CONSTELLATE

How does ChatGPT work?

CONSTELLATE

Part 2

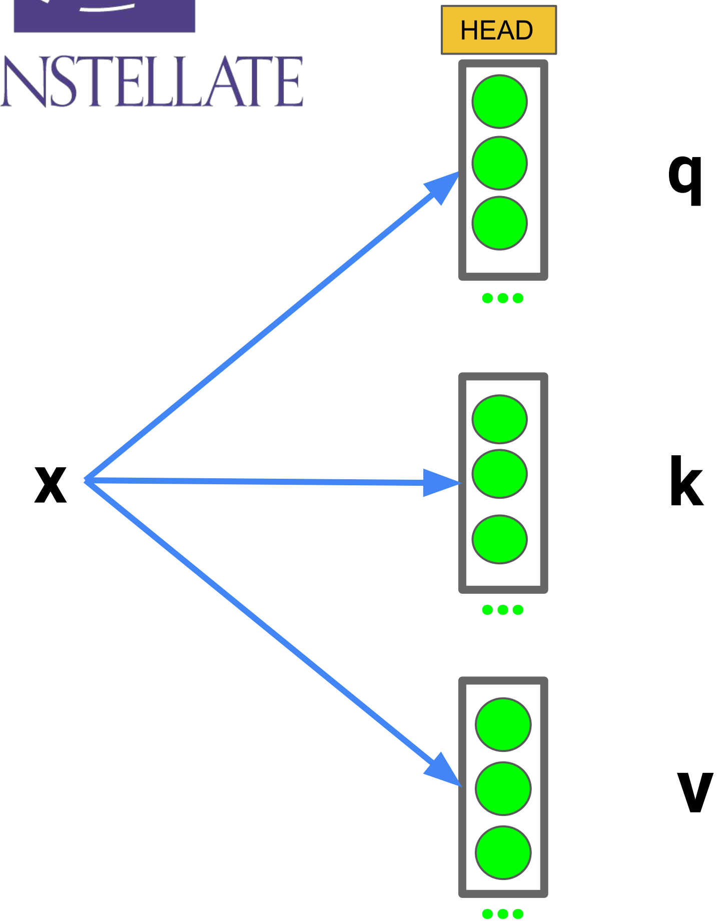
Learning objectives

- 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

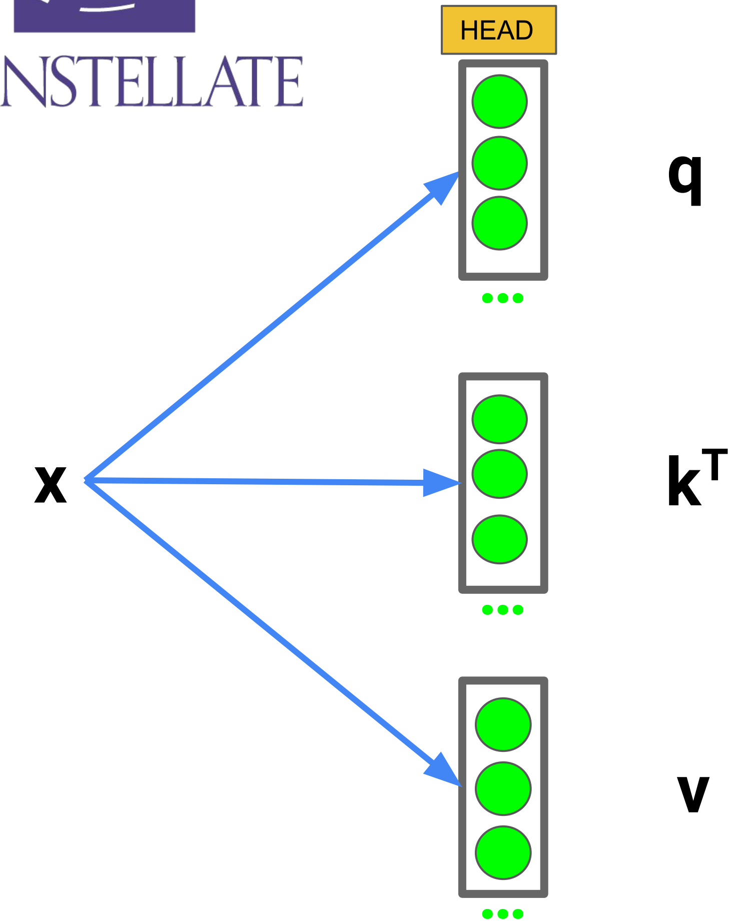
Last session

A single self-attention head

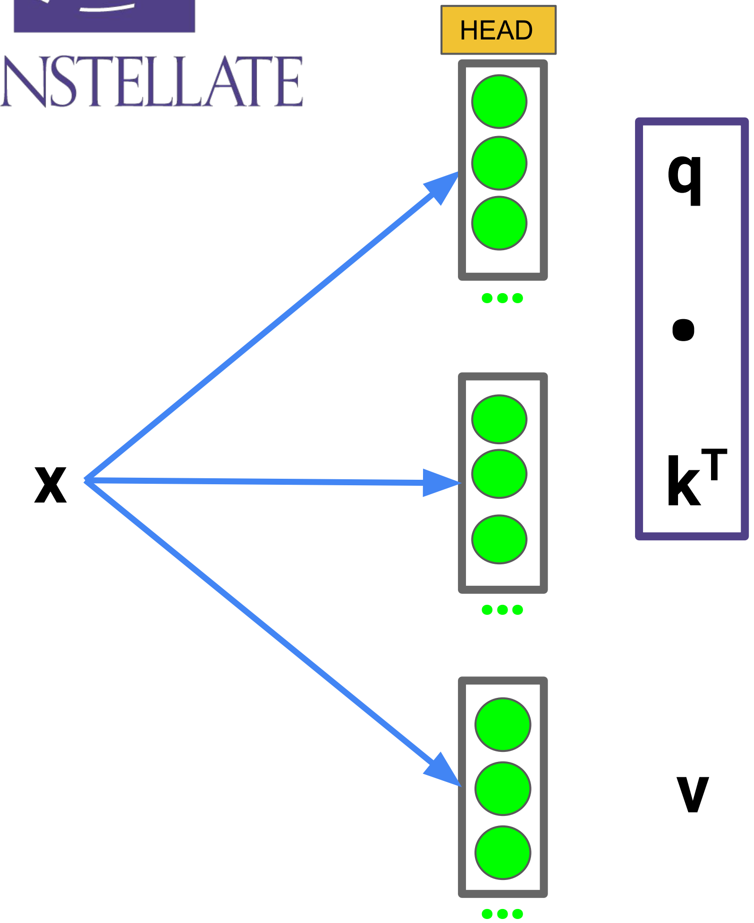
A visual of what we have learned so far



A visual of what we have learned so far

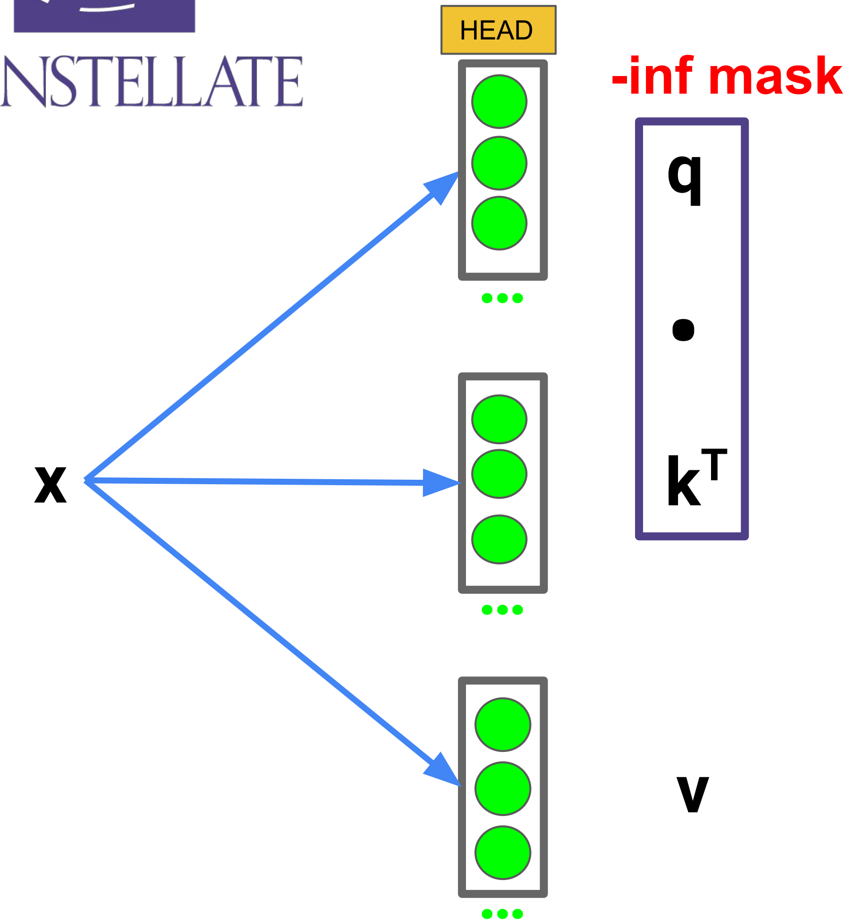


A visual of what we have learned so far



Attention scores

