouse ...
King
lived in Scotland

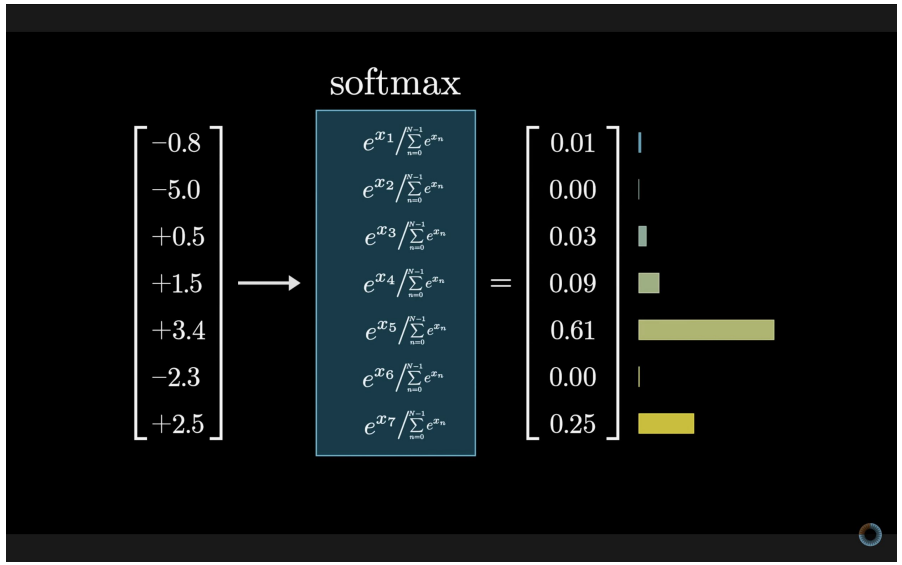

The King doth wake tonight and takes his rouse ...
King
lived in Scotland
murdered predecessor

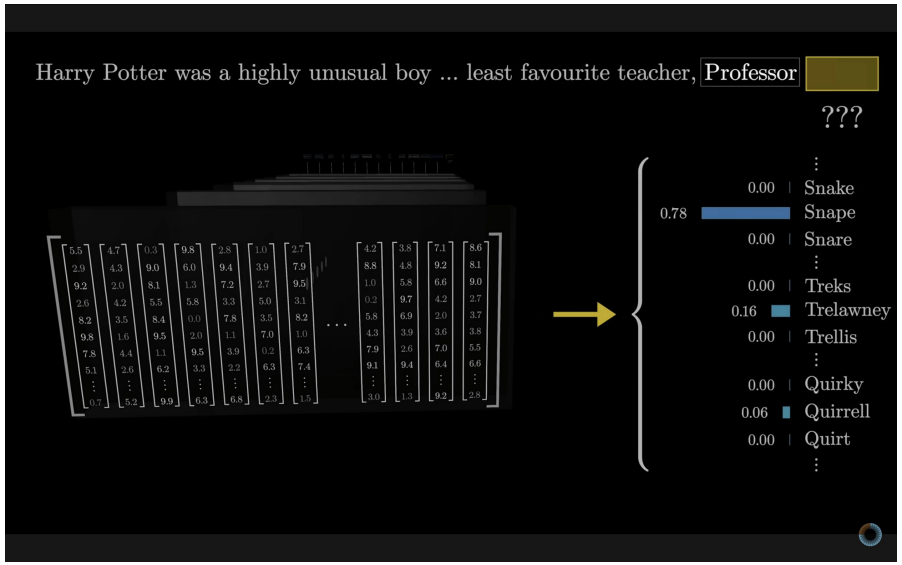

The King doth wake tonight and takes his rouse ...
in Shakespearean language
King
murdered predecessor
lived in Scotland

