CONSTELLATE

How does ChatGPT work?

Some clarifications

1. Conceptual questions vs.engineering questions
2. Human-engineered features vs. machine learning

Conceptual Questions
vs.
Engineering Questions

Skip gram

Conceptual question:
What determines the meaning of a word?

CONSTELLATE

Skip gram

Conceptual question:
What determines the meaning of a word?

The distributional hypothesis:
words that are surrounded by similar words have similar meanings

Skip gram

Sentence 1: I have to make sure that the cat gets fed.

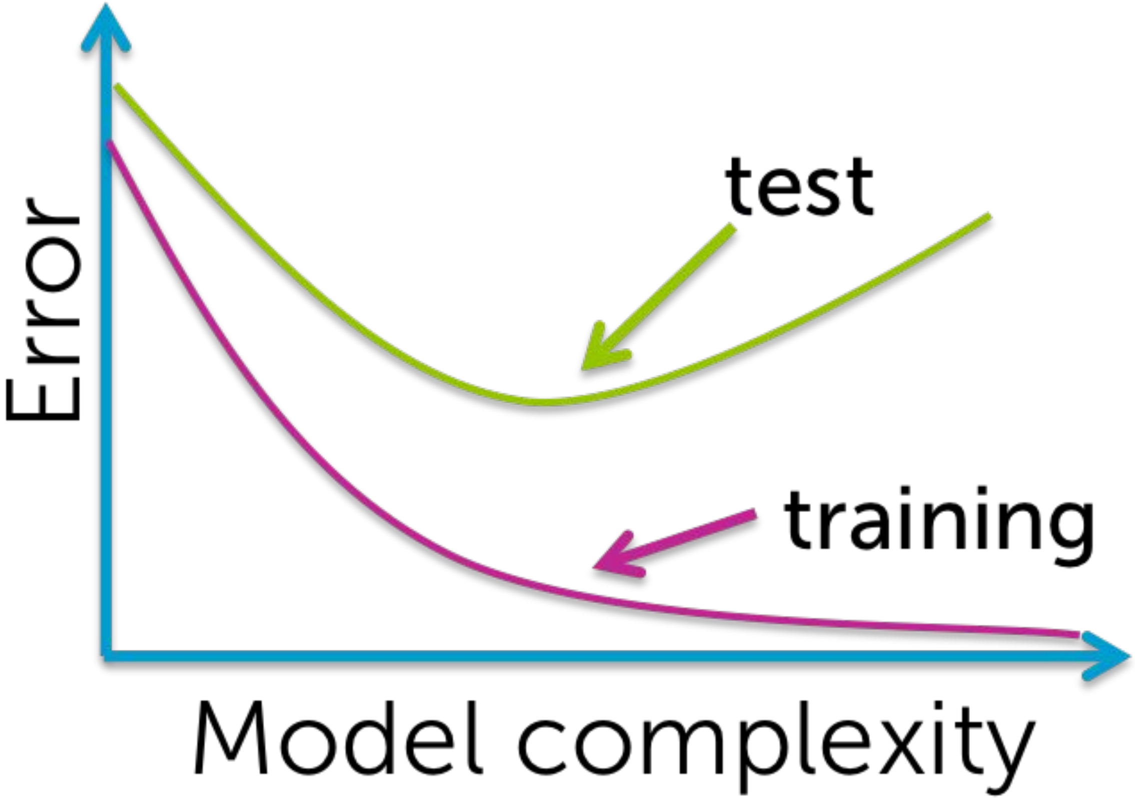
Sentence 2: I forgot to make sure that the dog gets fed.

Skip gram

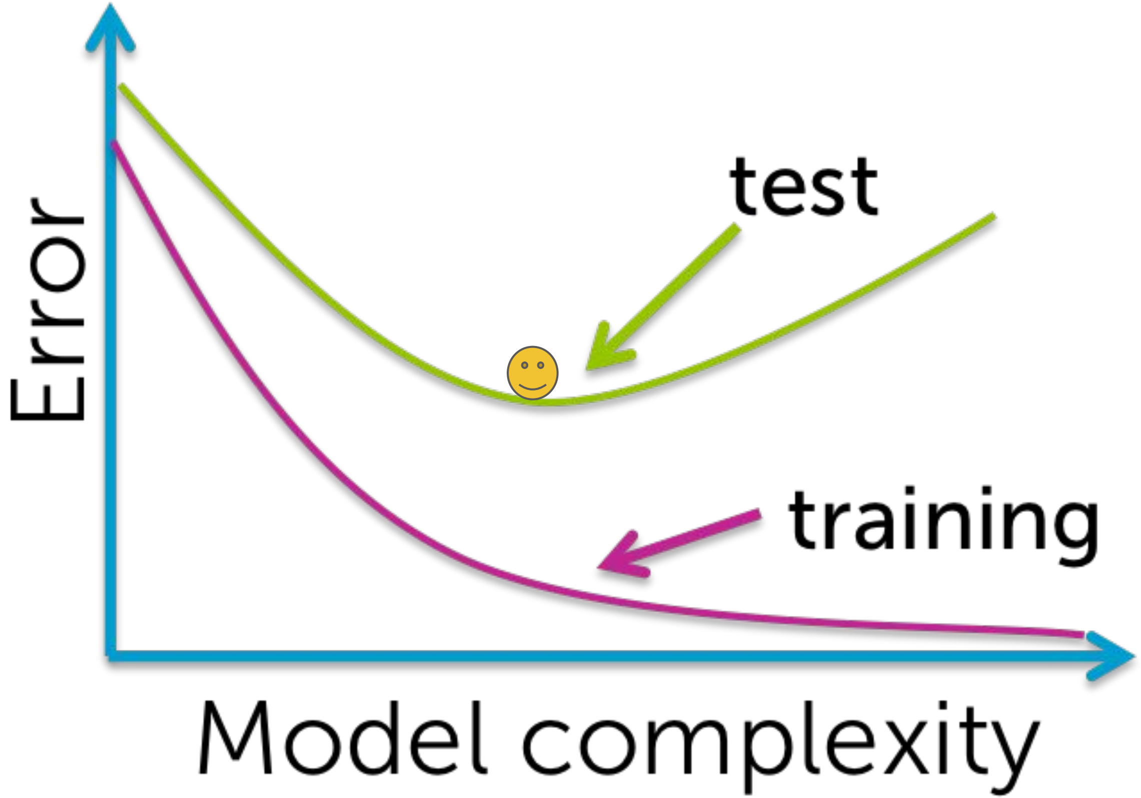
Engineering question:
Why the hidden layer has 300 neurons, not 200, not 500?