A visual of what we have learned so far



HEAD

-inf mask

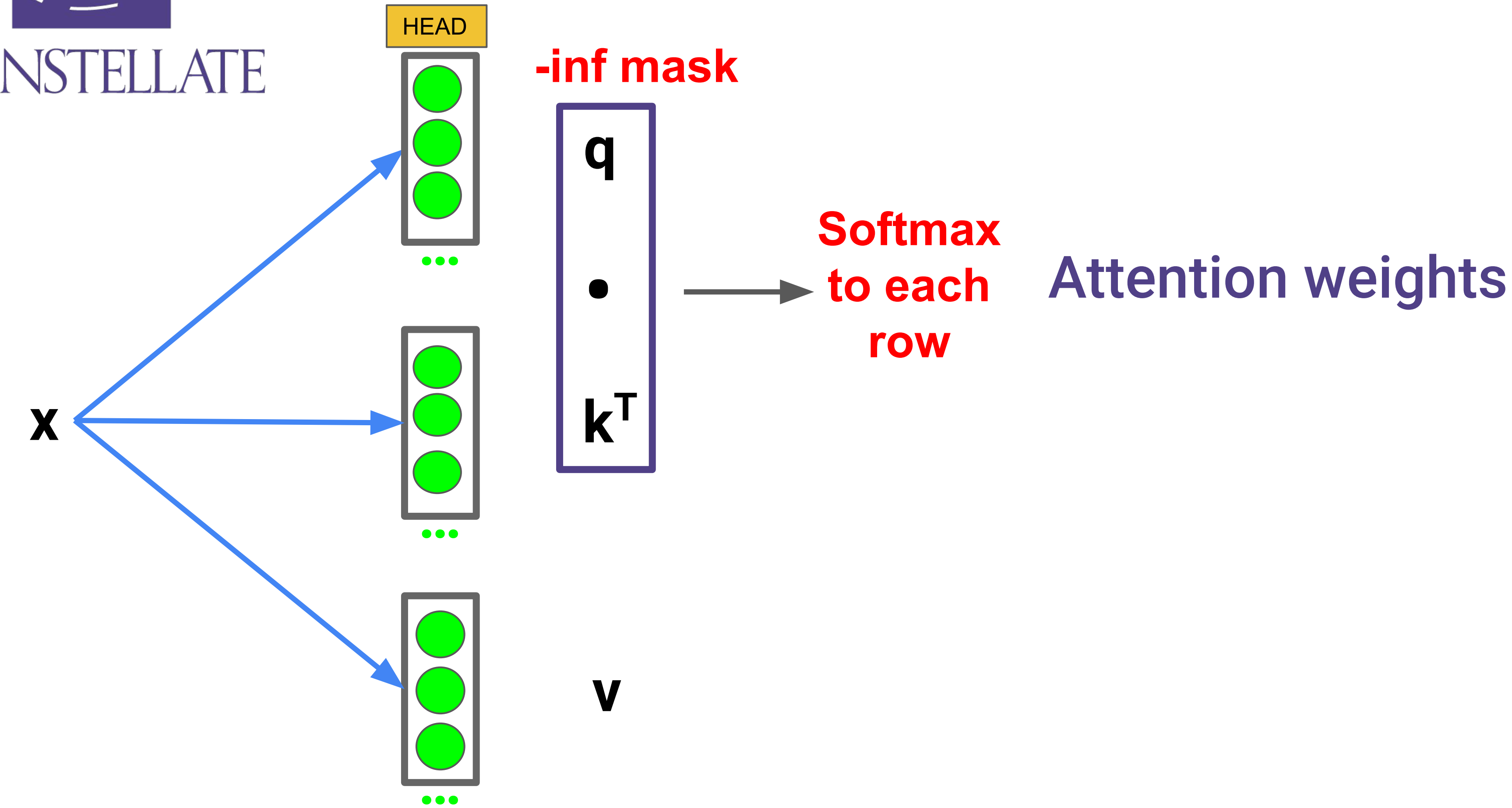
$q \bullet k^T$

x

v

A visual of what we have learned so far

CONSTELLATE



x

HEAD

-inf mask

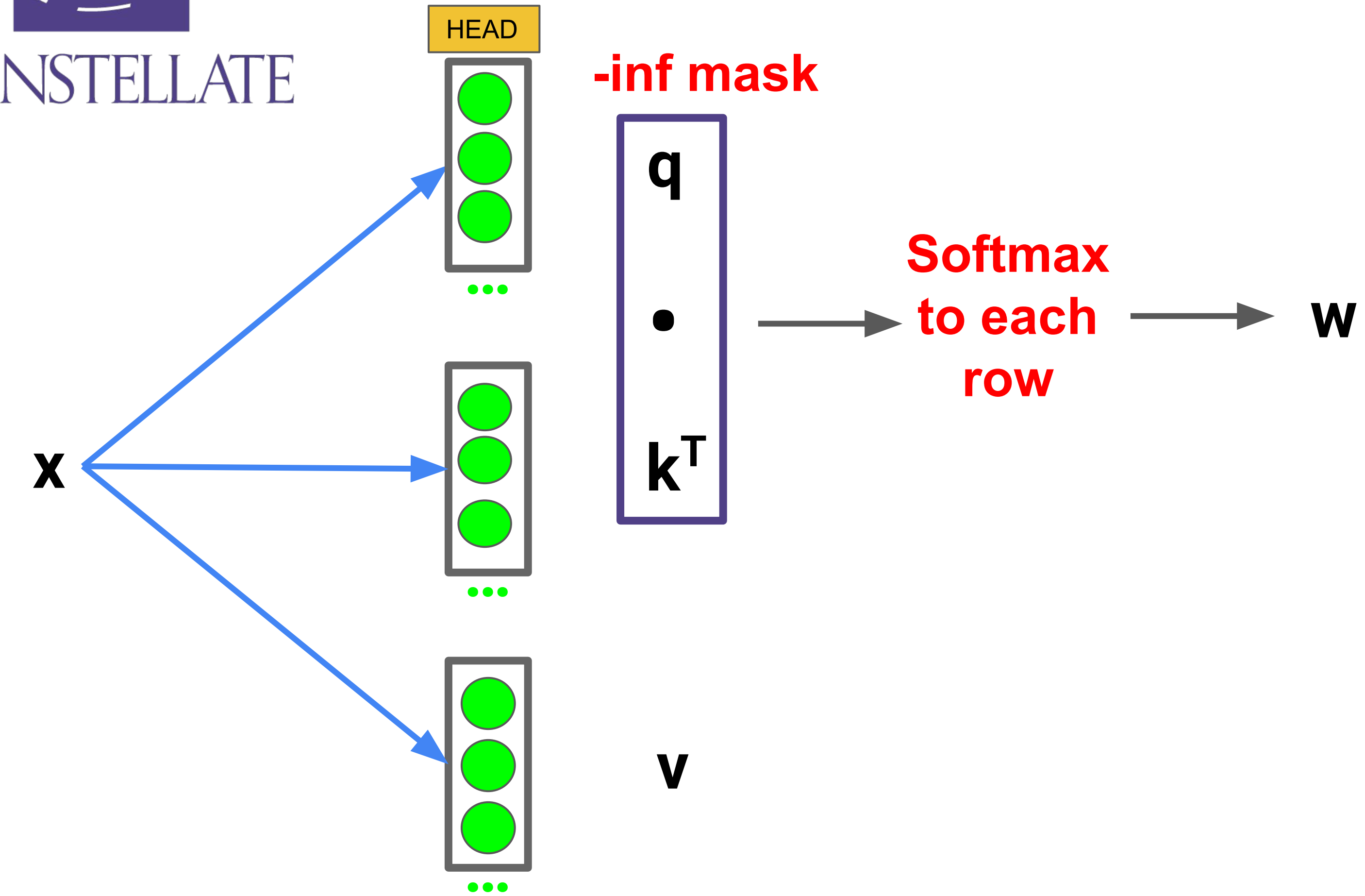
$q \cdot k^T$

v

Softmax to each row

Attention weights

A visual of what we have learned so far



HEAD

-inf mask

q

•

k^T

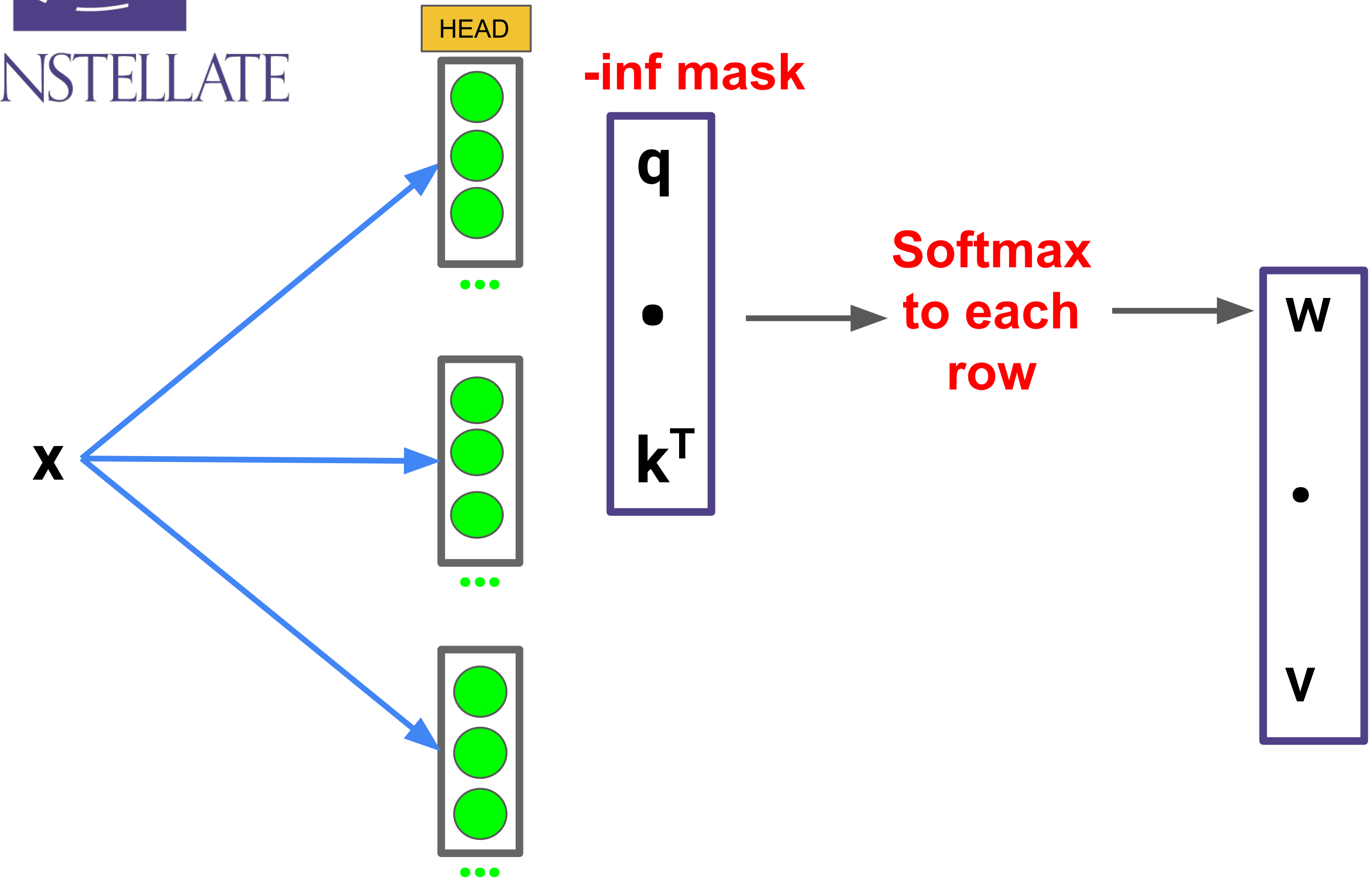
Softmax to each row

w

x

v

A visual of what we have learned so far



-inf mask

Softmax to each row

Any questions?