Embedding dimension

Vocab size

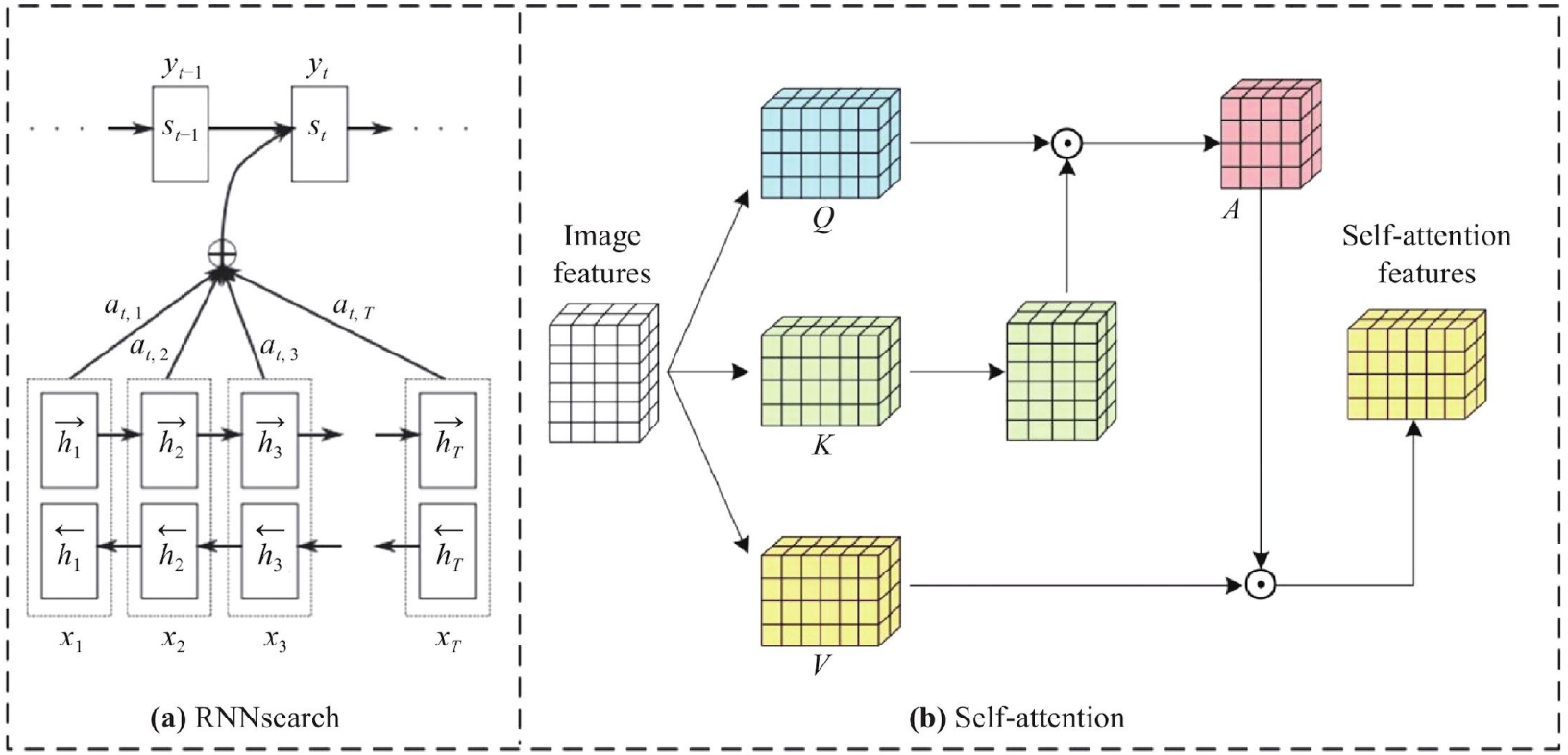




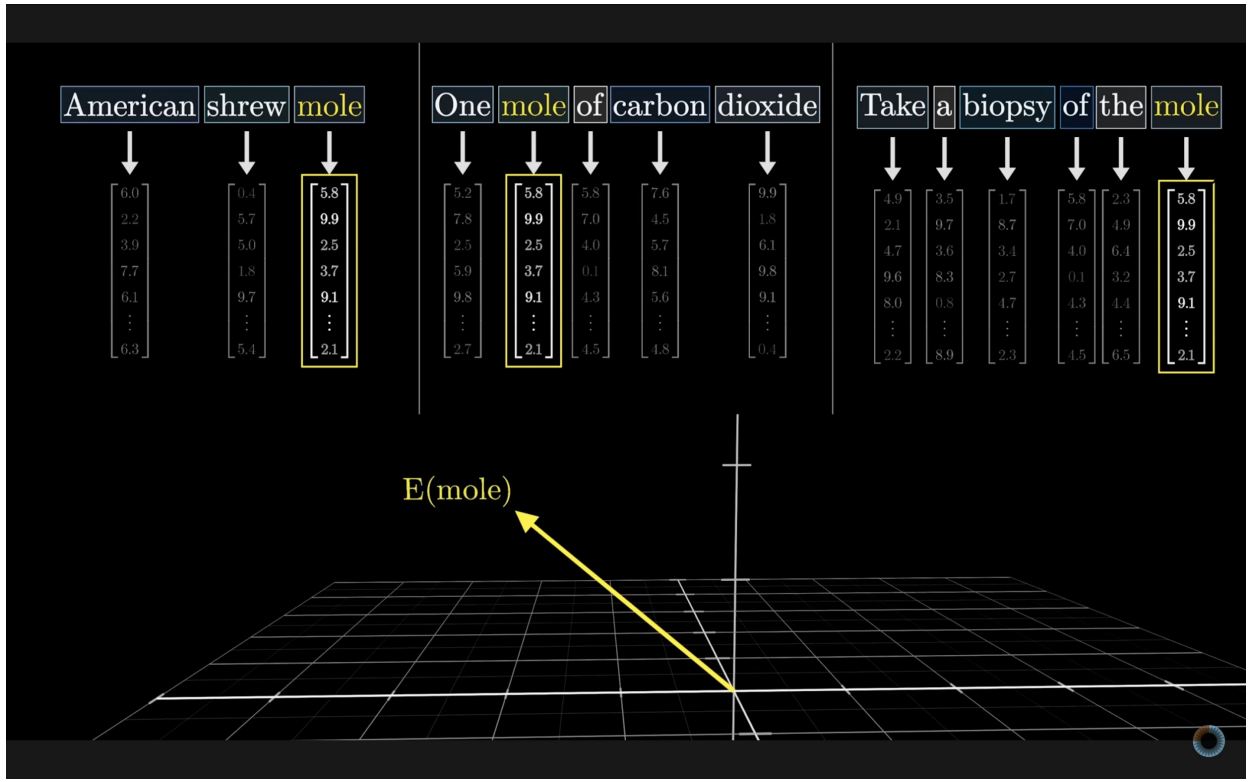


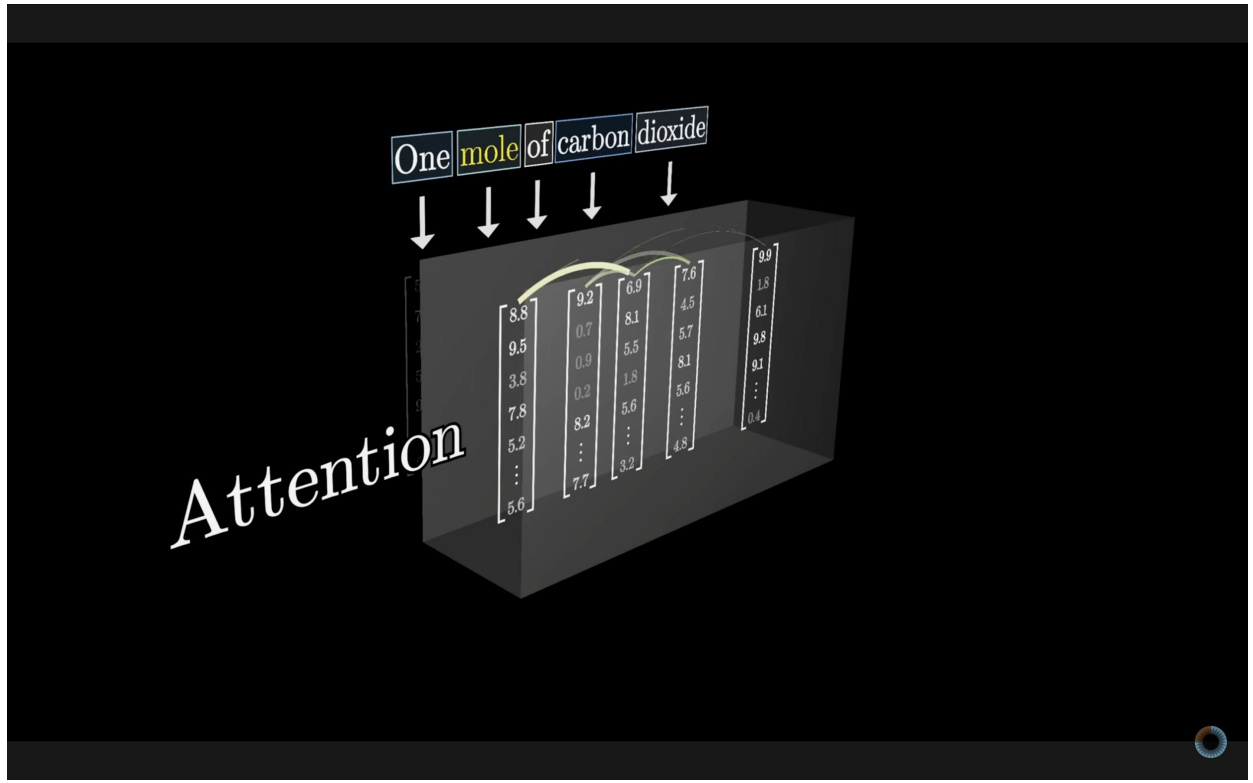
Attention mechanism

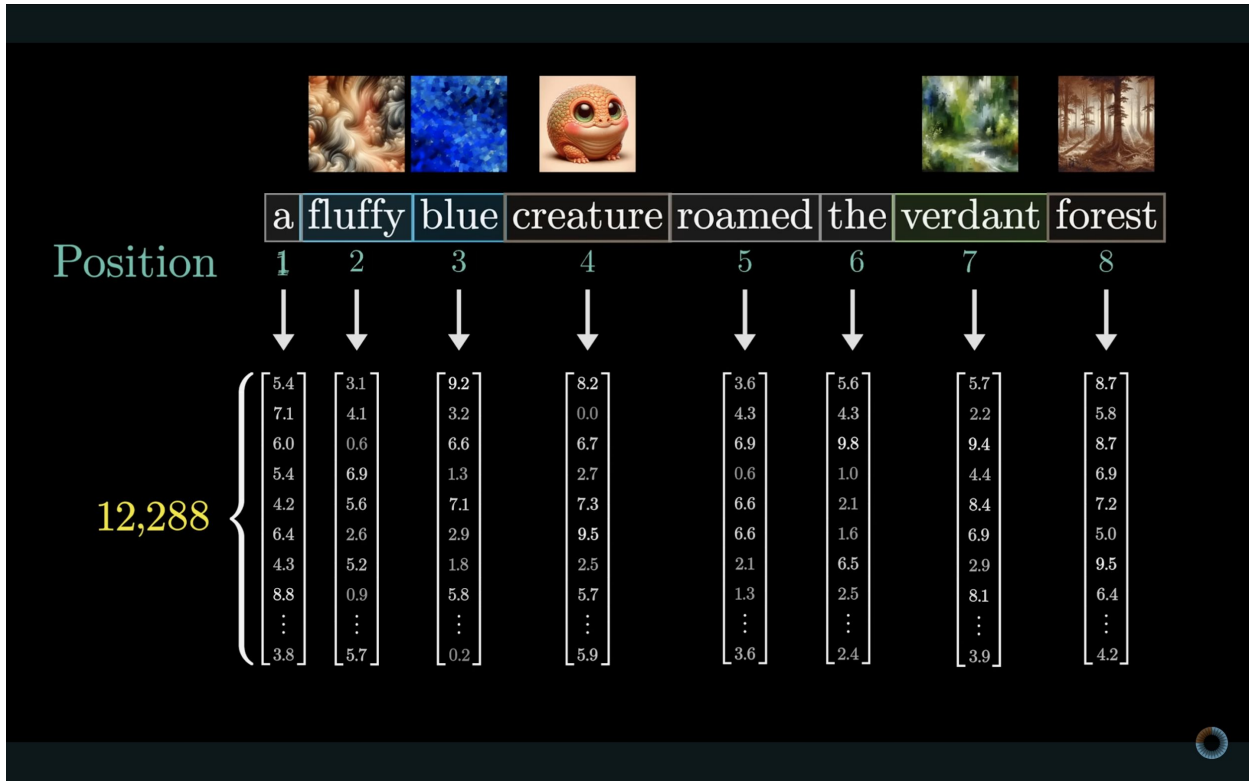
The teacher asked the class a question, and everyone turned to look at **her**