CONSTELLATE

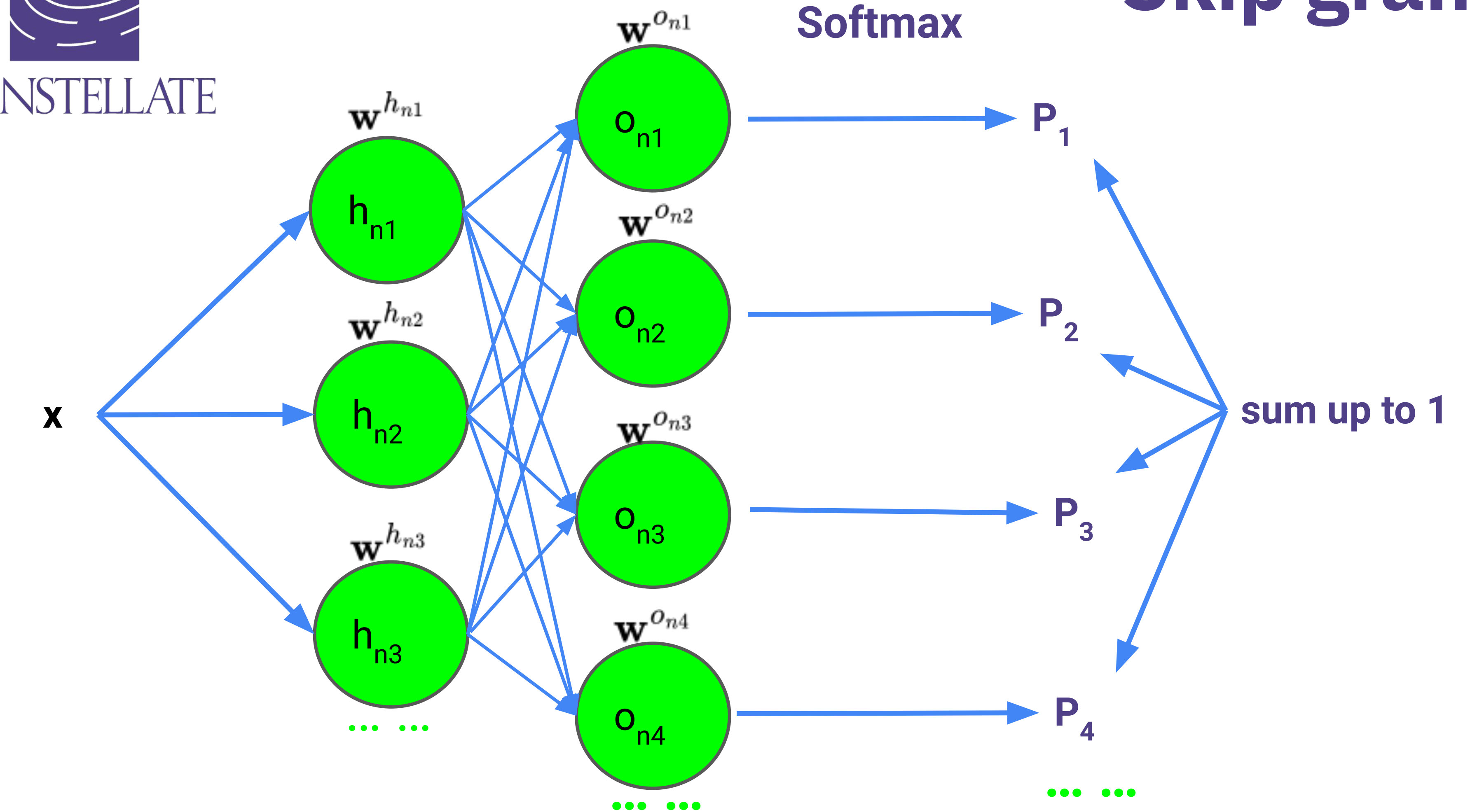
Skip gram



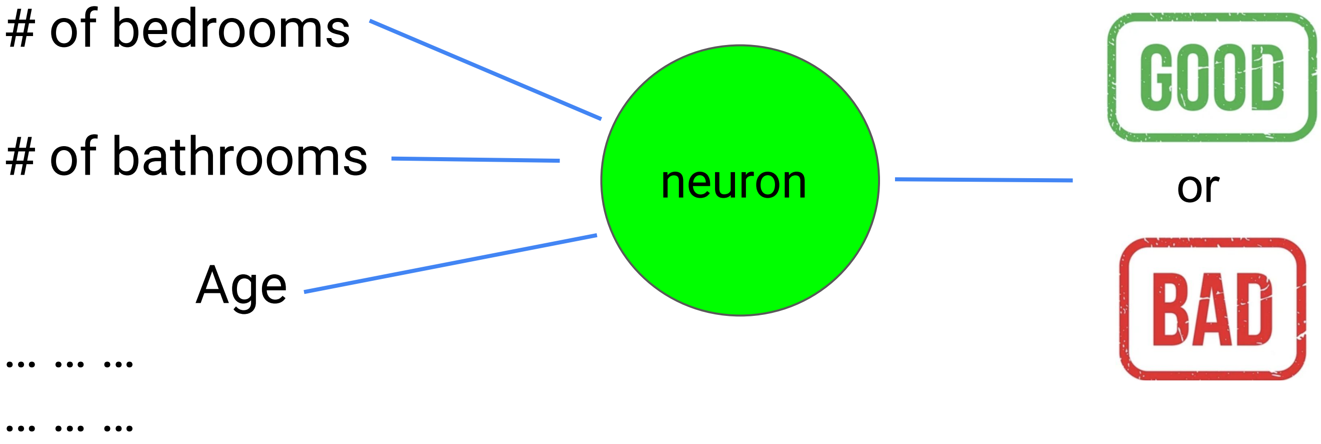
Skip gram



Skip gram


Softmax
$\mathbf{w}^{h_{n1}}$
$\mathbf{w}^{h_{n2}}$
$\mathbf{w}^{h_{n3}}$
$\mathbf{w}^{o_{n1}}$
$\mathbf{w}^{o_{n2}}$
$\mathbf{w}^{o_{n3}}$
$\mathbf{w}^{o_{n4}}$
x
h_{n1}
h_{n2}
h_{n3}
... ...
o_{n1}
o_{n2}
o_{n3}
o_{n4}
... ...
P_1
P_2
P_3
P_4
... ...
sum up to 1