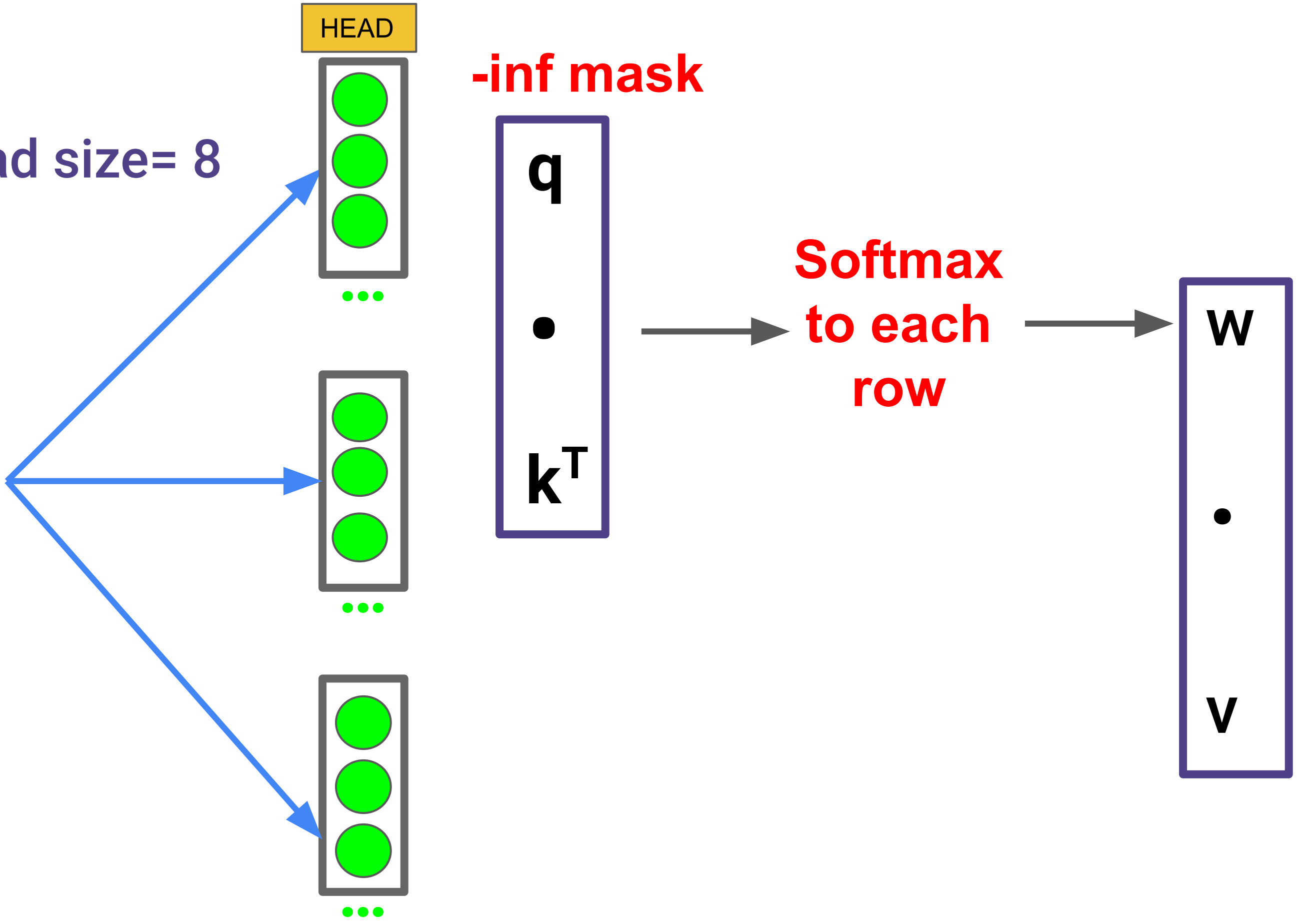
CONSTELLATE

Multi-head attention mechanism

A visual of what we have learned so far

$$\begin{bmatrix} x_1^{t1} & x_2^{t1} & \cdots & x_{32}^{t1} \\ x_1^{t2} & x_2^{t2} & \cdots & x_{32}^{t2} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{t8} & x_2^{t8} & \cdots & x_{32}^{t8} \end{bmatrix}$$

32 features



Head size= 8

HEAD

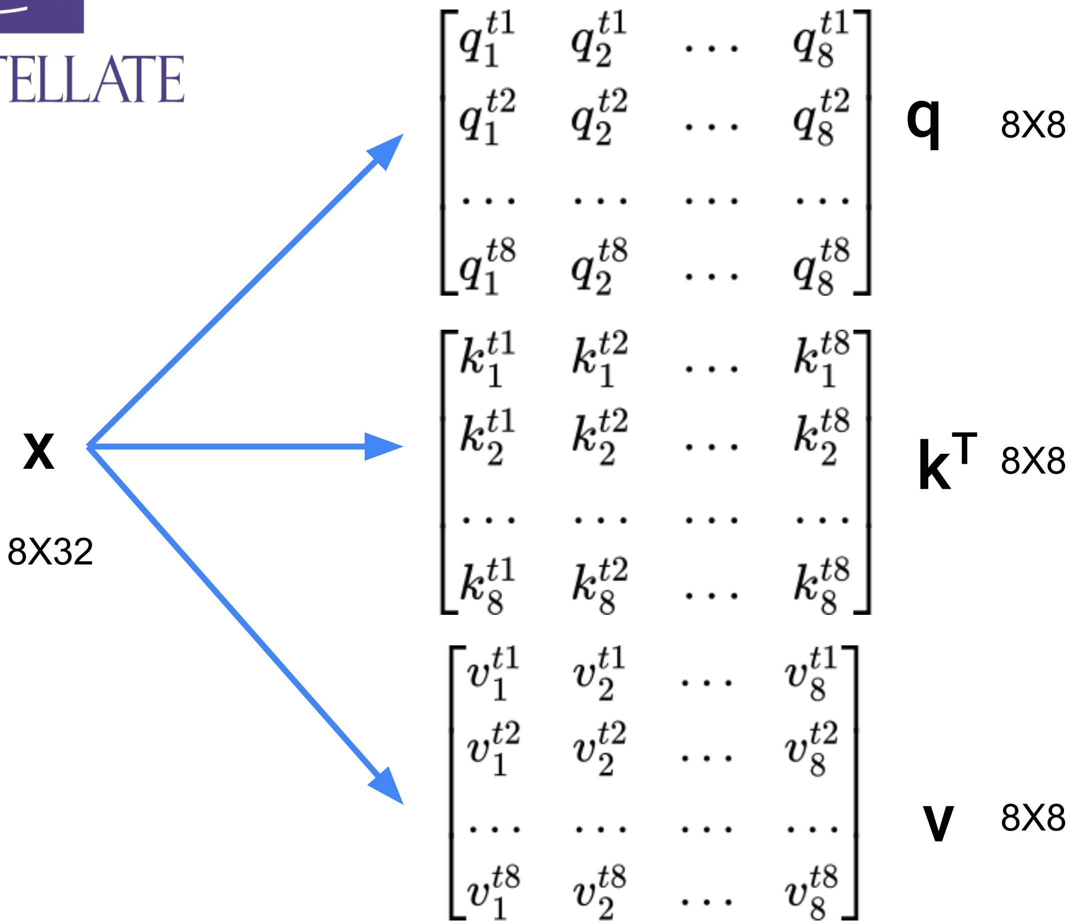
-inf mask

q • k^T

Softmax to each row

w • v

A single head

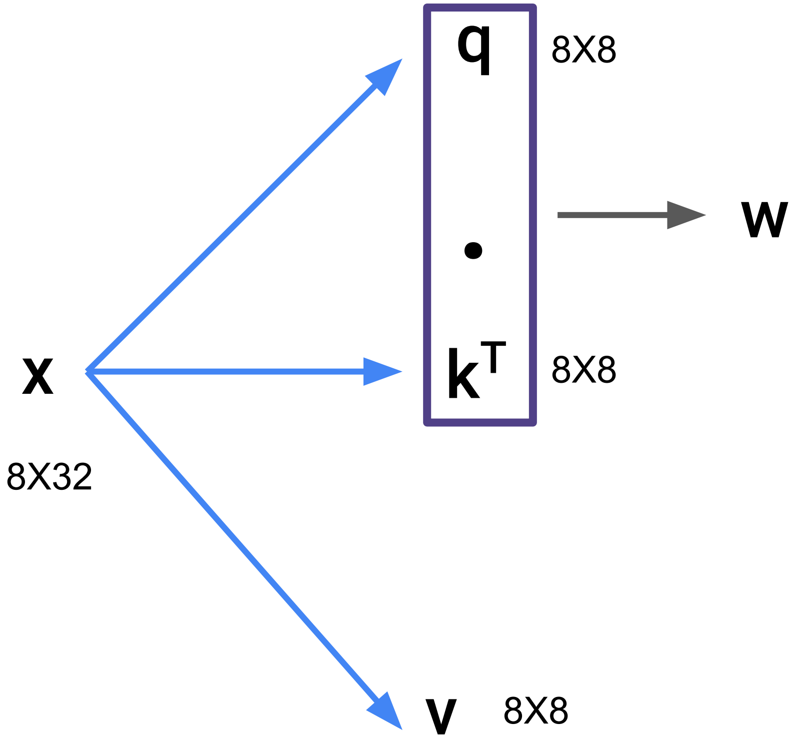


$$\mathbf{q} = \begin{bmatrix} q_1^{t1} & q_2^{t1} & \cdots & q_8^{t1} \\ q_1^{t2} & q_2^{t2} & \cdots & q_8^{t2} \\ \cdots & \cdots & \cdots & \cdots \\ q_1^{t8} & q_2^{t8} & \cdots & q_8^{t8} \end{bmatrix} \quad 8X8$$