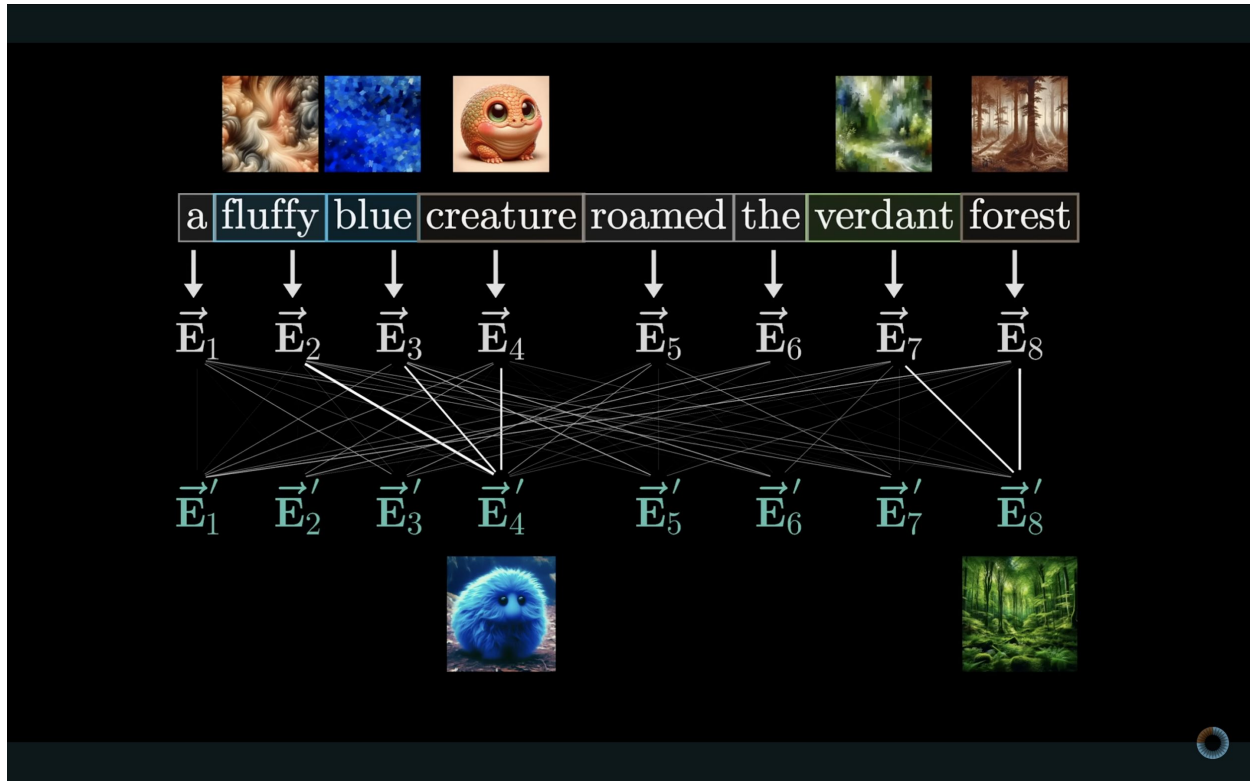


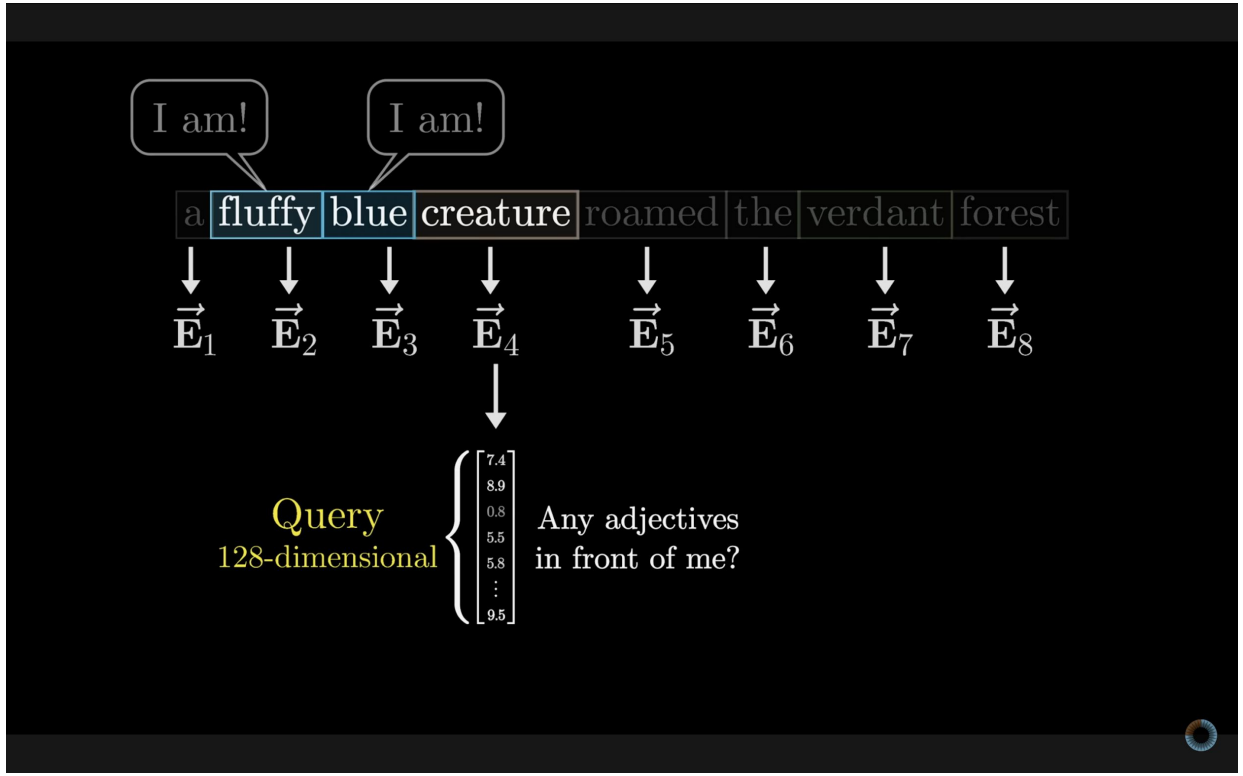
Li, X., Li, M., Yan, P., Li, G., Jiang, Y., Luo, H., & Yin, S. Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds. *International Journal of Network Dynamics and Intelligence*. 2023, *2*(1), 93–116.

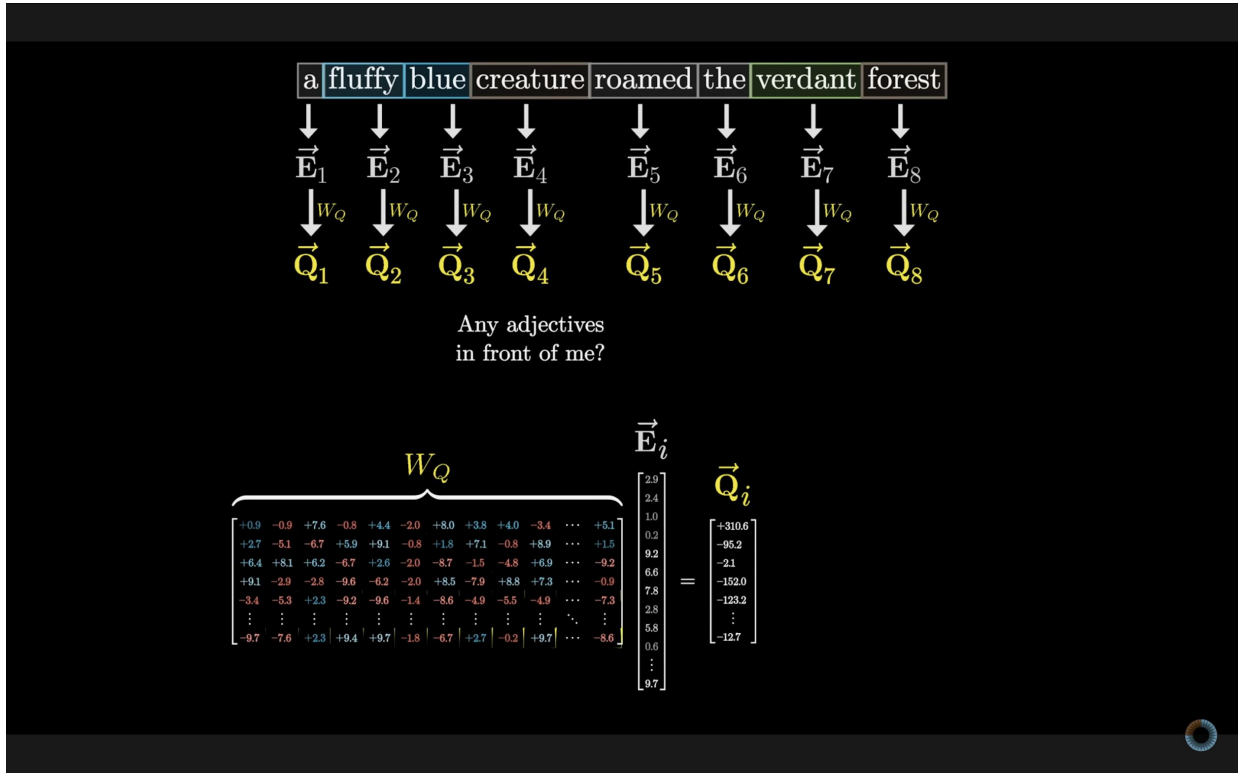

American shrew mole
One mole of carbon dioxide
Take a biopsy of the mole
E(mole)


