

# Multi-Head Attention

$$X = \begin{bmatrix} [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \end{bmatrix}$$

Sequence of 4 items where each item is represented as a vector with 1024 dimensions.  
Suppose number of heads  $h=8$

Our goal with MHA is to transform the initial sequence of uncontextualized embeddings into a sequence of contextualized embeddings.

## VISION TRANSFORMER

$$X = \begin{bmatrix} [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \end{bmatrix} \begin{matrix} \text{PATCH 1} \\ \text{PATCH 2} \\ \text{PATCH 3} \\ \text{PATCH 4} \end{matrix} \Rightarrow X = \begin{bmatrix} [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \end{bmatrix} \begin{matrix} \text{PATCH 1, 2, 3, 4} \\ \text{PATCH 1, 2, 3, 4} \\ \text{PATCH 1, 2, 3, 4} \\ \text{PATCH 1, 2, 3, 4} \end{matrix}$$

## LANGUAGE MODEL

$$X = \begin{bmatrix} [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \end{bmatrix} \begin{matrix} I \\ \text{LOVE} \\ \text{PEPPERONI} \\ \text{PIZZA} \end{matrix} \Rightarrow X = \begin{bmatrix} [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \end{bmatrix} \begin{matrix} I \\ I \text{ LOVE} \\ I \text{ LOVE PEPPERONI} \\ I \text{ LOVE PEPPERONI PIZZA} \end{matrix}$$

# STEP 1: from $x$ to $Q, K, V$

$$Q = X \times W_q = (4, 1024) \times (1024, 8, 128) = (4, 8, 128)$$

$$K = X \times W_k = (4, 1024) \times (1024, 8, 128) = (4, 8, 128)$$

$$V = X \times W_v = (4, 1024) \times (1024, 8, 128) = (4, 8, 128)$$

SEQUENCE

HIDDEN-SIZE

SEQUENCE N-HEAD

HEAD-DIM =  $1024 / N\text{-HEAD}$

$(4, 1024)$

$(1024, 8, 128)$

$(4, 8, 128)$

$$X = \begin{bmatrix} [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \\ [1 \dots 1024] \end{bmatrix} \times$$

INPUT SEQUENCE

$$\begin{bmatrix} \text{HEAD 1} & \text{HEAD 2} & \text{HEAD 3} \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ \vdots & \vdots & \vdots \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \end{bmatrix}$$

PARAMETERS  
 $W_q / W_k / W_v$

$$= \begin{bmatrix} \text{HEAD 1} & \text{HEAD 2} & \text{HEAD 3} \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [96 \dots 1024] & [96 \dots 1024] \end{bmatrix}$$

$Q / K / V$

# STEP 2: TREAT EACH HEAD INDEPENDENTLY!

$$\begin{array}{l} Q : (4, 8, 128) \\ K : (4, 8, 128) \\ V : (4, 8, 128) \end{array} \xrightarrow{\text{TRANSPOSE}} \begin{array}{l} (8, 4, 128) \\ (8, 4, 128) \\ (\underline{8}, \underline{4}, 128) \end{array}$$

each head...

... will compute the attention scores independently from other heads by using a part of the entire embedding.

$$\begin{array}{c} (4, 8, 128) \\ \left[ \begin{array}{c|c|c} \text{HEAD 1} & \text{HEAD 2} & \text{HEAD 3} \\ \hline [1 \dots 128], [129 \dots 256], \dots & [129 \dots 256] & [257 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [129 \dots 256] & [257 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [129 \dots 256] & [257 \dots 1024] \\ [1 \dots 128], [129 \dots 256], \dots & [129 \dots 256] & [257 \dots 1024] \end{array} \right] \end{array} \xrightarrow{\text{TRANSPOSE}} \begin{array}{c} \text{HEAD 1} \\ \text{HEAD 2} \\ \text{HEAD 7} \\ \text{HEAD 8} \end{array}$$

Q/K/V

$$\begin{array}{c} (8, 4, 128) \\ \left[ \begin{array}{c|c|c} \text{HEAD 1} & \text{HEAD 2} & \text{HEAD 7} & \text{HEAD 8} \\ \hline \begin{bmatrix} [1 \dots 128] \\ [129 \dots 256] \\ [257 \dots 1024] \end{bmatrix} & \dots & \dots & \begin{bmatrix} [996 \dots 1024] \\ [996 \dots 1024] \\ [996 \dots 1024] \end{bmatrix} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{array} \right] \end{array}$$

## WHY?

- 1) We want to parallelize the computation
- 2) Each head should learn to relate tokens (or patches) differently

# STEP 3: CALCULATE THE ATTENTION FOR EACH HEAD IN PARALLEL

$Q_{\text{HEAD}_1} = \begin{bmatrix} [1 & \dots & 128] \\ [1 & \dots & 128] \\ [1 & \dots & 128] \\ [1 & \dots & 128] \end{bmatrix}$ 