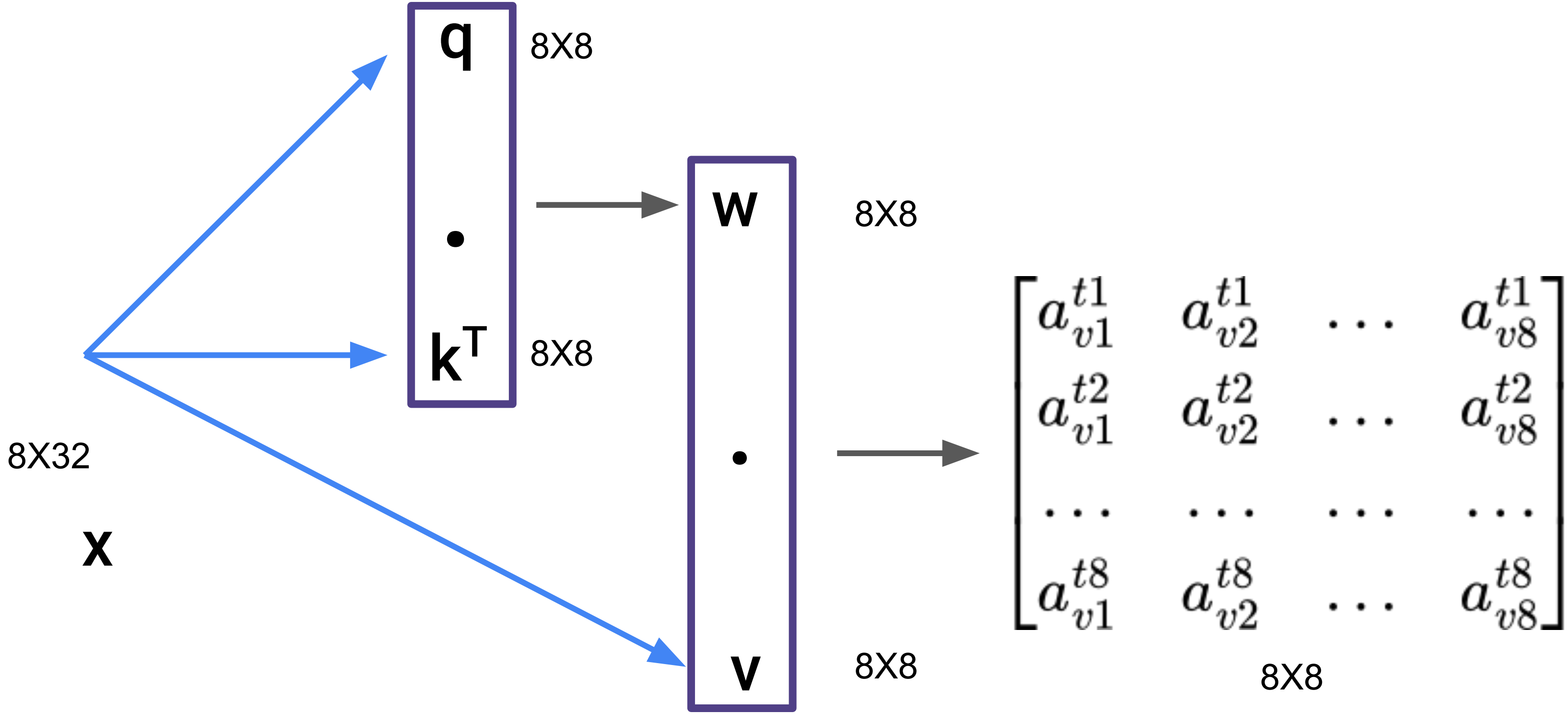
$$\mathbf{k}^T = \begin{bmatrix} k_1^{t1} & k_1^{t2} & \cdots & k_1^{t8} \\ k_2^{t1} & k_2^{t2} & \cdots & k_2^{t8} \\ \cdots & \cdots & \cdots & \cdots \\ k_8^{t1} & k_8^{t2} & \cdots & k_8^{t8} \end{bmatrix} \quad 8X8$$

$$\mathbf{v} = \begin{bmatrix} v_1^{t1} & v_2^{t1} & \cdots & v_8^{t1} \\ v_1^{t2} & v_2^{t2} & \cdots & v_8^{t2} \\ \cdots & \cdots & \cdots & \cdots \\ v_1^{t8} & v_2^{t8} & \cdots & v_8^{t8} \end{bmatrix} \quad 8X8$$

A single head



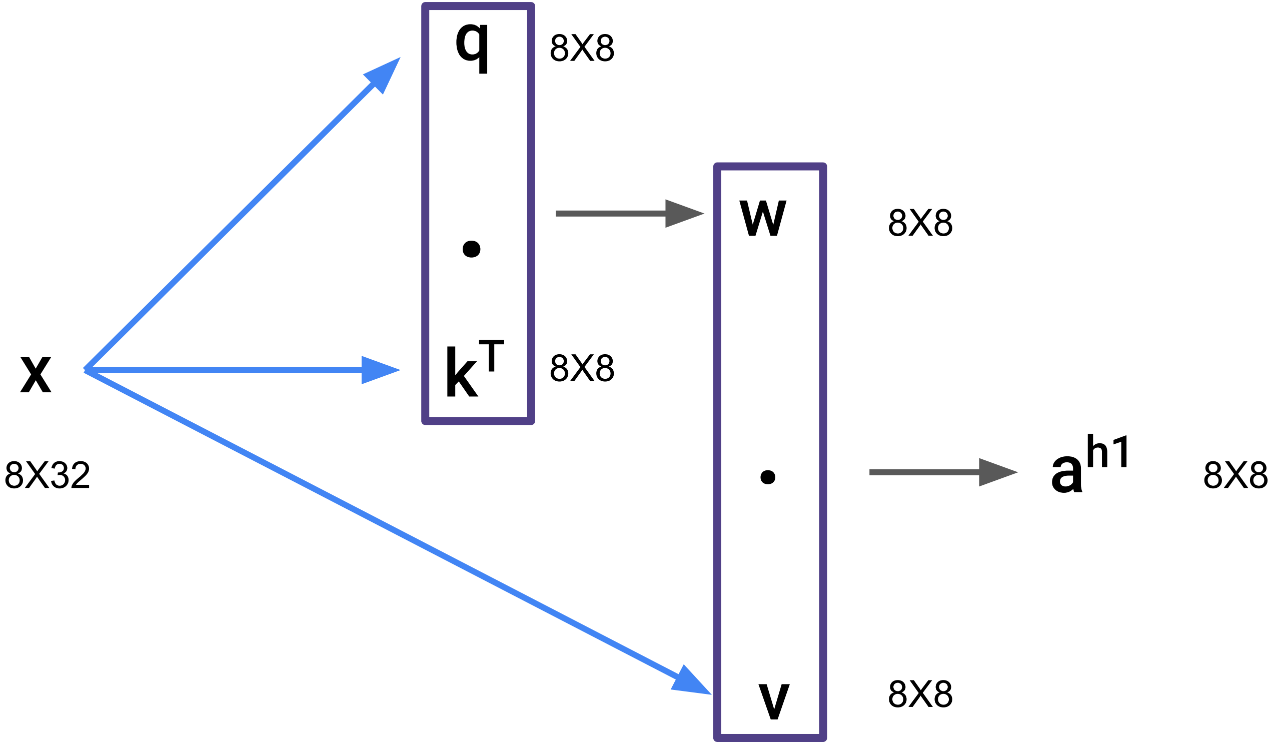
A single head



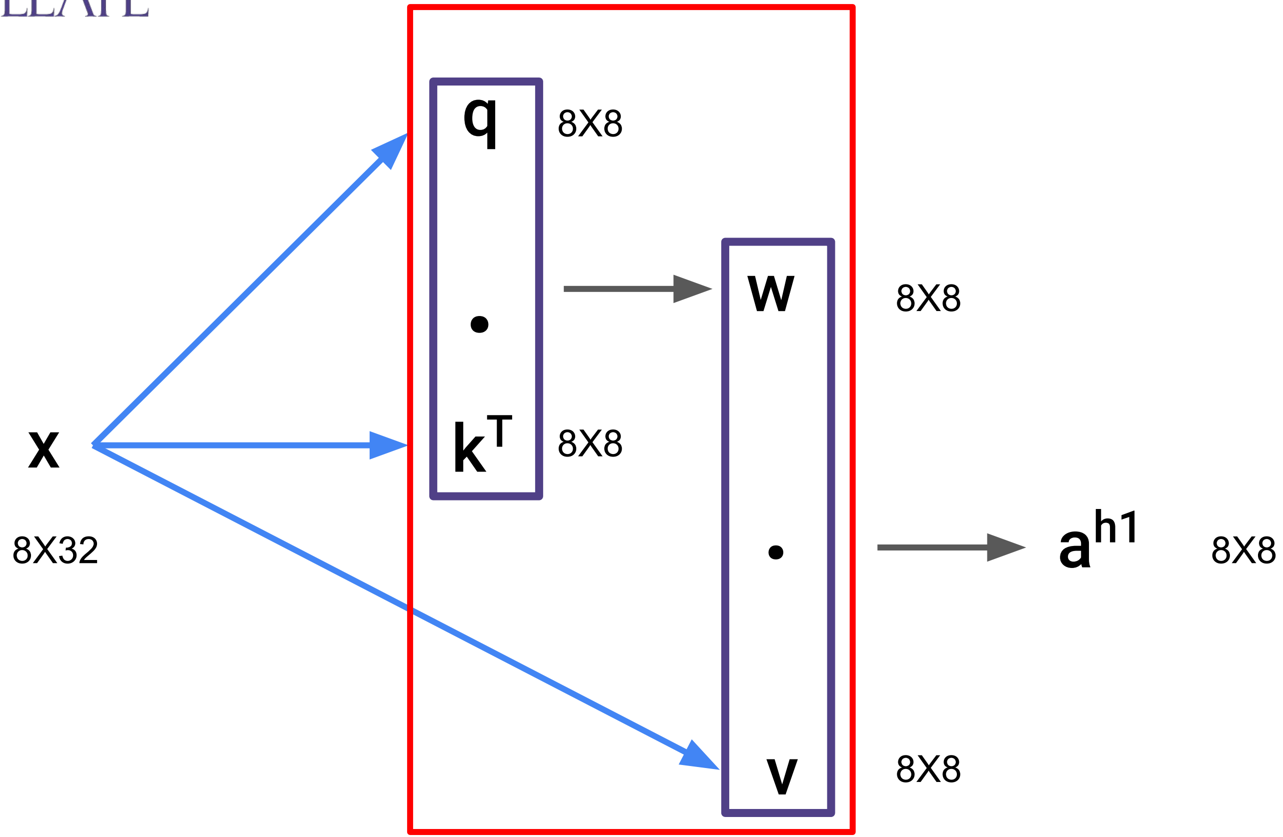
$$\begin{bmatrix} a_{v1}^{t1} & a_{v2}^{t1} & \dots & a_{v8}^{t1} \\ a_{v1}^{t2} & a_{v2}^{t2} & \dots & a_{v8}^{t2} \\ \dots & \dots & \dots & \dots \\ a_{v1}^{t8} & a_{v2}^{t8} & \dots & a_{v8}^{t8} \end{bmatrix}$$