Human-engineered features vs. machine learning



# of bedrooms

# of bathrooms

Age

… … …

… … …

neuron

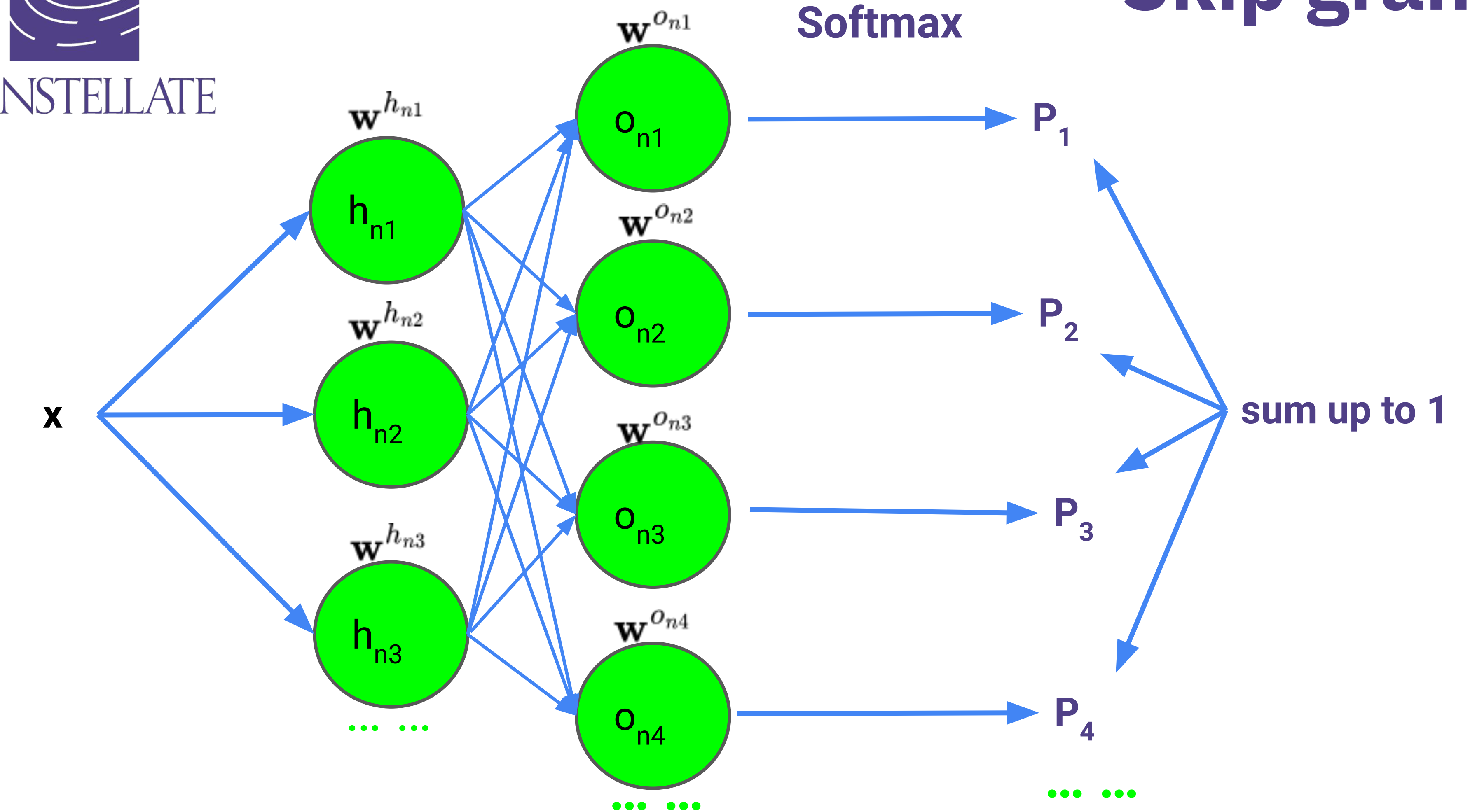
GOOD

or

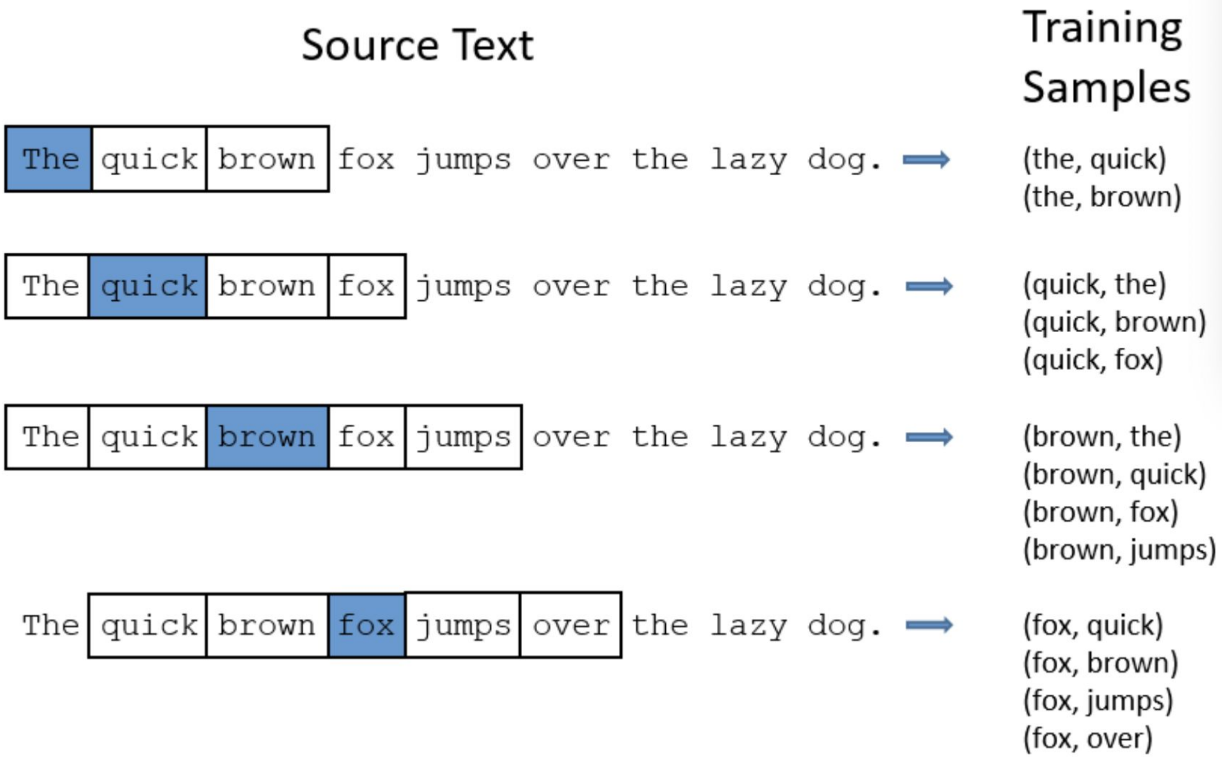
BAD

Features

Skip gram


Softmax
$\mathbf{w}^{o_{n1}}$
$\mathbf{w}^{o_{n2}}$
$\mathbf{w}^{o_{n3}}$
$\mathbf{w}^{o_{n4}}$
$\mathbf{w}^{h_{n1}}$
$\mathbf{w}^{h_{n2}}$
$\mathbf{w}^{h_{n3}}$
x
h_{n1}
h_{n2}
h_{n3}
... ...
o_{n1}
o_{n2}
o_{n3}
o_{n4}
... ...
P_1
P_2
P_3
P_4
... ...
sum up to 1

Skip gram



Source Text

Training Samples

The quick brown fox jumps over the lazy dog. ⟹

(the, quick)
(the, brown)

The quick brown fox jumps over the lazy dog. ⟹

(quick, the)
(quick, brown)
(quick, fox)

The quick brown fox jumps over the lazy dog. ⟹

(brown, the)
(brown, quick)
(brown, fox)
(brown, jumps)

The quick brown fox jumps over the lazy dog. ⟹

(fox, quick)
(fox, brown)
(fox, jumps)
(fox, over)

CONSTELLATE

Any questions?

CONSTELLATE

A brief review

CONSTELLATE

ChatGPT

GPT: generative pre-trained transformer

Transformer

a multi-layer neural network that relies on the parallel multi-head attention mechanism.

Last two sessions

Part 1: multi-layer neural network

From today

Part 2: multi-head attention mechanism

For today

A single self-attention head

CONSTELLATE

Part 2

Learning objectives

- Concepts
  - Understand the temporal dimension of language
  - Understand some basic concepts in attention
    - query, key, value, self-attention
  - Understand a head as a representation of a certain relationship between words
- Computation
  - Understand the power of matrix multiplication for parallel computation

CONSTELLATE

Roadmap

- The temporal dimension of language
- A single self-attention head

CONSTELLATE

The main task of LM

Given a sequence of words, what is the most likely word that follows the sequence?

The quick brown fox jumps over the lazy dog.

CONSTELLATE

The main task of LM

Given a sequence of words, what is the most likely word that follows the sequence?

$$w_1 w_2 w_3 w_4 \longrightarrow \quad w_5$$

Temporal dimension

The main task of LM

Given a sequence of words, what is the most likely word that follows the sequence?

$$w_1 w_2 w_3 w_4 \longrightarrow w_5$$

Temporal dimension

Assumption

Every word in the sequence plays a role in determining what word follows this sequence.

$$w_1w_2w_3w_4 \longrightarrow w_5$$

Temporal dimension

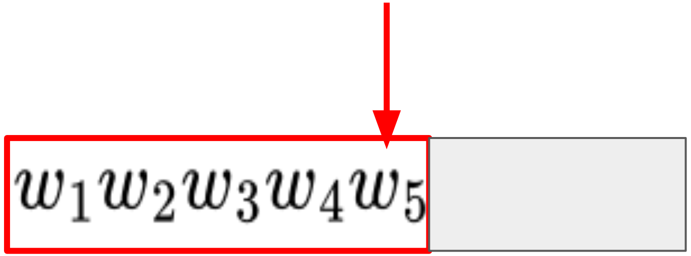
Assumption

In a word sequence, every word is influenced by the past words and itself, but never by the future words.

Temporal dimension

$$\boxed{w_1 w_2 w_3 w_4 \overset{\downarrow}{w_5}} w_6 w_7 w_8$$

Temporal dimension



Attention

$$w_1 w_2 w_3 w_4 \longrightarrow w_5$$

CONSTELLATE

Attention

Damian had a secret child, a girl, and had written in his will that his belongings will belong to ____

CONSTELLATE

Interim summary

1. Language has a temporal dimension.
2. When making predictions about future words, we need to pay more attention to certain words in the given sequence.

The task of ChatGPT

Given a sequence of words, generate the next word and continue this process.