One mole of carbon dioxide
Attention

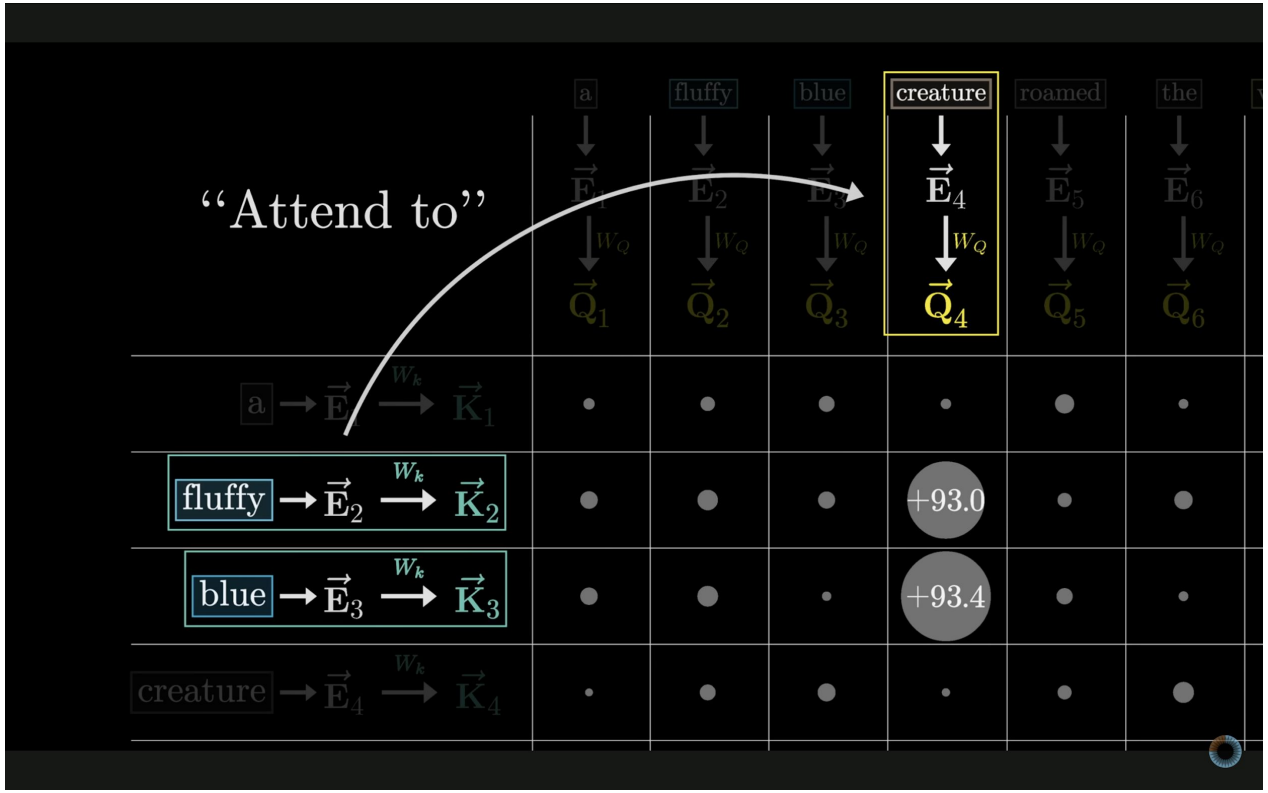

a fluffy blue creature roamed the verdant forest
Position
1 2 3 4 5 6 7 8
12,288
5.4 7.1 6.0 5.4 4.2 6.4 4.3 8.8 ⋮ 3.8
3.1 4.1 0.6 6.9 5.6 2.6 5.2 0.9 ⋮ 5.7
9.2 3.2 6.6 1.3 7.1 2.9 1.8 5.8 ⋮ 0.2
8.2 0.0 6.7 2.7 7.3 9.5 2.5 5.7 ⋮ 5.9
3.6 4.3 6.9 0.6 6.6 6.6 2.1 1.3 ⋮ 3.6
5.6 4.3 9.8 1.0 2.1 1.6 6.5 2.5 ⋮ 2.4
5.7 2.2 9.4 4.4 8.4 6.9 2.9 8.1 ⋮ 3.9
8.7 5.8 8.7 6.9 7.2 5.0 9.5 6.4 ⋮ 4.2


a fluffy blue creature roamed the verdant forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
$\vec{E}'_1$ $\vec{E}'_2$ $\vec{E}'_3$ $\vec{E}'_4$ $\vec{E}'_5$ $\vec{E}'_6$ $\vec{E}'_7$ $\vec{E}'_8$


I am!
I am!
a fluffy blue creature roamed the verdant forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
Query
128-dimensional
7.4
8.9
0.8
5.5
5.8
9.5
Any adjectives in front of me?


a fluffy blue creature roamed the verdant forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
W_Q
$\vec{Q}_1$ $\vec{Q}_2$ $\vec{Q}_3$ $\vec{Q}_4$ $\vec{Q}_5$ $\vec{Q}_6$ $\vec{Q}_7$ $\vec{Q}_8$
Any adjectives in front of me?
W_Q
$\vec{E}_i$
$\vec{Q}_i$


a
fluffy
blue
creature
roamed
the
verdant
forest
$\vec{E}_1$ $\vec{E}_2$ $\vec{E}_3$ $\vec{E}_4$ $\vec{E}_5$ $\vec{E}_6$ $\vec{E}_7$ $\vec{E}_8$
W_Q
$\vec{Q}_1$ $\vec{Q}_2$ $\vec{Q}_3$ $\vec{Q}_4$ $\vec{Q}_5$ $\vec{Q}_6$ $\vec{Q}_7$ $\vec{Q}_8$
a → $\vec{E}_1$ W_k → $\vec{K}_1$
fluffy → $\vec{E}_2$ W_k → $\vec{K}_2$
blue → $\vec{E}_3$ W_k → $\vec{K}_3$
creature → $\vec{E}_4$ W_k → $\vec{K}_4$
roamed → $\vec{E}_5$ W_k → $\vec{K}_5$
the → $\vec{E}_6$ W_k → $\vec{K}_6$
verdant → $\vec{E}_7$ W_k → $\vec{K}_7$
forest → $\vec{E}_8$ W_k → $\vec{K}_8$
$\vec{K}_1 \cdot \vec{Q}_1$ $\vec{K}_1 \cdot \vec{Q}_2$ $\vec{K}_1 \cdot \vec{Q}_3$ $\vec{K}_1 \cdot \vec{Q}_4$ $\vec{K}_1 \cdot \vec{Q}_5$ $\vec{K}_1 \cdot \vec{Q}_6$ $\vec{K}_1 \cdot \vec{Q}_7$ $\vec{K}_1 \cdot \vec{Q}_8$
$\vec{K}_2 \cdot \vec{Q}_1$ $\vec{K}_2 \cdot \vec{Q}_2$ $\vec{K}_2 \cdot \vec{Q}_3$ $\vec{K}_2 \cdot \vec{Q}_4$ $\vec{K}_2 \cdot \vec{Q}_5$ $\vec{K}_2 \cdot \vec{Q}_6$ $\vec{K}_2 \cdot \vec{Q}_7$ $\vec{K}_2 \cdot \vec{Q}_8$
$\vec{K}_3 \cdot \vec{Q}_1$ $\vec{K}_3 \cdot \vec{Q}_2$ $\vec{K}_3 \cdot \vec{Q}_3$ $\vec{K}_3 \cdot \vec{Q}_4$ $\vec{K}_3 \cdot \vec{Q}_5$ $\vec{K}_3 \cdot \vec{Q}_6$ $\vec{K}_3 \cdot \vec{Q}_7$ $\vec{K}_3 \cdot \vec{Q}_8$
$\vec{K}_4 \cdot \vec{Q}_1$ $\vec{K}_4 \cdot \vec{Q}_2$ $\vec{K}_4 \cdot \vec{Q}_3$ $\vec{K}_4 \cdot \vec{Q}_4$ $\vec{K}_4 \cdot \vec{Q}_5$ $\vec{K}_4 \cdot \vec{Q}_6$ $\vec{K}_4 \cdot \vec{Q}_7$ $\vec{K}_4 \cdot \vec{Q}_8$
$\vec{K}_5 \cdot \vec{Q}_1$ $\vec{K}_5 \cdot \vec{Q}_2$ $\vec{K}_5 \cdot \vec{Q}_3$ $\vec{K}_5 \cdot \vec{Q}_4$ $\vec{K}_5 \cdot \vec{Q}_5$ $\vec{K}_5 \cdot \vec{Q}_6$ $\vec{K}_5 \cdot \vec{Q}_7$ $\vec{K}_5 \cdot \vec{Q}_8$
$\vec{K}_6 \cdot \vec{Q}_1$ $\vec{K}_6 \cdot \vec{Q}_2$ $\vec{K}_6 \cdot \vec{Q}_3$ $\vec{K}_6 \cdot \vec{Q}_4$ $\vec{K}_6 \cdot \vec{Q}_5$ $\vec{K}_6 \cdot \vec{Q}_6$ $\vec{K}_6 \cdot \vec{Q}_7$ $\vec{K}_6 \cdot \vec{Q}_8$
$\vec{K}_7 \cdot \vec{Q}_1$ $\vec{K}_7 \cdot \vec{Q}_2$ $\vec{K}_7 \cdot \vec{Q}_3$ $\vec{K}_7 \cdot \vec{Q}_4$ $\vec{K}_7 \cdot \vec{Q}_5$ $\vec{K}_7 \cdot \vec{Q}_6$ $\vec{K}_7 \cdot \vec{Q}_7$ $\vec{K}_7 \cdot \vec{Q}_8$
$\vec{K}_8 \cdot \vec{Q}_1$ $\vec{K}_8 \cdot \vec{Q}_2$ $\vec{K}_8 \cdot \vec{Q}_3$ $\vec{K}_8 \cdot \vec{Q}_4$ $\vec{K}_8 \cdot \vec{Q}_5$ $\vec{K}_8 \cdot \vec{Q}_6$ $\vec{K}_8 \cdot \vec{Q}_7$ $\vec{K}_8 \cdot \vec{Q}_8$