CONSTELLATE

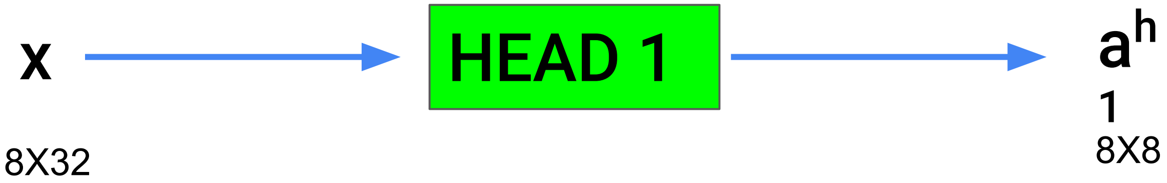
A single head



A single head



A single head

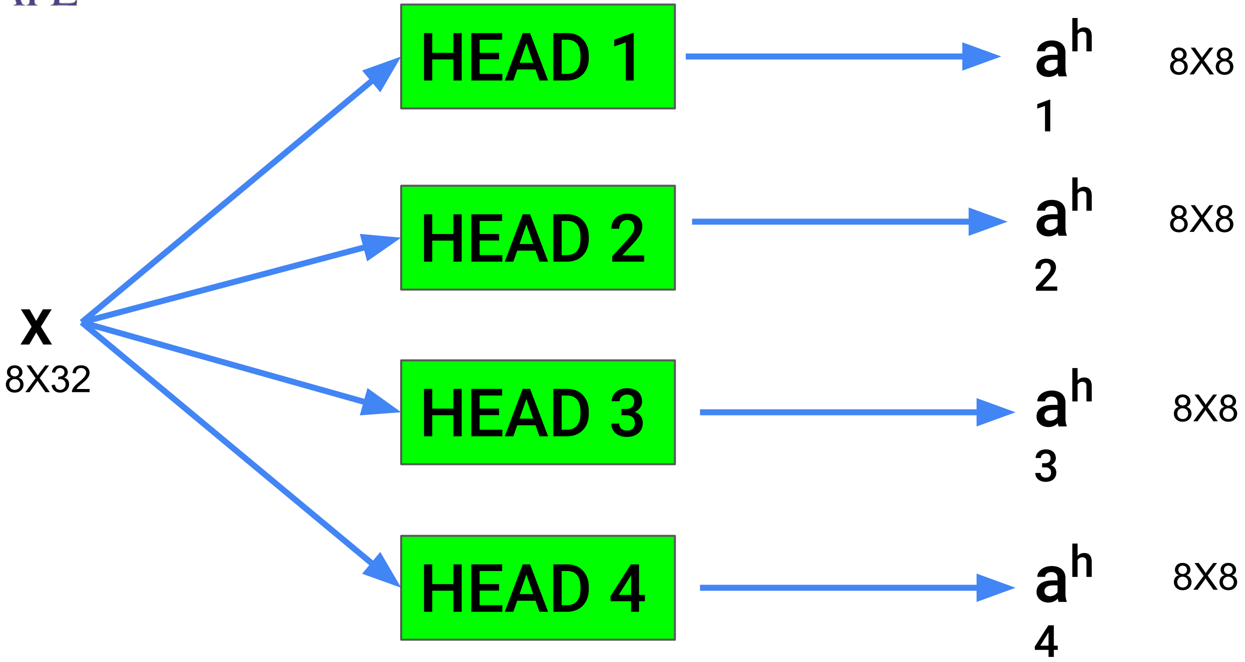


CONSTELLATE

Multi-head

CONSTELLATE

Multi-head



x

8X32

HEAD 1

HEAD 2

HEAD 3

HEAD 4

a^h_1 8X8

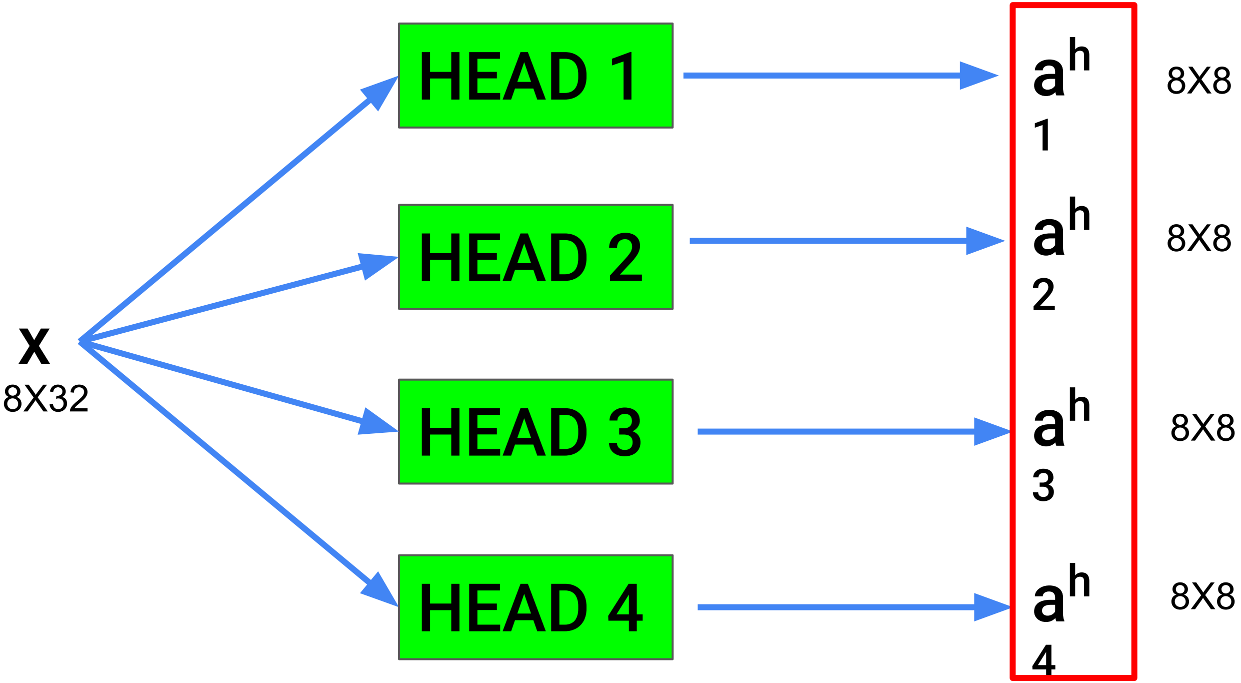
a^h_2 8X8

a^h_3 8X8

a^h_4 8X8

CONSTELLATE

Multi-head



x

8X32

HEAD 1

HEAD 2

HEAD 3

HEAD 4

a^h_1 8X8

a^h_2 8X8

a^h_3 8X8

a^h_4 8X8

Multi-head

$$\begin{bmatrix} a^{h1}_{t1_{v1}} & a^{h1}_{t1_{v2}} & \cdots & a^{h1}_{t1_{v8}} \\ a^{h1}_{t2_{v1}} & a^{h1}_{t2_{v2}} & \cdots & a^{h1}_{t2_{v8}} \\ \cdots & \cdots & \cdots & \cdots \\ a^{h1}_{t8_{v1}} & a^{h1}_{t8_{v2}} & \cdots & a^{h1}_{t8_{v8}} \end{bmatrix}$$

8X8

$$\begin{bmatrix} a^{h2}_{t1_{v1}} & a^{h2}_{t1_{v2}} & \cdots & a^{h2}_{t1_{v8}} \\ a^{h2}_{t2_{v1}} & a^{h2}_{t2_{v2}} & \cdots & a^{h2}_{t2_{v8}} \\ \cdots & \cdots & \cdots & \cdots \\ a^{h2}_{t8_{v1}} & a^{h2}_{t8_{v2}} & \cdots & a^{h2}_{t8_{v8}} \end{bmatrix}$$

8X8

Multi-head

$$\begin{bmatrix} a^{h1}_{t1_{v1}} & a^{h1}_{t1_{v2}} & \cdots & a^{h1}_{t1_{v8}} & a^{h2}_{t1_{v1}} & a^{h2}_{t1_{v2}} & \cdots & a^{h2}_{t1_{v8}} \\ a^{h1}_{t2_{v1}} & a^{h1}_{t2_{v2}} & \cdots & a^{h1}_{t2_{v8}} & a^{h2}_{t2_{v1}} & a^{h2}_{t2_{v2}} & \cdots & a^{h2}_{t2_{v8}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a^{h1}_{t8_{v1}} & a^{h1}_{t8_{v2}} & \cdots & a^{h1}_{t8_{v8}} & a^{h2}_{t8_{v1}} & a^{h2}_{t8_{v2}} & \cdots & a^{h2}_{t8_{v8}} \end{bmatrix}$$

8X16

Multi-head

8X32

$$\begin{bmatrix}
a^{h1}_{t1_{v1}} & \cdots & a^{h1}_{t1_{v8}} & a^{h2}_{t1_{v1}} & \cdots & a^{h2}_{t1_{v8}} & a^{h3}_{t1_{v1}} & \cdots & a^{h3}_{t1_{v8}} & a^{h4}_{t1_{v1}} & \cdots & a^{h4}_{t1_{v8}} \\
a^{h1}_{t2_{v1}} & \cdots & a^{h1}_{t2_{v8}} & a^{h2}_{t2_{v1}} & \cdots & a^{h2}_{t2_{v8}} & a^{h3}_{t2_{v1}} & \cdots & a^{h3}_{t2_{v8}} & a^{h4}_{t2_{v1}} & \cdots & a^{h4}_{t2_{v8}} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
a^{h1}_{t8_{v1}} & \cdots & a^{h1}_{t8_{v8}} & a^{h2}_{t8_{v1}} & \cdots & a^{h2}_{t8_{v8}} & a^{h3}_{t8_{v1}} & \cdots & a^{h3}_{t8_{v8}} & a^{h4}_{t8_{v1}} & \cdots & a^{h4}_{t8_{v8}}
\end{bmatrix}$$