CONSTELLATE

The task of ChatGPT

$$w_1 w_2 \ldots w_8 \rightarrow w_9$$

$$w_2 w_3 \ldots w_9 \rightarrow w_{10}$$

$$w_3 w_4 \ldots w_{10} \rightarrow w_{11}$$

Context length:8

We are using the past words to predict the future words.

CONSTELLATE

The question becomes:

How do we encode the relationship between the words in a given sequence?

Recall that...

A word is represented by a feature vector

$$\begin{bmatrix} w_1^{f_1} & w_1^{f_2} & \dots & w_1^{f_{32}} \end{bmatrix}$$

$$1 \times 32$$

If we have a word sequence of len 8

$$\begin{bmatrix} w_1^{f_1} & w_1^{f_2} & \dots & w_1^{f_{32}} \\ w_2^{f_1} & w_2^{f_2} & \dots & w_2^{f_{32}} \\ \dots & \dots & \dots & \dots \\ w_8^{f_1} & w_8^{f_2} & \dots & w_8^{f_{32}} \end{bmatrix}$$

$$8 \times 32$$

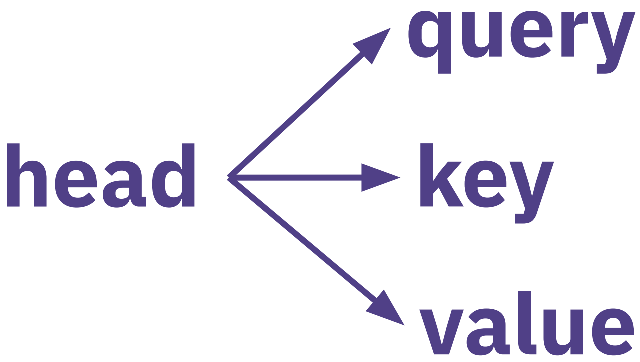
CONSTELLATE

How do we encode the relationship between the words in a given sequence?

We use attention heads!

Self-attention head

For each word token



head

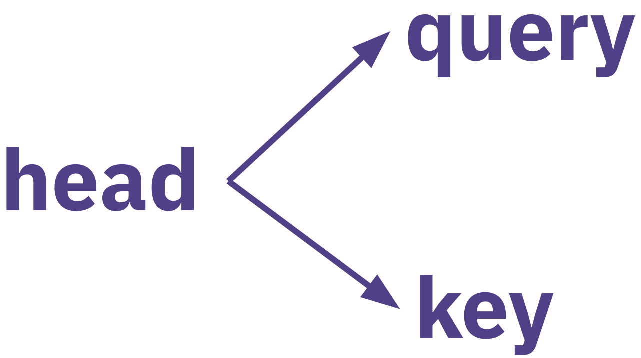
query

key

value

Self-attention head

For each word token

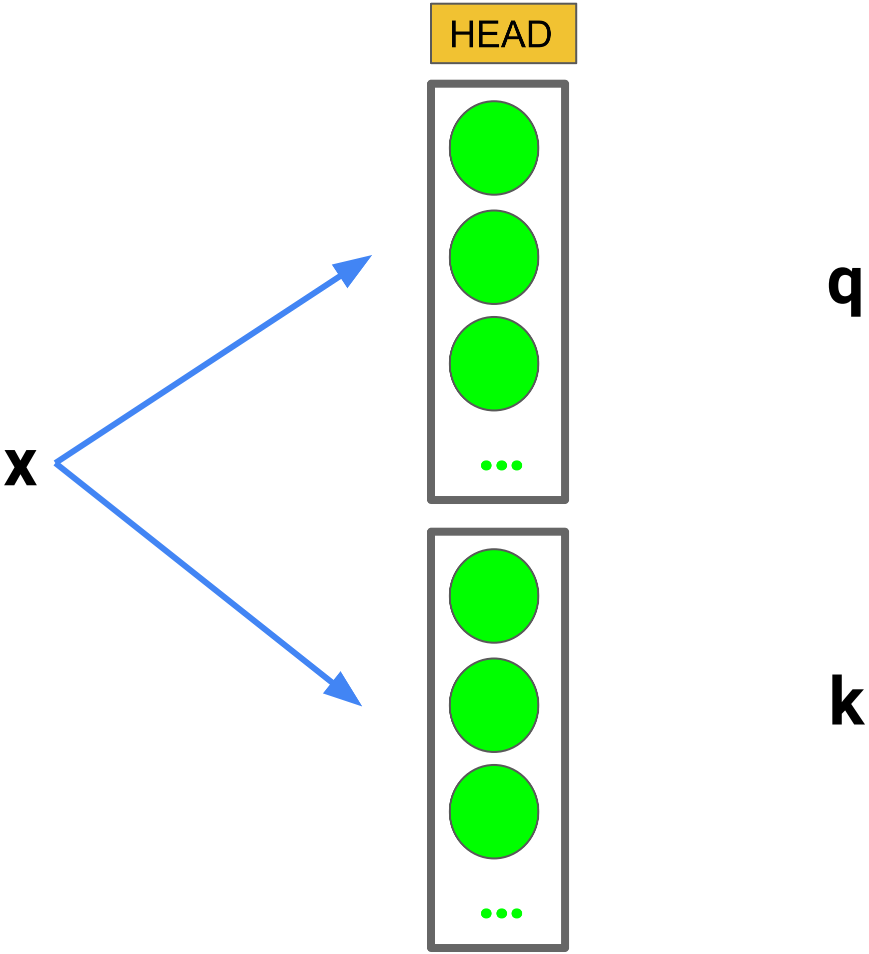


head

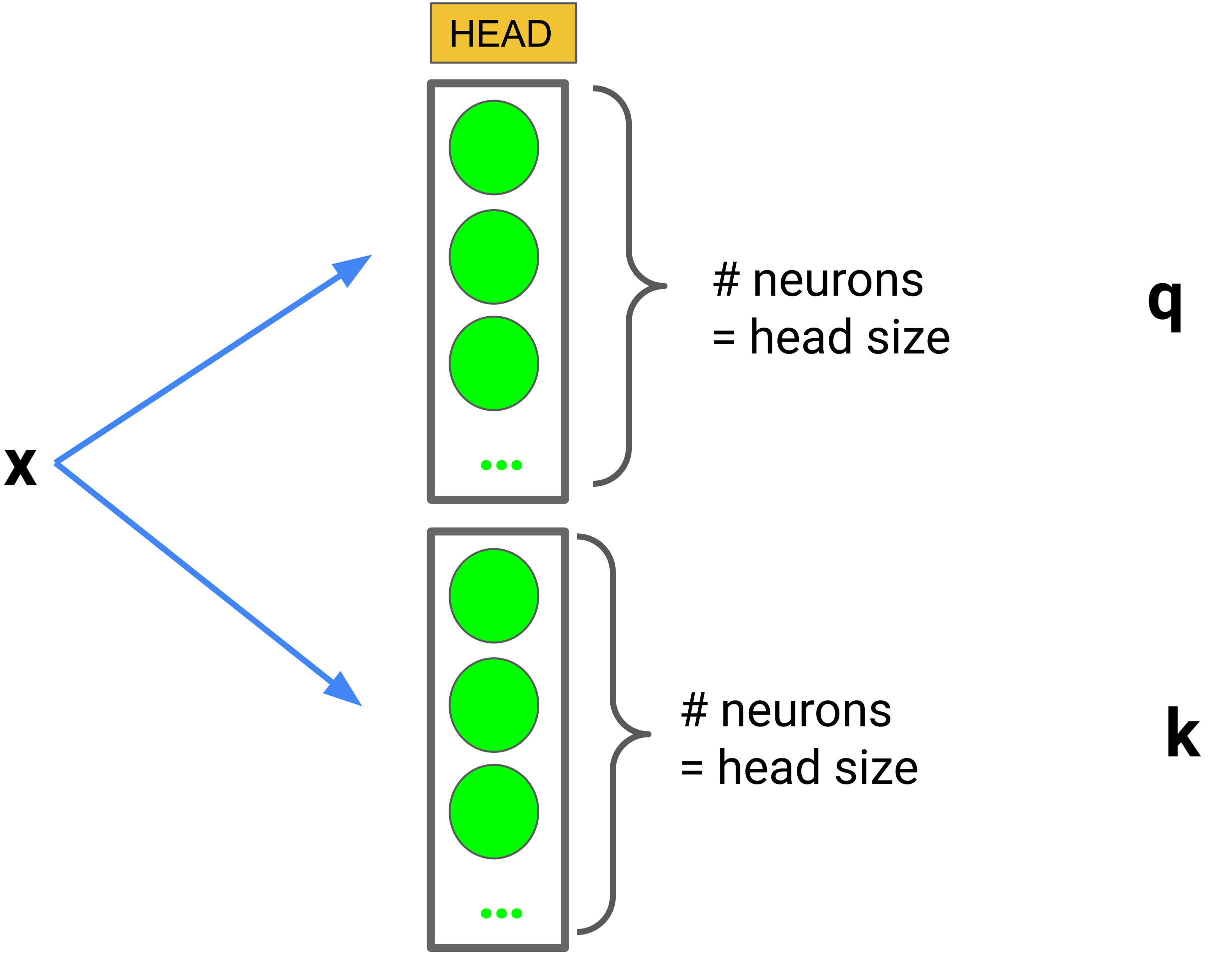
query What is the word looking for?