“Attend to”
a
fluffy
blue
creature
roamed
the
$\vec{E}_4$
W_Q
$\vec{Q}_4$
a → $\vec{E}_1$ W_k → $\vec{K}_1$
fluffy → $\vec{E}_2$ W_k → $\vec{K}_2$
blue → $\vec{E}_3$ W_k → $\vec{K}_3$
creature → $\vec{E}_4$ W_k → $\vec{K}_4$
+93.0
+93.4

	a $\vec{E}_1 \xrightarrow{W_Q} \vec{Q}_1$	fluffy $\vec{E}_2 \xrightarrow{W_Q} \vec{Q}_2$	blue $\vec{E}_3 \xrightarrow{W_Q} \vec{Q}_3$	creature $\vec{E}_4 \xrightarrow{W_Q} \vec{Q}_4$	roamed $\vec{E}_5 \xrightarrow{W_Q} \vec{Q}_5$	the $\vec{E}_6 \xrightarrow{W_Q} \vec{Q}_6$	verdant $\vec{E}_7 \xrightarrow{W_Q} \vec{Q}_7$	forest $\vec{E}_8 \xrightarrow{W_Q} \vec{Q}_8$
a $\vec{E}_1 \xrightarrow{W_K} \vec{K}_1$	+0.7	−83.7	−24.7	−27.8	−5.2	−89.3	−45.2	−36.1
fluffy $\vec{E}_2 \xrightarrow{W_K} \vec{K}_2$	−73.4	+2.9	−5.4	+93.0	−48.2	−87.3	−49.7	+7.8
blue $\vec{E}_3 \xrightarrow{W_K} \vec{K}_3$	−53.4	−5.7	+1.8	+93.4	−55.6	−56.0	−26.1	−62.1
creature $\vec{E}_4 \xrightarrow{W_K} \vec{K}_4$	−21.5	−29.7	−56.1	+4.9	−32.4	−92.3	−9.5	−28.1
roamed $\vec{E}_5 \xrightarrow{W_K} \vec{K}_5$	−20.1	−40.9	−87.8	−55.4	+0.6	−64.7	−96.7	−18.9
the $\vec{E}_6 \xrightarrow{W_K} \vec{K}_6$	−87.9	−33.3	−22.6	−31.4	+5.5	+0.6	−4.6	−96.8
verdant $\vec{E}_7 \xrightarrow{W_K} \vec{K}_7$	−41.2	−55.5	−42.3	−59.8	−79.0	−97.9	+3.7	+93.8
forest $\vec{E}_8 \xrightarrow{W_K} \vec{K}_8$	−58.9	−75.5	−91.1	−90.6	−75.6	−89.0	−70.8	+4.7

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

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

	Q_1	Q_2	Q_3	Q_4	Q_5	$\cdots$	Q_n
K_1	$\frac{Q_1 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_1}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_1}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_1}{\sqrt{d_k}}$
K_2	$\frac{Q_1 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_2}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_2}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_2}{\sqrt{d_k}}$
K_3	$\frac{Q_1 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_3}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_3}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_3}{\sqrt{d_k}}$
K_4	$\frac{Q_1 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_4}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_4}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_4}{\sqrt{d_k}}$
K_5	$\frac{Q_1 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_2 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_3 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_4 \cdot K_5}{\sqrt{d_k}}$	$\frac{Q_5 \cdot K_5}{\sqrt{d_k}}$	$\cdots$	$\frac{Q_n \cdot K_5}{\sqrt{d_k}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$

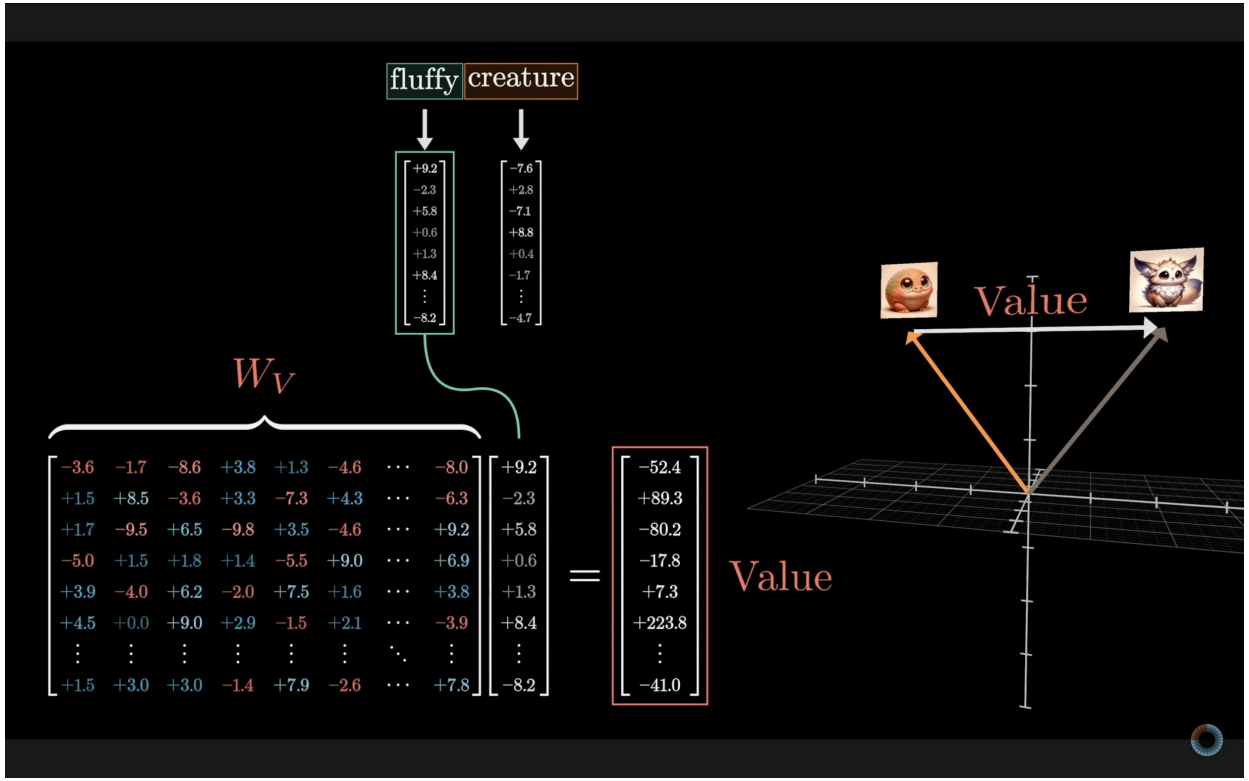
Unnormalized Attention Pattern

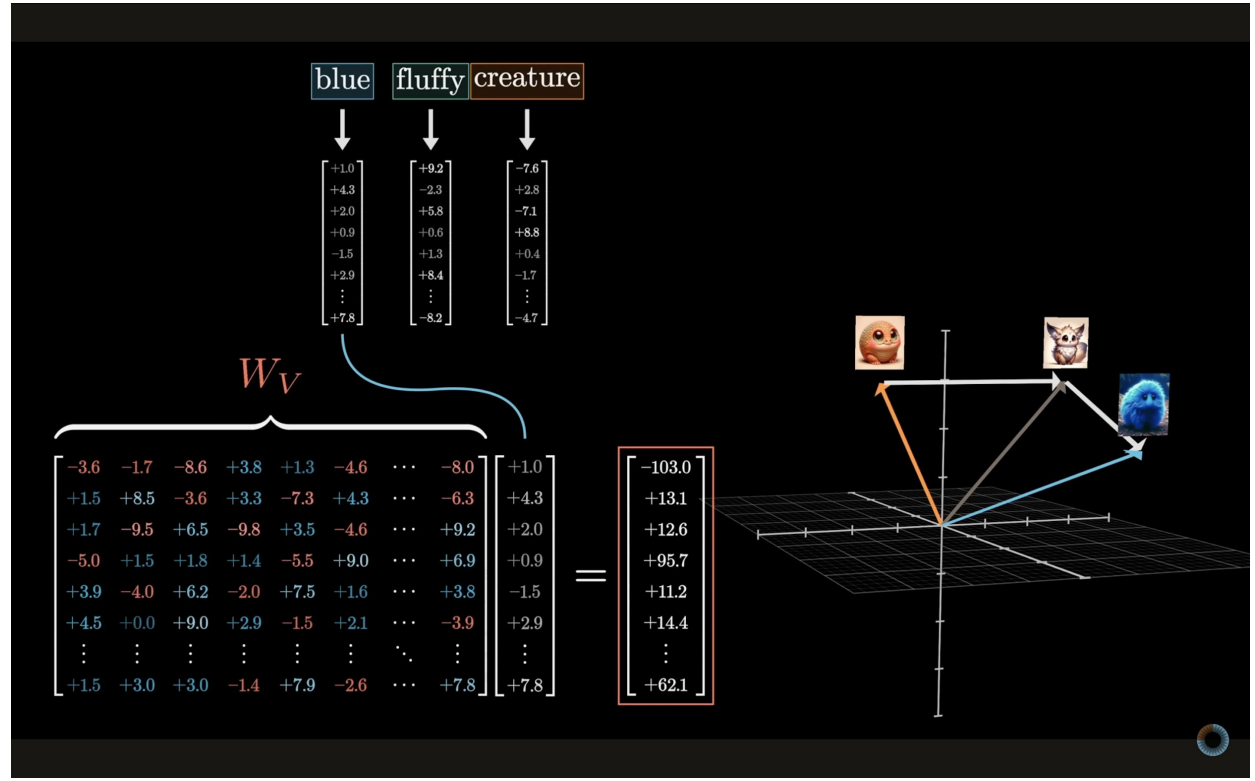
+3.53	+0.80	+1.96	+4.48	+3.74	−1.95
−∞	−0.30	−0.21	+0.82	+0.29	+2.91
−∞	−∞	+0.89	+0.67	+2.99	−0.41
−∞	−∞	−∞	+1.31	+1.73	−1.48
−∞	−∞	−∞	−∞	+3.07	+2.94
−∞	−∞	−∞	−∞	−∞	+0.31