HEAD 1

HEAD 2

HEAD 3

HEAD 4

Multi-head

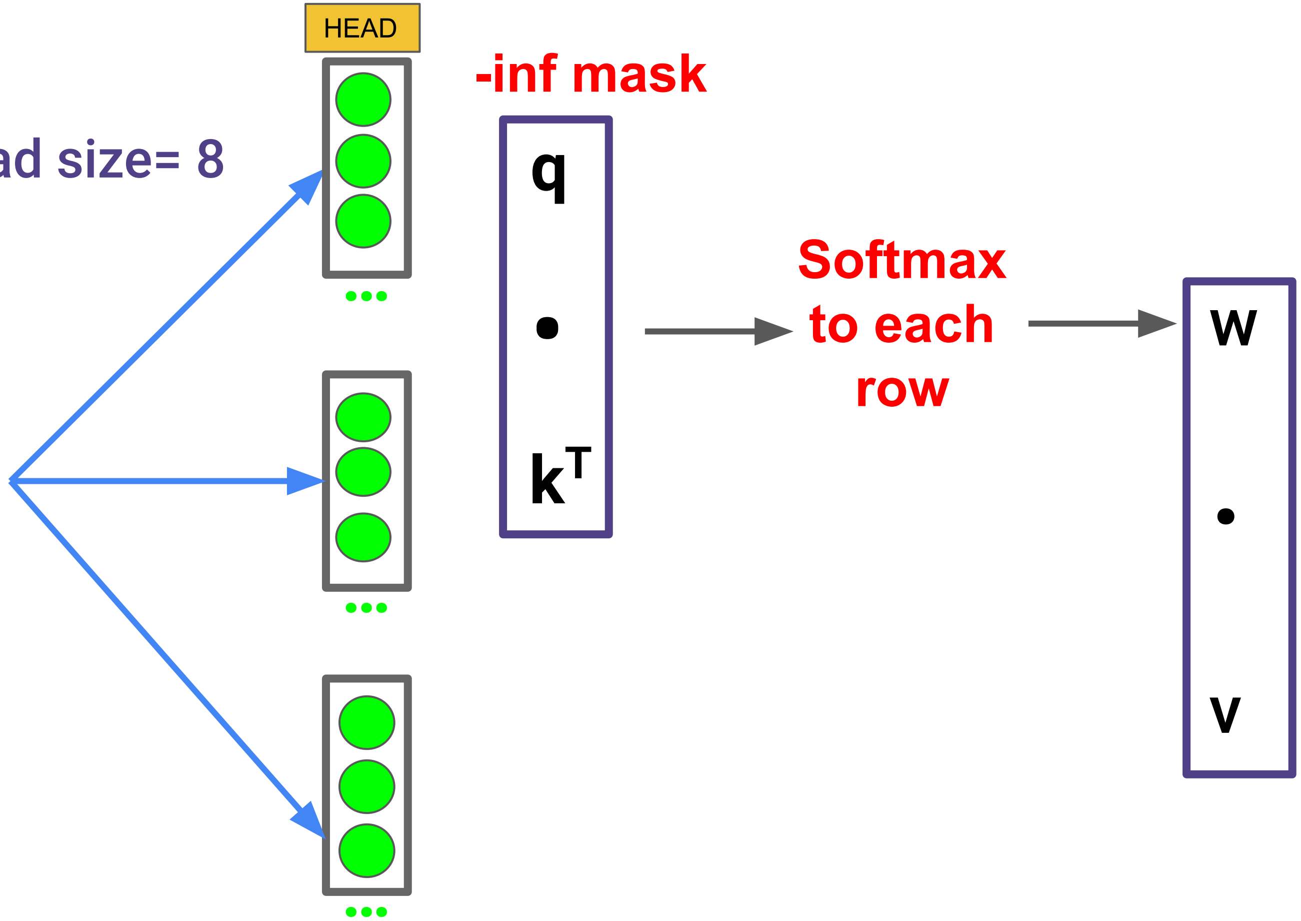
Still remember the shape of the original input?

A visual of what we have learned so far

$$\begin{bmatrix} x_1^{t1} & x_2^{t1} & \cdots & x_{32}^{t1} \\ x_1^{t2} & x_2^{t2} & \cdots & x_{32}^{t2} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{t8} & x_2^{t8} & \cdots & x_{32}^{t8} \end{bmatrix}$$

8X32

Head size= 8



HEAD

-inf mask

$q \bullet k^T$

Softmax to each row

$w \bullet v$

Multi-head

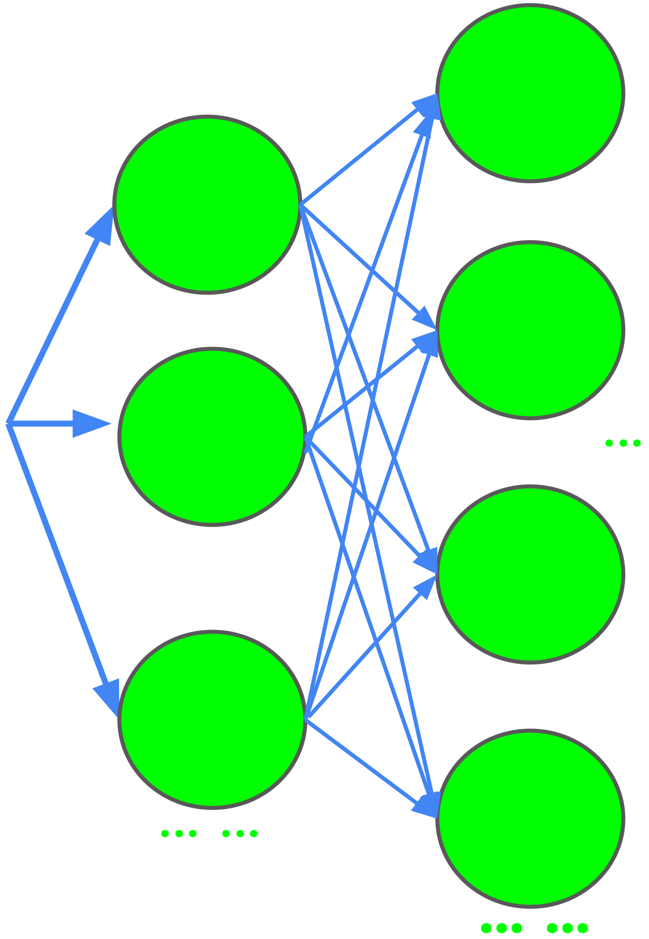
8X32

$$\left[\underbrace{\begin{matrix} a^{h1}_{t1_{v1}} & \cdots & a^{h1}_{t1_{v8}} \\ a^{h1}_{t2_{v1}} & \cdots & a^{h1}_{t2_{v8}} \\ \cdots & \cdots & \cdots \\ a^{h1}_{t8_{v1}} & \cdots & a^{h1}_{t8_{v8}} \end{matrix}}_{\text{HEAD 1}} \quad \underbrace{\begin{matrix} a^{h2}_{t1_{v1}} & \cdots & a^{h2}_{t1_{v8}} \\ a^{h2}_{t2_{v1}} & \cdots & a^{h2}_{t2_{v8}} \\ \cdots & \cdots & \cdots \\ a^{h2}_{t8_{v1}} & \cdots & a^{h2}_{t8_{v8}} \end{matrix}}_{\text{HEAD 2}} \quad \underbrace{\begin{matrix} a^{h3}_{t1_{v1}} & \cdots & a^{h3}_{t1_{v8}} \\ a^{h3}_{t2_{v1}} & \cdots & a^{h3}_{t2_{v8}} \\ \cdots & \cdots & \cdots \\ a^{h3}_{t8_{v1}} & \cdots & a^{h3}_{t8_{v8}} \end{matrix}}_{\text{HEAD 3}} \quad \underbrace{\begin{matrix} a^{h4}_{t1_{v1}} & \cdots & a^{h4}_{t1_{v8}} \\ a^{h4}_{t2_{v1}} & \cdots & a^{h4}_{t2_{v8}} \\ \cdots & \cdots & \cdots \\ a^{h4}_{t8_{v1}} & \cdots & a^{h4}_{t8_{v8}} \end{matrix}}_{\text{HEAD 4}}\right]$$