 $\rightarrow$  dimension 1...128 of token 1  
 $\rightarrow$  dimension 1...128 of token 2  
 $\rightarrow$  dimension 1...128 of token 4

$K_{\text{HEAD}_1}^T = \begin{bmatrix} [1] & [1] & [1] & [1] \\ \vdots & \vdots & \vdots & \vdots \\ [128] & [128] & [128] & [128] \end{bmatrix}$ 

 $\downarrow$  dimension 1...128 of token 1  
 $\downarrow$  dimension 1...128 of token 4

$$\frac{Q \times K^T}{\sqrt{d_{\text{head}}}} = Q$$

	K				
	13.9	21.1	-100.3	17.5	1
	-5.0	3.14	1.2	75.3	LOVE
	...	...	...	...	PEPPERONI
	...	...	...	...	PIZZA
	1	LOVE	PEPPERONI	PIZZA	

$$\text{softmax} \left( \frac{Q \times K^T}{\sqrt{d_{\text{head}}}} \right) = Q$$

	K				
	0.1	0.2	0.5	0.3	1
	0.4	0.1	0.3	0.2	LOVE
	...	...	...	...	PEPPERONI
	...	...	...	...	PIZZA
	1	LOVE	PEPPERONI	PIZZA	

BRO, WHERE IS YOUR MASK?

$$\text{softmax} \left( \frac{Q \times k^T}{\sqrt{d_{\text{head}}}} + \text{MASK} \right) = Q$$

1.0	0	0	0	1
0.6	0.4	0	0	LOVE
0.2	0.4	0.4	0	PEPPERONI
0.4	0.2	0.3	0.1	PIZZA
	1	LOVE	PEPPERONI	PIZZA

STEP 4: MULTIPLY BY THE V SEQUENCE

1.0	0	0	0	1
0.6	0.4	0	0	LOVE
0.2	0.4	0.4	0	PEPPERONI
0.4	0.2	0.3	0.1	PIZZA
	1	LOVE	PEPPERONI	PIZZA

 $\times$ 

[1 ... 128]	1
[1 ... 128]	LOVE
[1 ... 128]	PEPPERONI
[1 ... 128]	PIZZA

(4, 4)                      (4, 128)

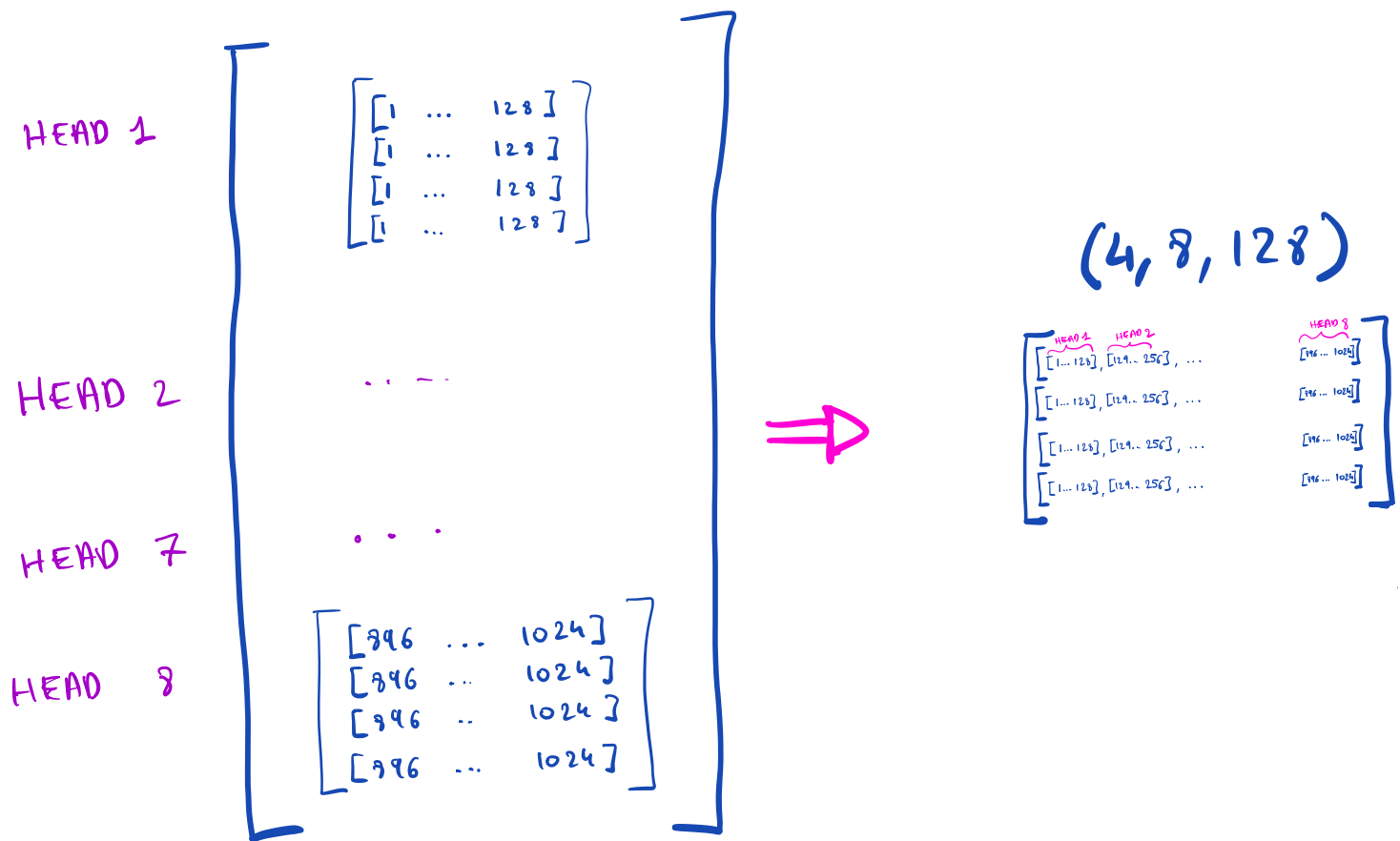
EACH ROW REPRESENTS A WEIGHTED SUM OF:

$$= \begin{bmatrix} [1 \dots 128] \\ [1 \dots 128] \\ [1 \dots 128] \\ [1 \dots 128] \end{bmatrix} \begin{matrix} \rightarrow 1 \\ \rightarrow 1 \text{ LOVE} \\ \rightarrow 1 \text{ LOVE PEPPERONI} \\ \rightarrow 1 \text{ LOVE PEPPERONI PIZZA} \end{matrix}$$

(4, 128)

# STEP 5: TRANSPOSE BACK

(8, 4, 128)

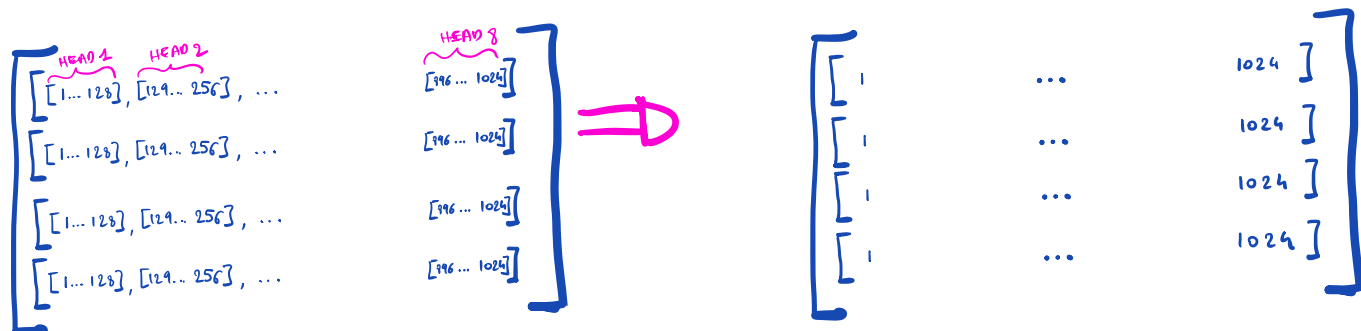


## STEP 6: CONCATENATE ALL THE HEADS

Given that each head is computing the contextualized embeddings using a "part" of each token we can concatenate all the result of all the heads back together

(4, 8, 128)

(4, 1024)



# STEP 7: MULTIPLY BY $W_0$

PARAMETER

$W_0$

$$\begin{array}{c}
 (4, 1024) \\
 \begin{bmatrix} \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \end{bmatrix}
 \end{array}
 \begin{array}{c}
 \text{I} \\
 \text{I LOVE} \\
 \text{I LOVE PEPPERONI} \\
 \text{I LOVE PEPPERONI PIZZA}
 \end{array}
 \times
 \begin{array}{c}
 (1024, 1024) \\
 \begin{bmatrix} \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 (4, 1024) \\
 \begin{bmatrix} \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \\ \vdots & \dots & 1024 \end{bmatrix}
 \end{array}
 \begin{array}{c}
 \text{I} \\
 \text{I LOVE} \\
 \text{I LOVE PEPPERONI} \\
 \text{I LOVE PEPPERONI PIZZA}
 \end{array}$$

## WHY MULTIPLY BY $W_0$ ?

If we don't multiply by  $W_0$ , each group of 128 dimensions will be independent from each other, as they are the result of the concatenation of independent heads. Multiplying by  $W_0$  gives the chance to each head to "mix" with other heads.