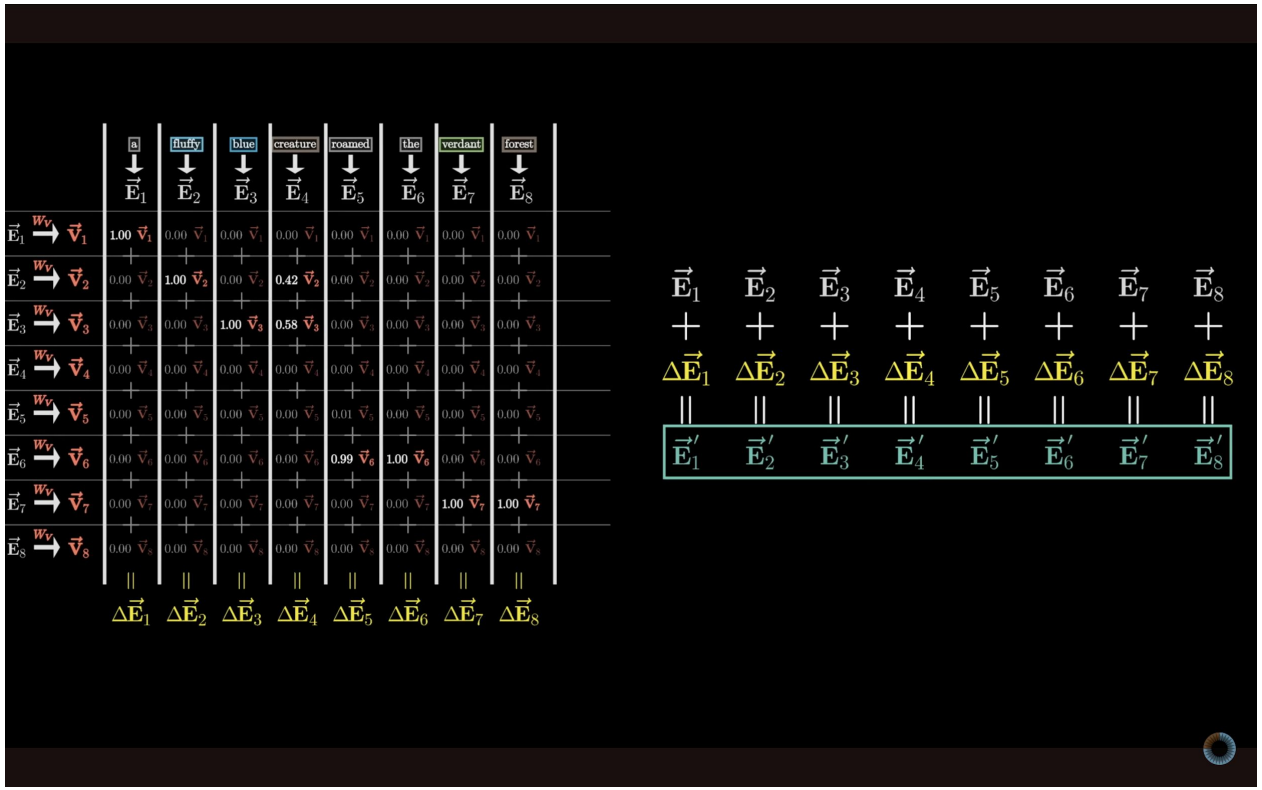
softmax →

Normalized Attention Pattern

1.00	0.75	0.69	0.92	0.46	0.00
0.00	0.25	0.08	0.02	0.01	0.46
0.00	0.00	0.24	0.02	0.22	0.02
0.00	0.00	0.00	0.04	0.06	0.01
0.00	0.00	0.00	0.00	0.24	0.48
0.00	0.00	0.00	0.00	0.00	0.03


fluffy
creature
W_V
Value
Value


blue
fluffy
creature
W_V



Query

$$\begin{bmatrix} +1.0 & +4.3 & +2.0 & +0.9 & -1.5 & +2.9 & -1.2 & +7.8 & \cdots & -2.3 \\ +5.8 & +0.6 & +1.3 & +8.4 & -8.5 & -8.2 & -9.5 & +6.6 & \cdots & +7.3 \\ +9.5 & +5.9 & -0.8 & +5.6 & -7.6 & +2.8 & -7.1 & +8.8 & \cdots & -1.7 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -2.8 & -1.2 & +3.9 & -8.7 & +3.3 & +3.4 & -5.7 & -7.3 & \cdots & -2.7 \end{bmatrix}$$

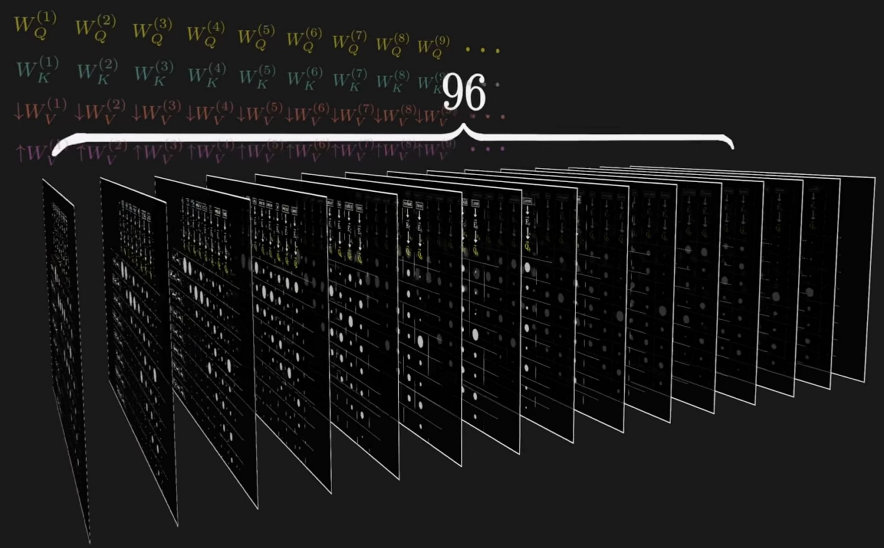
Key

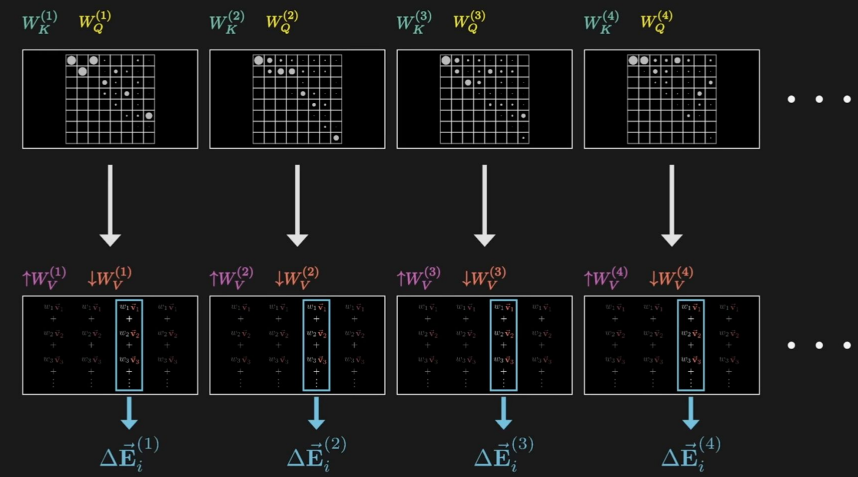
$$\begin{bmatrix} +1.4 & -1.2 & +9.7 & -7.9 & -5.8 & -6.7 & +3.0 & -4.9 & \cdots & -5.1 \\ -6.8 & -7.7 & +3.1 & -7.2 & -6.0 & -2.6 & +6.4 & -8.0 & \cdots & -8.0 \\ +9.4 & -0.6 & +9.4 & +2.1 & +4.7 & -9.1 & -4.3 & -7.5 & \cdots & -7.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -3.6 & +3.3 & -7.3 & +4.3 & -4.2 & -6.3 & +1.7 & -9.5 & \cdots & -9.8 \end{bmatrix}$$