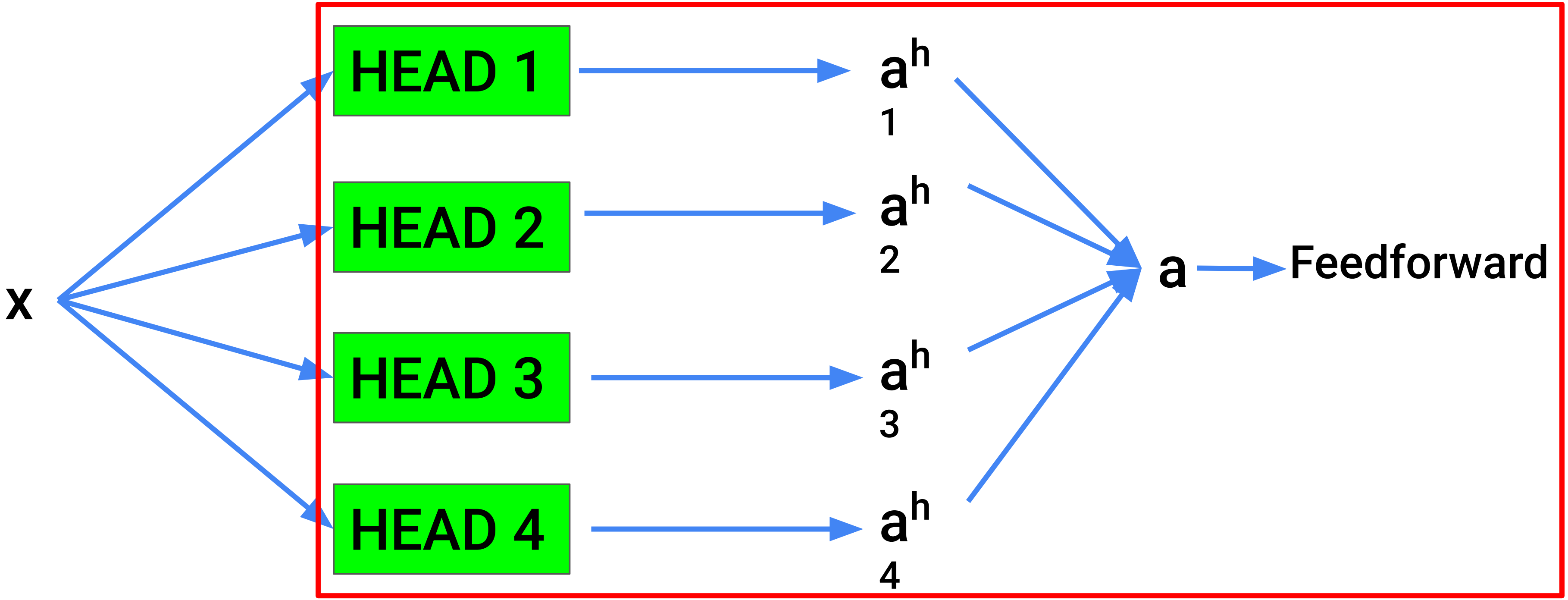
Feedforward

$$\begin{bmatrix}
a^{h1}_{t1_{v1}} & \dots & a^{h1}_{t1_{v8}} & a^{h2}_{t1_{v1}} & \dots & a^{h2}_{t1_{v8}} & a^{h3}_{t1_{v1}} & \dots & a^{h3}_{t1_{v8}} & a^{h4}_{t1_{v1}} & \dots & a^{h4}_{t1_{v8}} \\
a^{h1}_{t2_{v1}} & \dots & a^{h1}_{t2_{v8}} & a^{h2}_{t2_{v1}} & \dots & a^{h2}_{t2_{v8}} & a^{h3}_{t2_{v1}} & \dots & a^{h3}_{t2_{v8}} & a^{h4}_{t2_{v1}} & \dots & a^{h4}_{t2_{v8}} \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
a^{h1}_{t8_{v1}} & \dots & a^{h1}_{t8_{v8}} & a^{h2}_{t8_{v1}} & \dots & a^{h2}_{t8_{v8}} & a^{h3}_{t8_{v1}} & \dots & a^{h3}_{t8_{v8}} & a^{h4}_{t8_{v1}} & \dots & a^{h4}_{t8_{v8}}
\end{bmatrix}$$

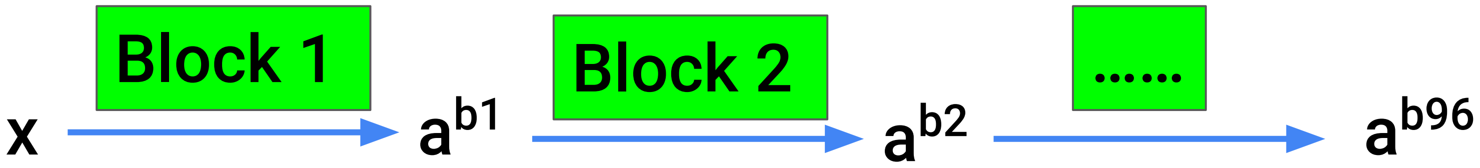
8X32



Attention block



N attention blocks



$$x \xrightarrow{\text{Block 1}} a^{b1} \xrightarrow{\text{Block 2}} a^{b2} \xrightarrow{\text{......}} a^{b96}$$

Any questions?

Positional encoding

Motivation

The order of the words in a sequence matters.

Positional encoding

Even though she did not win the award, she was satisfied.

Even though she did win the award, she was not satisfied.

Positional encoding

$$\begin{bmatrix} x_1^{t1} & x_2^{t1} & \dots & x_{32}^{t1} \\ x_1^{t2} & x_2^{t2} & \dots & x_{32}^{t2} \\ \dots & \dots & \dots & \dots \\ x_1^{t8} & x_2^{t8} & \dots & x_{32}^{t8} \end{bmatrix} + \begin{bmatrix} p_1^{t1} & p_2^{t1} & \dots & p_{32}^{t1} \\ p_1^{t2} & p_2^{t2} & \dots & p_{32}^{t2} \\ \dots & \dots & \dots & \dots \\ p_1^{t8} & p_2^{t8} & \dots & p_{32}^{t8} \end{bmatrix}$$

Positional encoding

$$\begin{bmatrix} x_1^{t1} & x_2^{t1} & \dots & x_{32}^{t1} \\ x_1^{t2} & x_2^{t2} & \dots & x_{32}^{t2} \\ \dots & \dots & \dots & \dots \\ x_1^{t8} & x_2^{t8} & \dots & x_{32}^{t8} \end{bmatrix} + \begin{bmatrix} p_1^{t1} & p_2^{t1} & \dots & p_{32}^{t1} \\ p_1^{t2} & p_2^{t2} & \dots & p_{32}^{t2} \\ \dots & \dots & \dots & \dots \\ p_1^{t8} & p_2^{t8} & \dots & p_{32}^{t8} \end{bmatrix}$$

Positional encoding

$$\begin{bmatrix} x_1^{t1} & x_2^{t1} & \dots & x_{32}^{t1} \\ x_1^{t2} & x_2^{t2} & \dots & x_{32}^{t2} \\ \dots & \dots & \dots & \dots \\ x_1^{t8} & x_2^{t8} & \dots & x_{32}^{t8} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \dots & 0 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 7 & 7 & \dots & 7 \end{bmatrix}$$

Positional encoding

$$\begin{array}{cccc} i=0 & i=1 & \ldots\ldots & i=32 \\ \left[x_1^{t1}\right. & x_2^{t1} & \ldots & \left.x_{32}^{t1}\right] \end{array}$$

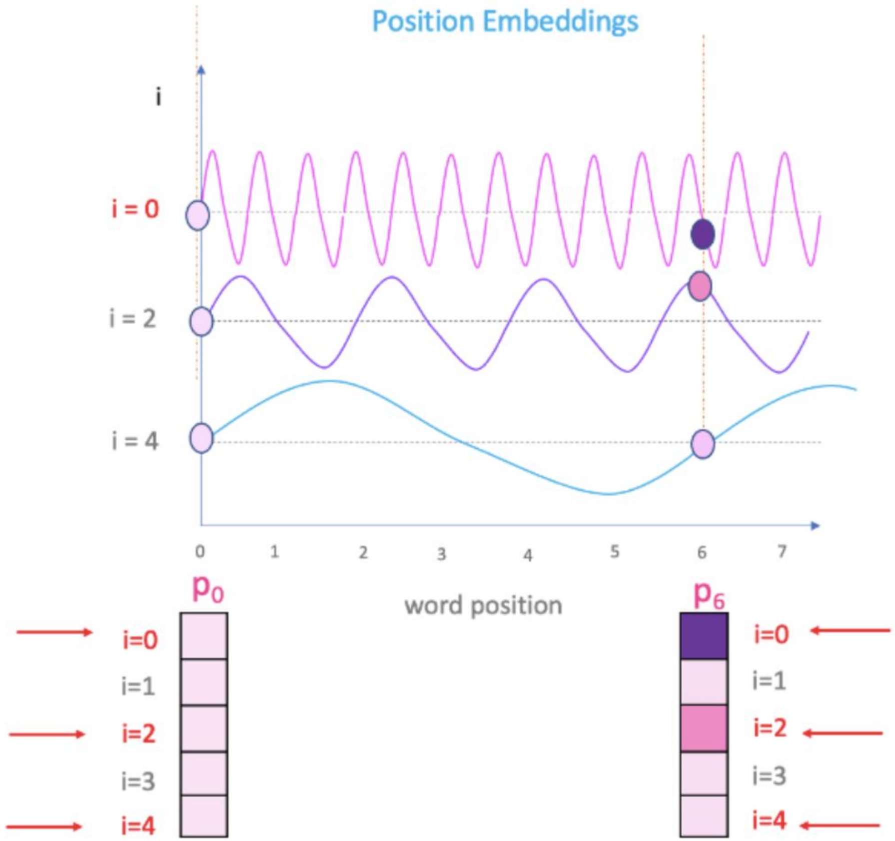
$$PE_{(}pos, 2i) = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

d=32

$$PE_{(}pos, 2i+1) = cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

Positional encoding

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$



Positional encoding

Even though she did not win the award, she was satisfied.

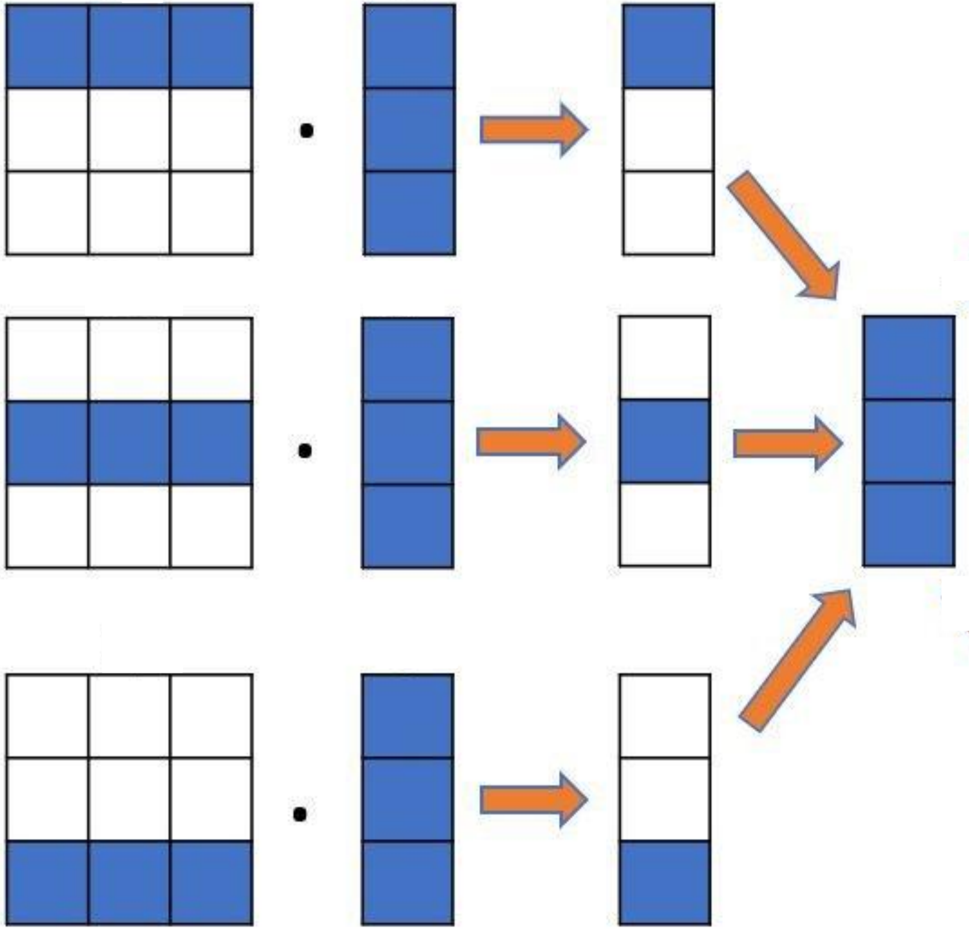
Even though she did win the award, she was not satisfied.