key What information does the word contain?

Self-attention head



Head size



CONSTELLATE

Query

$$\begin{array}{cc} \begin{array}{c} t1 \\ t2 \\ \dots \\ t8 \end{array} & \begin{bmatrix} w_1^{f_1} & w_1^{f_2} & \dots & w_1^{f_{32}} \\ w_2^{f_1} & w_2^{f_2} & \dots & w_2^{f_{32}} \\ \dots & \dots & \dots & \dots \\ w_8^{f_1} & w_8^{f_2} & \dots & w_8^{f_{32}} \end{bmatrix} \end{array}$$

Word sequence

$$\begin{bmatrix} q_1^{t1} & q_2^{t1} & \dots & q_{32}^{t1} \\ q_1^{t2} & q_2^{t2} & \dots & q_{32}^{t2} \\ \dots & \dots & \dots & \dots \\ q_1^{t8} & q_2^{t8} & \dots & q_{32}^{t8} \end{bmatrix}$$

Query matrix

CONSTELLATE

Key

$$\begin{array}{c} t1 \\ t2 \\ \ldots \\ t8 \end{array} \begin{bmatrix} w_1^{f_1} & w_1^{f_2} & \ldots & w_1^{f_{32}} \\ w_2^{f_1} & w_2^{f_2} & \ldots & w_2^{f_{32}} \\ \ldots & \ldots & \ldots & \ldots \\ w_8^{f_1} & w_8^{f_2} & \ldots & w_8^{f_{32}} \end{bmatrix}$$

Word sequence

$$\begin{bmatrix} k_1^{t1} & k_2^{t1} & \ldots & k_{32}^{t1} \\ k_1^{t2} & k_2^{t2} & \ldots & k_{32}^{t2} \\ \ldots & \ldots & \ldots & \ldots \\ k_1^{t8} & k_2^{t8} & \ldots & k_{32}^{t8} \end{bmatrix}$$

Key matrix

Key^T

$$\begin{array}{c} t1 \\ t2 \\ \dots \\ t8 \end{array} \begin{bmatrix} w_1^{f_1} & w_1^{f_2} & \dots & w_1^{f_{32}} \\ w_2^{f_1} & w_2^{f_2} & \dots & w_2^{f_{32}} \\ \dots & \dots & \dots & \dots \\ w_8^{f_1} & w_8^{f_2} & \dots & w_8^{f_{32}} \end{bmatrix}$$

Word sequence

$$\begin{bmatrix} k_1^{t1} & k_2^{t2} & \dots & k_8^{t8} \\ k_2^{t1} & k_2^{t2} & \dots & k_2^{t8} \\ \dots & \dots & \dots & \dots \\ k_{32}^{t1} & k_{32}^{t2} & \dots & k_{32}^{t8} \end{bmatrix}$$

Key matrix

Query · Key^T

$$\begin{array}{c|cccc} q^{t1} & q_1^{t1} & q_2^{t1} & \cdots & q_{32}^{t1} \\ q^{t2} & q_1^{t2} & q_2^{t2} & \cdots & q_{32}^{t2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ q^{t8} & q_1^{t8} & q_2^{t8} & \cdots & q_{32}^{t8} \end{array} \quad \bullet \quad \begin{array}{cccc} k^{t1} & k^{t2} & \cdots & k^{t8} \\ \hline k_1^{t1} & k_2^{t2} & \cdots & k_8^{t8} \\ k_2^{t1} & k_2^{t2} & \cdots & k_2^{t8} \\ \cdots & \cdots & \cdots & \cdots \\ k_{32}^{t1} & k_{32}^{t2} & \cdots & k_{32}^{t8} \end{array}$$

Query matrix

Key matrix

Attention score

$$\begin{bmatrix} \mathbf{q}^{t1} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t1} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t1} \cdot \mathbf{k}^{t8} \\ \mathbf{q}^{t2} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t8} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{q}^{t8} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t8} \end{bmatrix}$$

CONSTELLATE

But, what about the temporal dimension of language?

CONSTELLATE

Attention score

$$\begin{bmatrix} \mathbf{q}^{t1} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t1} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t1} \cdot \mathbf{k}^{t8} \\ \mathbf{q}^{t2} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t8} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{q}^{t8} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t8} \end{bmatrix}$$

Attention score

$$\begin{bmatrix} \mathbf{q}^{t1} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t1} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t1} \cdot \mathbf{k}^{t8} \\ \mathbf{q}^{t2} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t8} \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{q}^{t8} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t8} \end{bmatrix}$$

Attention score

-inf

$$\begin{bmatrix} \mathbf{q}^{t1} \cdot \mathbf{k}^{t1}, & \cancel{\mathbf{q}^{t1} \cdot \mathbf{k}^{t2},} & \cancel{\ldots,} & \cancel{\mathbf{q}^{t1} \cdot \mathbf{k}^{t8}} \\ \mathbf{q}^{t2} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t2} \cdot \mathbf{k}^{t2}, & \cancel{\ldots,} & \cancel{\mathbf{q}^{t2} \cdot \mathbf{k}^{t8}} \\ \ldots & \ldots & \ldots & \cancel{\ldots} \\ \mathbf{q}^{t8} \cdot \mathbf{k}^{t1}, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t2}, & \ldots, & \mathbf{q}^{t8} \cdot \mathbf{k}^{t8} \end{bmatrix}$$

Attention score

$$\begin{bmatrix} \mathbf{q}^{t1} \cdot \mathbf{k}^{t1} & \textbf{-inf} & \ldots & \ldots & \textbf{-inf} \\ \mathbf{q}^{t2} \cdot \mathbf{k}^{t1} & \mathbf{q}^{t2} \cdot \mathbf{k}^{t2} & \textbf{-inf} & \ldots & \textbf{-inf} \\ \ldots & \ldots & \ldots & \ldots & \textbf{-inf} \\ \mathbf{q}^{t8} \cdot \mathbf{k}^{t1} & \mathbf{q}^{t8} \cdot \mathbf{k}^{t2} & \mathbf{q}^{t8} \cdot \mathbf{k}^{t3} & \ldots & \mathbf{q}^{t8} \cdot \mathbf{k}^{t8} \end{bmatrix}$$