Value

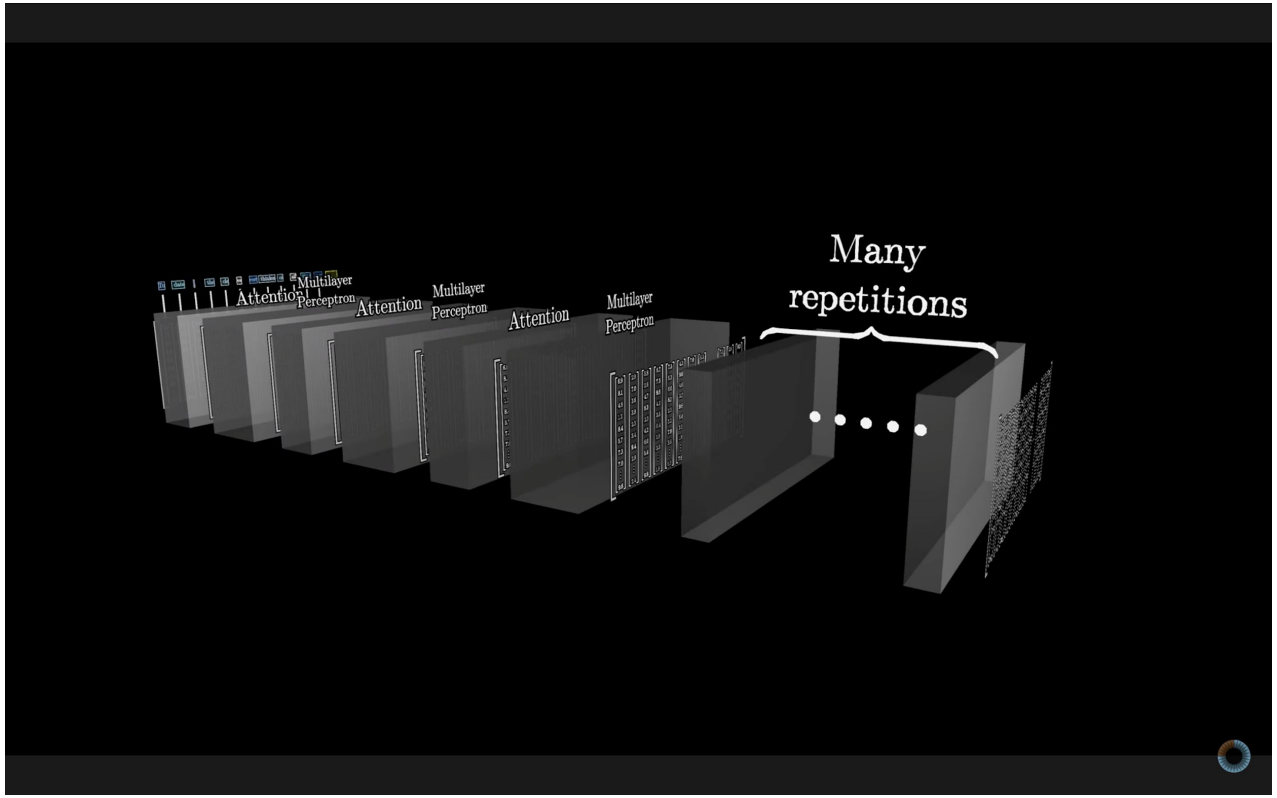
$$\begin{bmatrix} +3.5 & -4.6 & +4.7 & +9.2 & -5.0 & +1.5 & +1.8 & +1.4 & \cdots & +9.0 \\ -1.0 & +6.9 & +3.9 & -4.0 & +6.2 & -2.0 & +7.5 & +1.6 & \cdots & +3.8 \\ +4.5 & +0.0 & +9.0 & +2.9 & -1.5 & +2.1 & -9.5 & -3.9 & \cdots & -4.2 \\ +2.3 & -1.4 & -7.2 & -4.0 & +1.4 & +1.8 & +1.5 & +3.0 & \cdots & -1.4 \\ +7.9 & -2.6 & -1.3 & +7.8 & +6.1 & +4.0 & -7.9 & +8.3 & \cdots & +9.9 \\ -6.9 & +7.3 & -6.7 & +2.3 & -7.4 & +6.9 & +6.1 & +1.4 & \cdots & -8.5 \\ +3.9 & -0.9 & +4.4 & +7.3 & +9.4 & +7.0 & -9.7 & -2.8 & \cdots & -6.5 \\ +0.4 & -8.8 & -5.9 & -9.5 & +5.8 & -5.5 & -3.1 & +8.5 & \cdots & -9.3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -2.0 & -5.7 & -6.2 & +8.8 & +4.7 & -0.2 & -5.4 & -4.9 & \cdots & -1.3 \end{bmatrix}$$

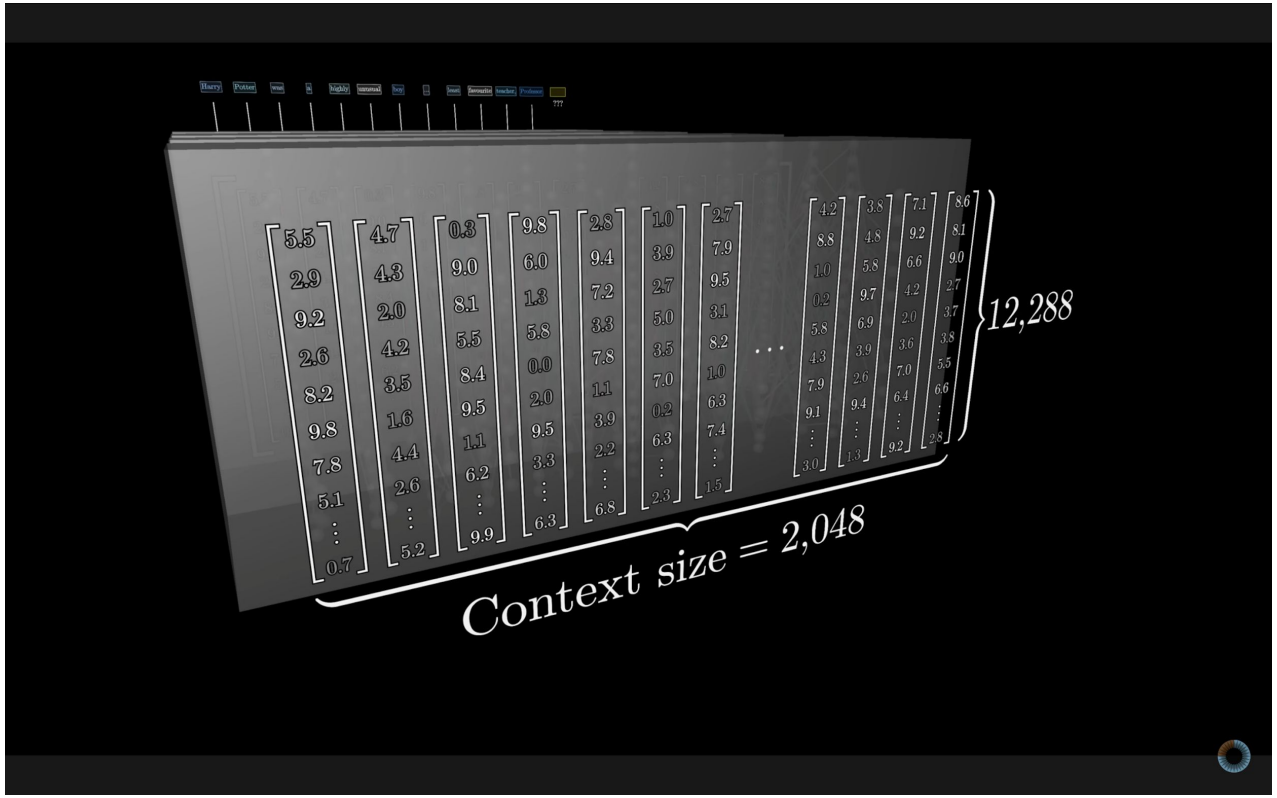
Multi-headed attention





Original embedding $\vec{\mathbf{E}}_i + \Delta\vec{\mathbf{E}}_i^{(1)} + \Delta\vec{\mathbf{E}}_i^{(2)} + \Delta\vec{\mathbf{E}}_i^{(3)} + \Delta\vec{\mathbf{E}}_i^{(4)} + \cdots$


Attention
Multilayer Perceptron
Attention
Multilayer Perceptron
Attention
Multilayer Perceptron
Many repetitions

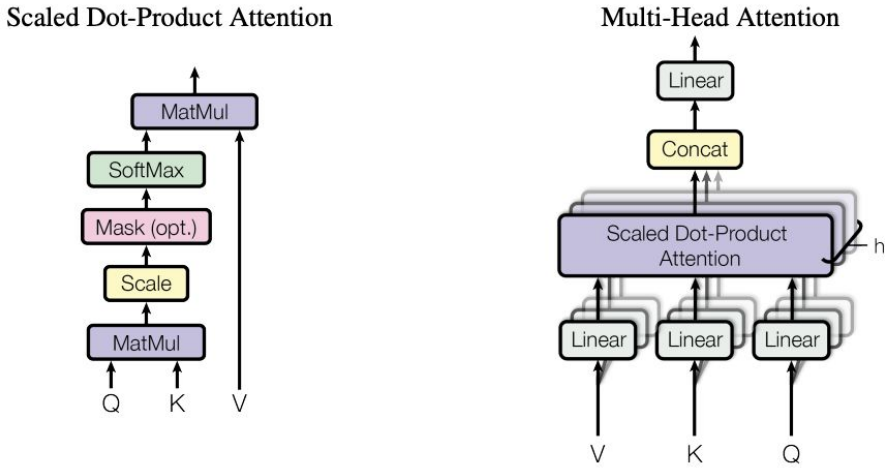

Harry
Potter
was
a
highly
unusual
boy
in
many
ways
.
The
12,288
Context size = 2,048

Transformers

Advantage: we can calculate attention for all the embeddings in a sequence simultaneously. This speeds up computation and allows for massive scalability

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).