CONSTELLATE

Any questions?

CONSTELLATE

Parallel computation



CONSTELLATE

Calculate all observations at one time: Matrix multiplication

$$\begin{array}{c|cccc} & & & & \\ \mathbf{x}^{(1)} & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ \mathbf{x}^{(2)} & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ & \dots & \dots & \dots & \dots \\ \mathbf{x}^{(m)} & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{array} \bullet \begin{array}{cccc} \mathbf{W}^{c}_{1} & \mathbf{W}^{c}_{2} & & \mathbf{W}^{cK} \\ \hline w_1^{c1} & w_1^{c2} & \dots & w_1^{cK} \\ w_2^{c1} & w_2^{c2} & \dots & w_2^{cK} \\ \dots & \dots & \dots & \dots \\ w_n^{c1} & w_n^{c2} & \dots & w_n^{cK} \end{array} + \begin{bmatrix} \boxed{b^{c1}} & b^{c2} & \dots & b^{cK} \end{bmatrix}$$

$$= \begin{bmatrix} \boxed{\mathbf{w}^{c1}\mathbf{x}^{(1)} + b^{c1}} & \mathbf{w}^{c2}\mathbf{x}^{(1)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(1)} + b^{cK} \\ \mathbf{w}^{c1}\mathbf{x}^{(2)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(2)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(2)} + b^{cK} \\ \dots & \dots & \dots & \dots \\ \mathbf{w}^{c1}\mathbf{x}^{(m)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(m)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(m)} + b^{cK} \end{bmatrix}$$

Calculate all observations at one time: Matrix multiplication

$$
\begin{array}{c}
\mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \\ \mathbf{x}^{(m)}
\end{array}
\begin{bmatrix}
x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\
\dots & \dots & \dots & \dots \\
x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)}
\end{bmatrix}
\bullet
\begin{array}{cccc}
\mathbf{w}^{\mathrm{c}}_{1} & \mathbf{w}^{\mathrm{c}}_{2} & & \mathbf{w}^{\mathrm{cK}} \\
\left[\begin{matrix}
w_1^{c1} & w_1^{c2} & \dots & w_1^{cK} \\
w_2^{c1} & w_2^{c2} & \dots & w_2^{cK} \\
\dots & \dots & \dots & \dots \\
w_n^{c1} & w_n^{c2} & \dots & w_n^{cK}
\end{matrix}\right]
\end{array}
+ \begin{bmatrix} b^{c1} & b^{c2} & \dots & \boxed{b^{cK}} \end{bmatrix}
$$

$$
= \begin{bmatrix}
\mathbf{w}^{c1}\mathbf{x}^{(1)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(1)} + b^{c2} & \dots & \boxed{\mathbf{w}^{cK}\mathbf{x}^{(1)} + b^{cK}} \\
\mathbf{w}^{c1}\mathbf{x}^{(2)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(2)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(2)} + b^{cK} \\
\dots & \dots & \dots & \dots \\
\mathbf{w}^{c1}\mathbf{x}^{(m)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(m)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(m)} + b^{cK}
\end{bmatrix}
$$

Calculate all observations at one time: Matrix multiplication

$$
\begin{array}{r}
\mathbf{x}^{(1)} \\[1em] \mathbf{x}^{(2)} \\[1em] \\[1em] \mathbf{x}^{(m)}
\end{array}
\begin{bmatrix}
x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\
x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\
\dots & \dots & \dots & \dots \\
x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)}
\end{bmatrix}
\bullet
\begin{array}{c}
\begin{array}{cccc}
\mathbf{w}^{c}_{1} & \mathbf{w}^{c}_{2} & & \mathbf{w}^{cK}
\end{array} \\
\begin{bmatrix}
w_1^{c1} & w_1^{c2} & \dots & w_1^{cK} \\
w_2^{c1} & w_2^{c2} & \dots & w_2^{cK} \\
\dots & \dots & \dots & \dots \\
w_n^{c1} & w_n^{c2} & \dots & w_n^{cK}
\end{bmatrix}
\end{array}
+
\begin{bmatrix}
\boxed{b^{c1}} & b^{c2} & \dots & b^{cK}
\end{bmatrix}
$$

$$
=
\begin{bmatrix}
\mathbf{w}^{c1}\mathbf{x}^{(1)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(1)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(1)} + b^{cK} \\
\boxed{\mathbf{w}^{c1}\mathbf{x}^{(2)} + b^{c1}} & \mathbf{w}^{c2}\mathbf{x}^{(2)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(2)} + b^{cK} \\
\dots & \dots & \dots & \dots \\
\mathbf{w}^{c1}\mathbf{x}^{(m)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(m)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(m)} + b^{cK}
\end{bmatrix}
$$

CONSTELLATE

Calculate all observations at one time: Matrix multiplication

$$\begin{array}{c}\mathbf{x}^{(1)}\\ \mathbf{x}^{(2)}\\ \\ \mathbf{x}^{(m)}\end{array}\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ \boxed{x_1^{(2)} \quad x_2^{(2)} \quad \dots \quad x_n^{(2)}} & & & \\ \dots & \dots & \dots & \dots \\ x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix} \bullet \begin{array}{c}\begin{array}{cccc}\mathbf{w}^{c}_{1} & \mathbf{w}^{c}_{2} & & \mathbf{w}^{cK}\end{array}\\ \begin{bmatrix} w_1^{c1} & w_1^{c2} & \dots & w_1^{cK} \\ w_2^{c1} & w_2^{c2} & \dots & w_2^{cK} \\ \dots & \dots & \dots & \dots \\ w_n^{c1} & w_n^{c2} & \dots & w_n^{cK} \end{bmatrix}\end{array} + \begin{bmatrix} b^{c1} & b^{c2} & \dots & \boxed{b^{cK}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{w}^{c1}\mathbf{x}^{(1)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(1)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(1)} + b^{cK} \\ \mathbf{w}^{c1}\mathbf{x}^{(2)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(2)} + b^{c2} & \dots & \boxed{\mathbf{w}^{cK}\mathbf{x}^{(2)} + b^{cK}} \\ \dots & \dots & \dots & \dots \\ \mathbf{w}^{c1}\mathbf{x}^{(m)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(m)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(m)} + b^{cK} \end{bmatrix}$$

Calculate all observations at one time: Matrix multiplication

$$\begin{array}{c|cccc} & & & & \\ \mathbf{x}^{(1)} & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ \mathbf{x}^{(2)} & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ & \dots & \dots & \dots & \dots \\ \mathbf{x}^{(m)} & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{array} \bullet \begin{array}{cccc} \mathbf{w}^{c}_{1} & \mathbf{w}^{c}_{2} & & \mathbf{w}^{cK} \\ \hline w_1^{c1} & w_1^{c2} & \dots & w_1^{cK} \\ w_2^{c1} & w_2^{c2} & \dots & w_2^{cK} \\ \dots & \dots & \dots & \dots \\ w_n^{c1} & w_n^{c2} & \dots & w_n^{cK} \end{array} + \begin{bmatrix} \boxed{b^{c1}} & b^{c2} & \dots & b^{cK} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{w}^{c1}\mathbf{x}^{(1)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(1)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(1)} + b^{cK} \\ \mathbf{w}^{c1}\mathbf{x}^{(2)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(2)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(2)} + b^{cK} \\ \dots & \dots & \dots & \dots \\ \boxed{\mathbf{w}^{c1}\mathbf{x}^{(m)} + b^{c1}} & \mathbf{w}^{c2}\mathbf{x}^{(m)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(m)} + b^{cK} \end{bmatrix}$$

Calculate all observations at one time: Matrix multiplication

$$\begin{array}{c} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \\ \mathbf{x}^{(m)} \end{array} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ \boxed{x_1^{(m)}} & \boxed{x_2^{(m)}} & \boxed{\dots} & \boxed{x_n^{(m)}} \end{bmatrix} \bullet \begin{array}{cccc} \mathbf{w}^{c}_{1} & \mathbf{w}^{c}_{2} & & \mathbf{w}^{cK} \\ \multicolumn{4}{c}{\begin{bmatrix} w_1^{c1} & w_1^{c2} & \dots & \boxed{w_1^{cK}} \\ w_2^{c1} & w_2^{c2} & \dots & \boxed{w_2^{cK}} \\ \dots & \dots & \dots & \boxed{\dots} \\ w_n^{c1} & w_n^{c2} & \dots & \boxed{w_n^{cK}} \end{bmatrix}} \end{array} + \begin{bmatrix} b^{c1} & b^{c2} & \dots & \boxed{b^{cK}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{w}^{c1}\mathbf{x}^{(1)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(1)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(1)} + b^{cK} \\ \mathbf{w}^{c1}\mathbf{x}^{(2)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(2)} + b^{c2} & \dots & \mathbf{w}^{cK}\mathbf{x}^{(2)} + b^{cK} \\ \dots & \dots & \dots & \dots \\ \mathbf{w}^{c1}\mathbf{x}^{(m)} + b^{c1} & \mathbf{w}^{c2}\mathbf{x}^{(m)} + b^{c2} & \dots & \boxed{\mathbf{w}^{cK}\mathbf{x}^{(m)} + b^{cK}} \end{bmatrix}$$

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

Vaswani, Ashish et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.