Attention scores to weights

$$\left[\begin{array}{ccccc} \mathbf{q}^{t1}\cdot\mathbf{k}^{t1} & \textbf{-inf} & \dots & \dots & \textbf{-inf} \\ \mathbf{q}^{t2}\cdot\mathbf{k}^{t1} & \mathbf{q}^{t2}\cdot\mathbf{k}^{t2} & \textbf{-inf} & \dots & \textbf{-inf} \\ \dots & \dots & \dots & \dots & \textbf{-inf} \\ \mathbf{q}^{t8}\cdot\mathbf{k}^{t1} & \mathbf{q}^{t8}\cdot\mathbf{k}^{t2} & \mathbf{q}^{t8}\cdot\mathbf{k}^{t3} & \dots & \mathbf{q}^{t8}\cdot\mathbf{k}^{t8} \end{array}\right] \begin{array}{l} \textbf{softmax} \\ \textbf{softmax} \\ \textbf{softmax} \\ \textbf{softmax} \end{array}$$

Attention scores to weights

$$\begin{array}{cl}
\begin{bmatrix}
P_{\mathbf{q}^{t1}\cdot\mathbf{k}^{t1}} & 0 & \dots & \dots & 0 \\
P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}} & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & 0 \\
P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t2}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t3}} & \dots & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t8}}
\end{bmatrix}
&
\begin{array}{l}
\text{sum up to 1} \\
\text{sum up to 1} \\
\text{sum up to 1} \\
\text{sum up to 1}
\end{array}
\end{array}$$

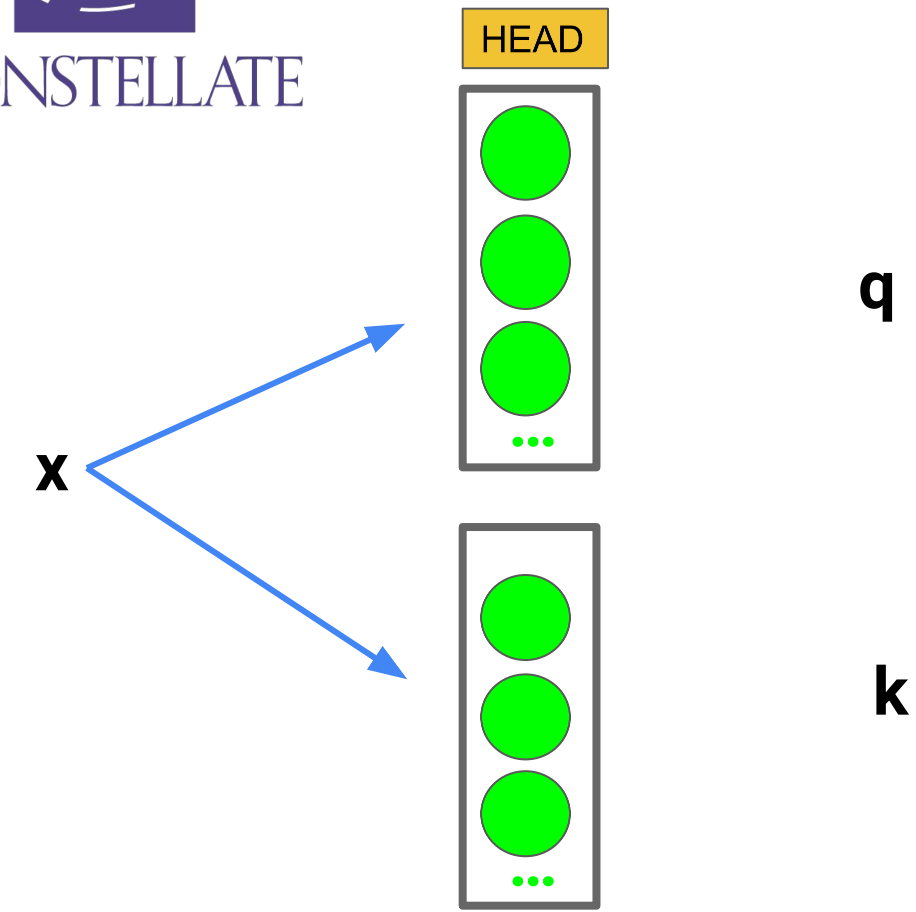
CONSTELLATE

Attention scores to weights

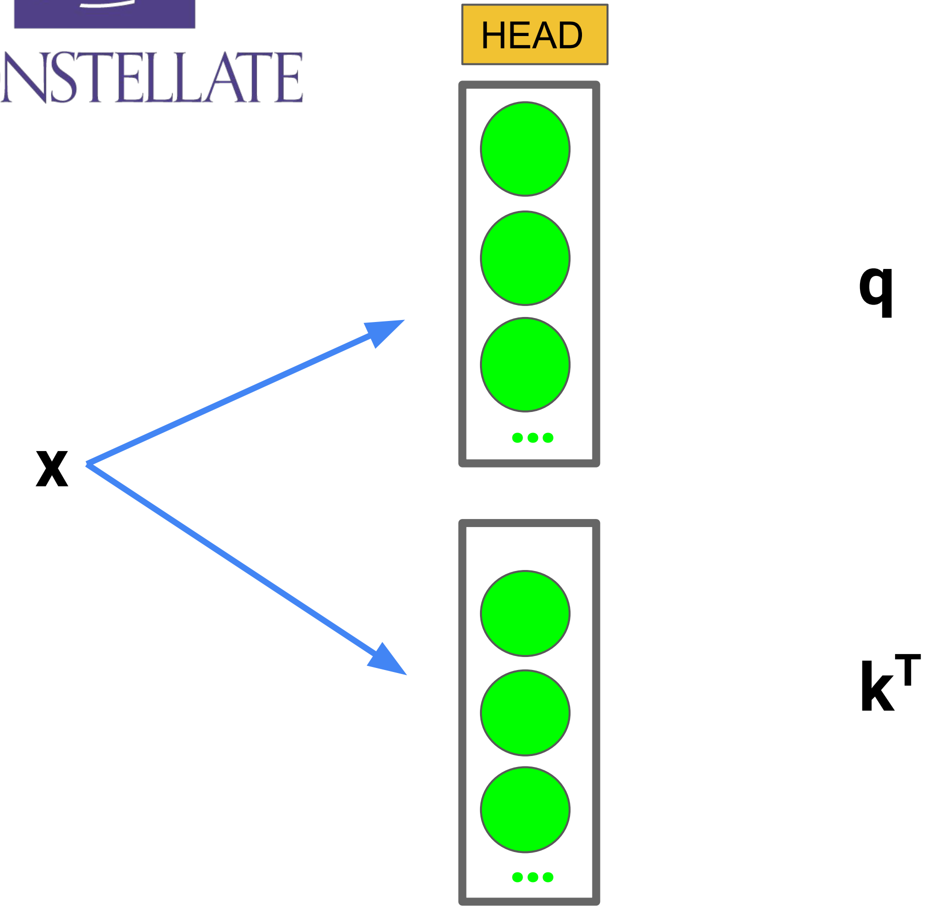
$$\left[\begin{array}{ccccc} 1 & 0 & 0 & \dots & 0 \\ 0.22 & 0.78 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0.02 & 0.12 & 0.24 & \dots & 0.45 \end{array}\right] \begin{array}{l} \textbf{sum up to 1} \\ \textbf{sum up to 1} \\ \textbf{sum up to 1} \\ \textbf{sum up to 1} \end{array}$$