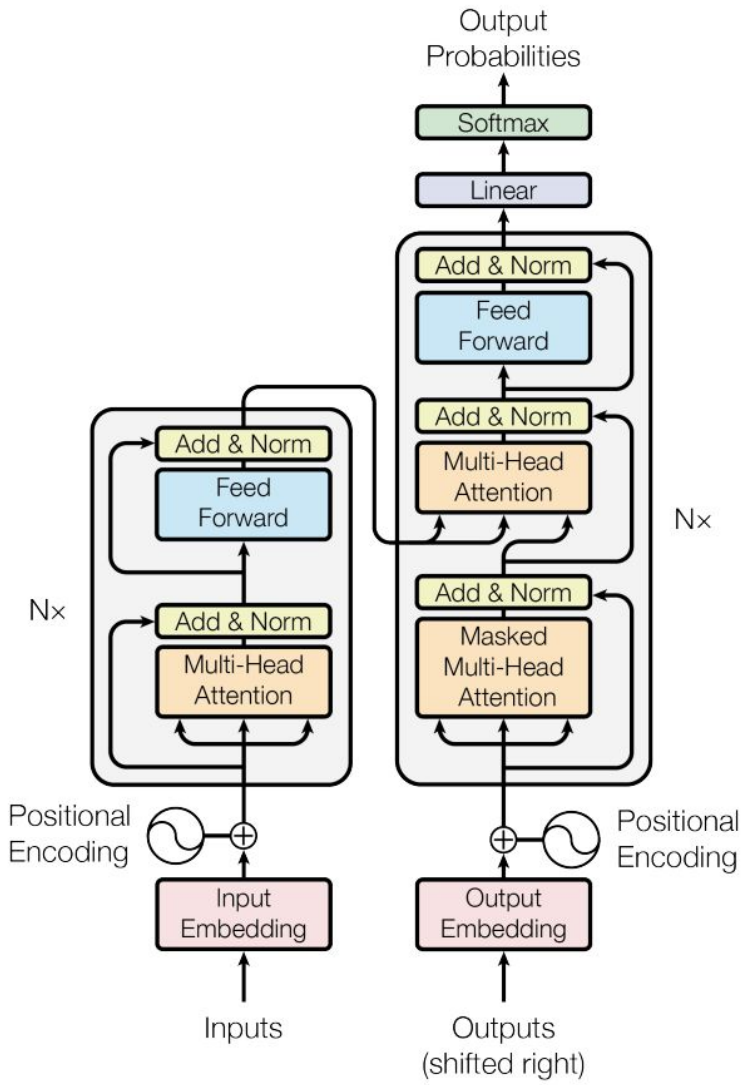
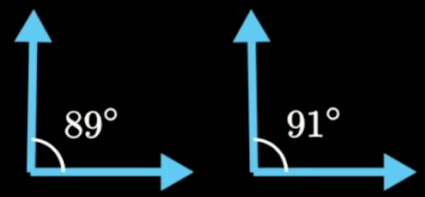


Figure 1: The Transformer - model architecture.

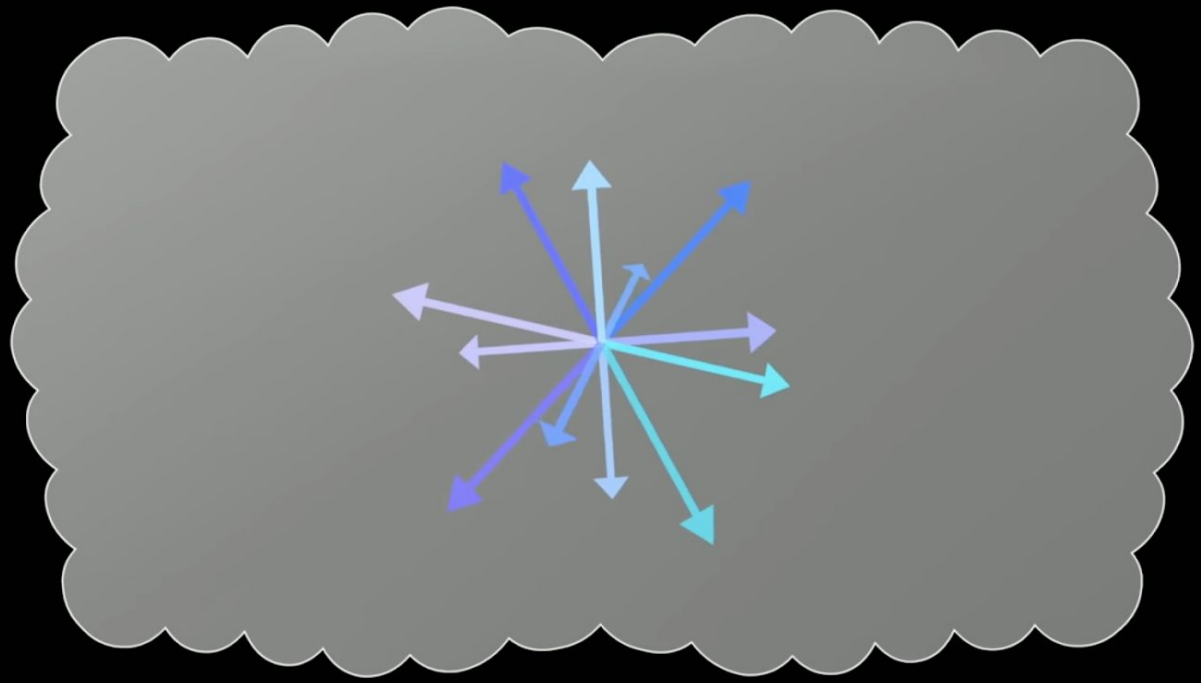


Choose multiple vectors,
each pair between 89° and 91° apart

Johnson–Lindenstrauss Lemma

$\Longrightarrow$ Maximum # of vectors: $\approx \exp(\epsilon \cdot N)$

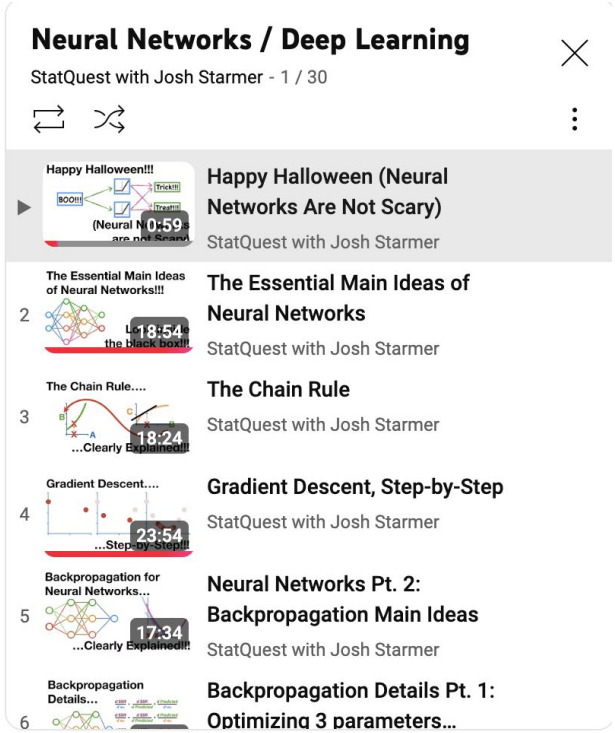
N-dimensional Space

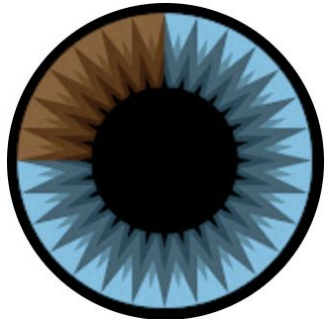


YouTube channels

StatQuest with Josh Starmer

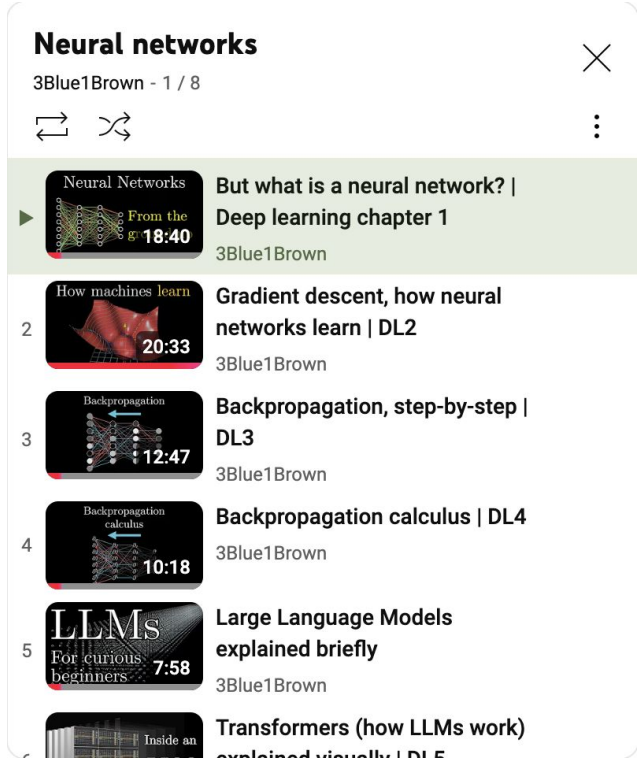
@statquest





3Blue1Brown

@3blue1brown



DeepLearningAI

@Deeplearningai

vcubingx

@vcubingx

Serrano.Academy

@SerranoAcademy