weight matrix

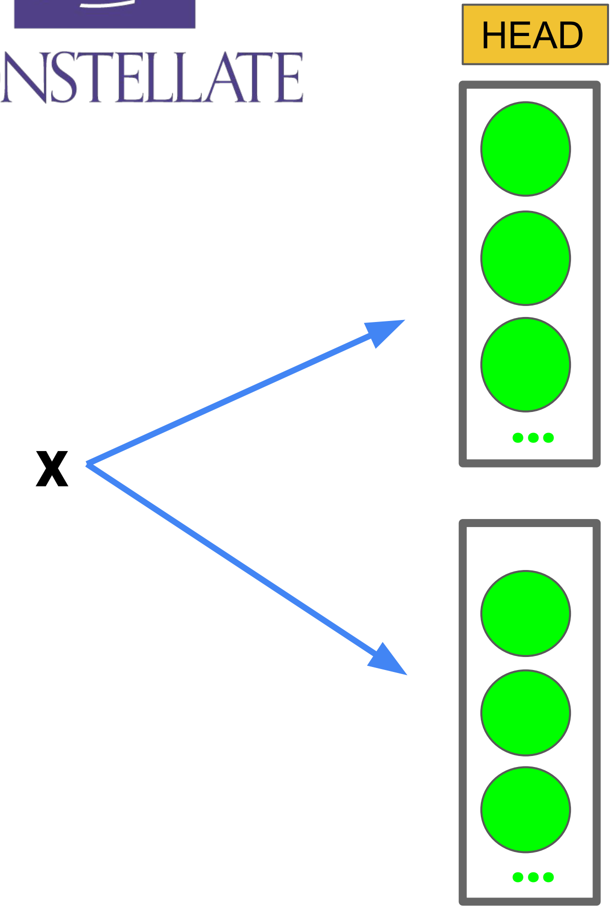
A visual of what we have learned so far



A visual of what we have learned so far



A visual of what we have learned so far

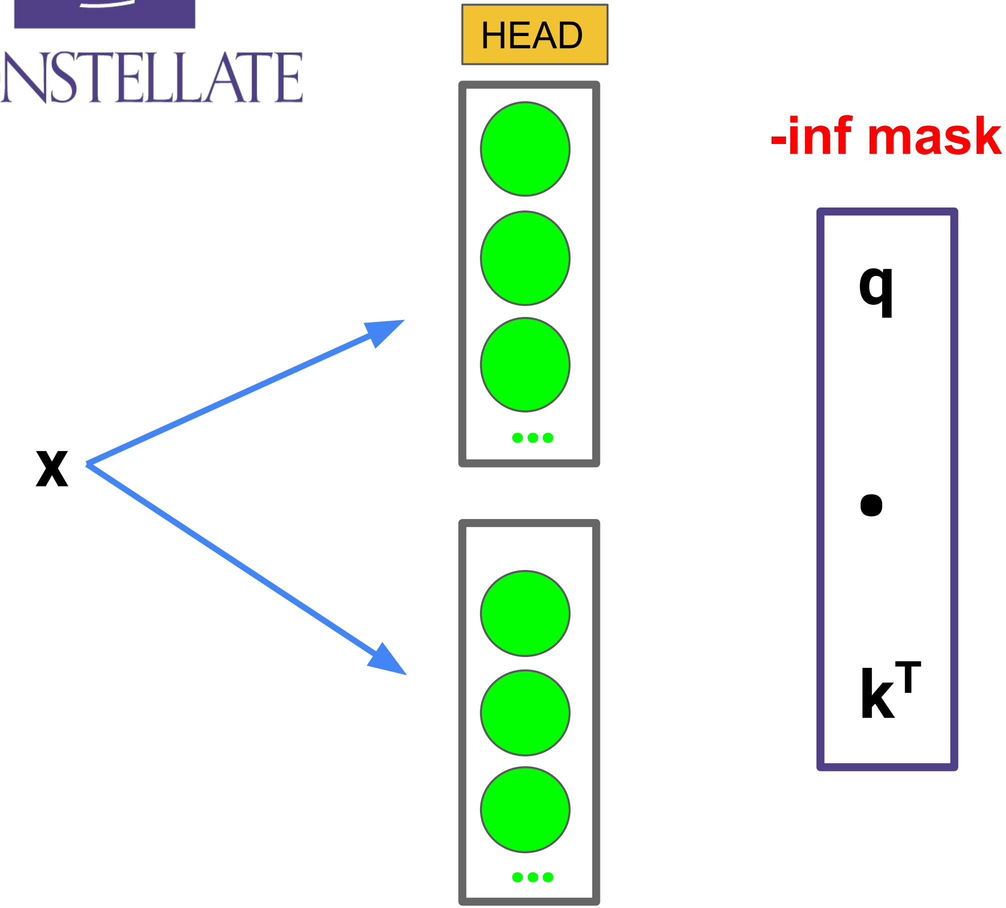


q

•

$\mathbf{k}^T$

A visual of what we have learned so far



HEAD

x

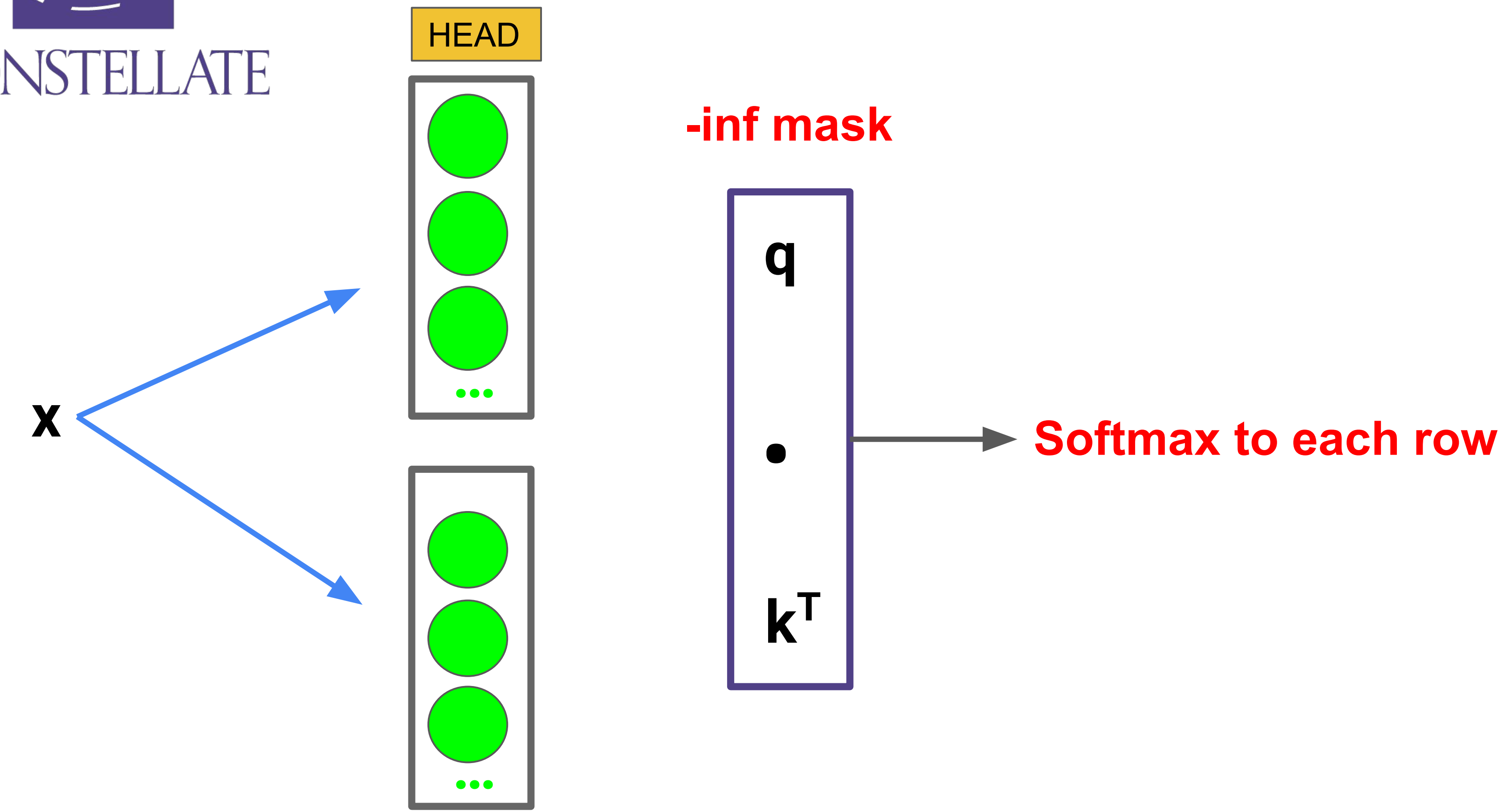
-inf mask

q

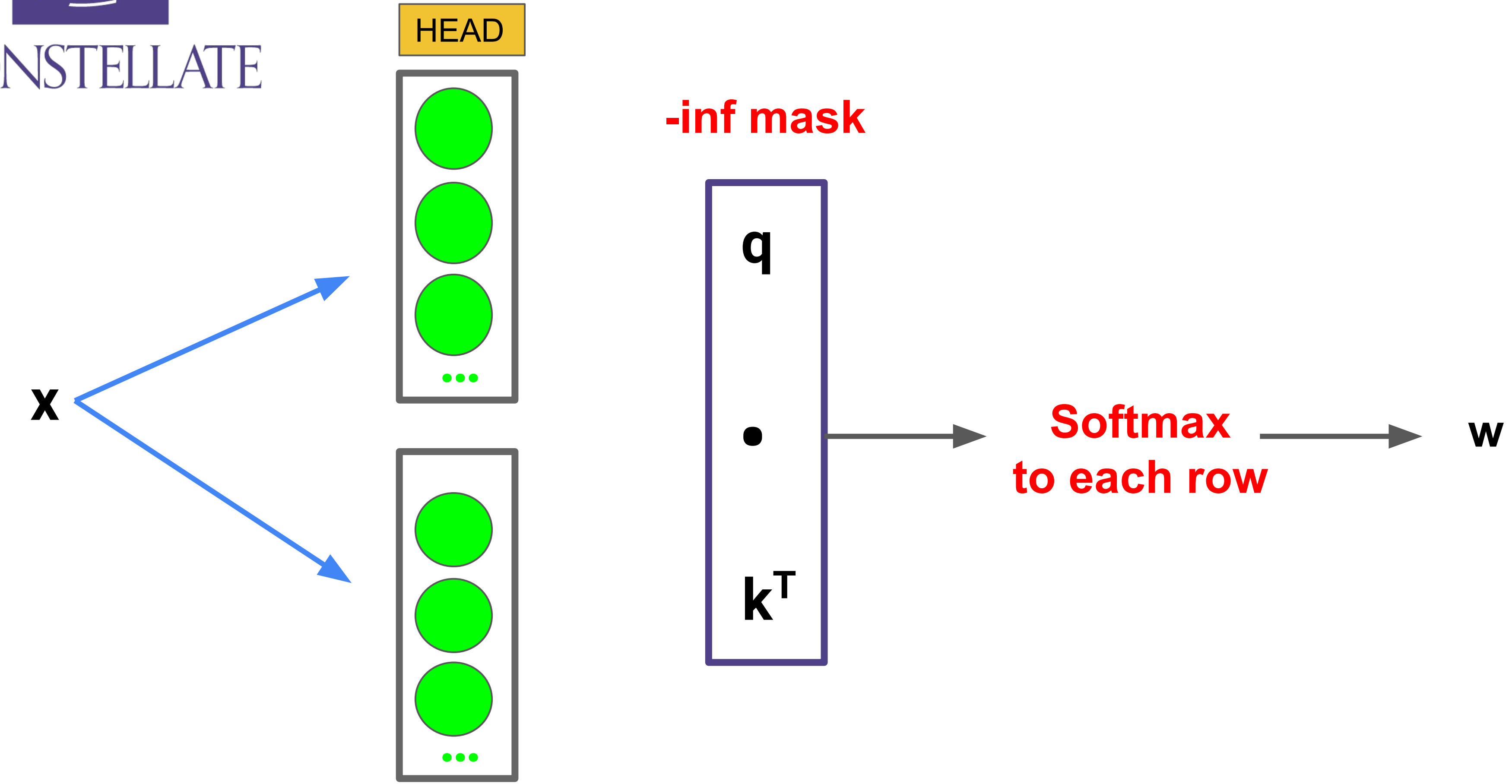
•

k^T

A visual of what we have learned so far



A visual of what we have learned so far



HEAD

x

-inf mask

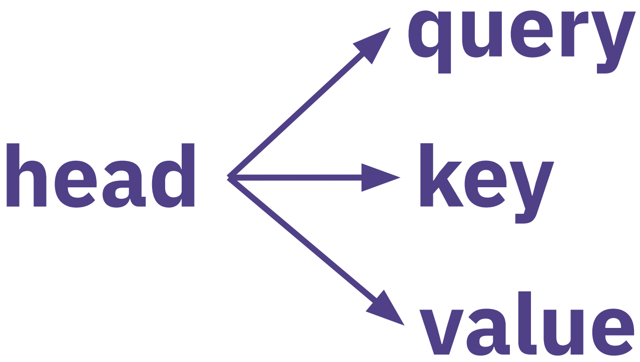
$q \cdot k^T$

Softmax
to each row

w

Self-attention head

For each word token



head

query

key

value

Value matrix

$$\begin{bmatrix} v_1^{t1} & v_2^{t1} & \dots & v_8^{t1} \\ v_1^{t2} & v_2^{t2} & \dots & v_8^{t2} \\ \dots & \dots & \dots & \dots \\ v_1^{t8} & v_2^{t8} & \dots & v_8^{t8} \end{bmatrix}$$

weight · value

$$\begin{bmatrix} P_{\mathbf{q}^{t1}\cdot\mathbf{k}^{t1}} & 0 & \dots & \dots & 0 \\ P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t2}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t3}} & \dots & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t8}} \end{bmatrix} \bullet \begin{bmatrix} v_1^{t1} & v_2^{t1} & \dots & v_8^{t1} \\ v_1^{t2} & v_2^{t2} & \dots & v_8^{t2} \\ \dots & \dots & \dots & \dots \\ v_1^{t8} & v_2^{t8} & \dots & v_8^{t8} \end{bmatrix}$$

weight · value

$$\begin{bmatrix} P_{\mathbf{q}^{t1}\cdot\mathbf{k}^{t1}} & 0 & \dots & \dots & 0 \\ \boxed{P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}}} & \boxed{P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}}} & \boxed{0} & \boxed{\dots} & \boxed{0} \\ \dots & \dots & \dots & \dots & 0 \\ P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t2}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t3}} & \dots & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t8}} \end{bmatrix} \bullet \begin{bmatrix} v_1^{t1} & v_2^{t1} & \dots & v_8^{t1} \\ v_1^{t2} & v_2^{t2} & \dots & v_8^{t2} \\ \dots & \dots & \dots & \dots \\ v_1^{t8} & v_2^{t8} & \dots & v_8^{t8} \end{bmatrix}$$

weight · value

$$\begin{bmatrix} P_{\mathbf{q}^{t1}\cdot\mathbf{k}^{t1}} & 0 & \dots & \dots & 0 \\ P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t1}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t2}} & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t3}} & \dots & P_{\mathbf{q}^{t8}\cdot\mathbf{k}^{t8}} \end{bmatrix} \bullet \begin{bmatrix} v_1^{t1} & v_2^{t1} & \dots & v_8^{t1} \\ v_1^{t2} & v_2^{t2} & \dots & v_8^{t2} \\ \dots & \dots & \dots & \dots \\ v_1^{t8} & v_2^{t8} & \dots & v_8^{t8} \end{bmatrix}$$

$$= \left[P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}} \cdot v_1^{t1} + P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}} \cdot v_1^{t2} \quad P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}} \cdot v_2^{t1} + P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}} \cdot v_2^{t2} \quad \dots \quad P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t1}} \cdot v_8^{t1} + P_{\mathbf{q}^{t2}\cdot\mathbf{k}^{t2}} \cdot v_8^{t2} \right]$$

References

Jurafsky, Daniel, and James H. Martin. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.*

Karpathy, Andrej. (2023), GitHub repository, https://github.com/karpathy/nanoGPT

Karpathy Andrej. (2023). Let's build GPT: from scratch, in code, spelled out. [Andrej Karpathy]. YouTube. Retrieved September 5, 2023 from https://www.youtube.com/watch?v=kCc8FmEb1nY