Attention (P-7) (LLM notes)

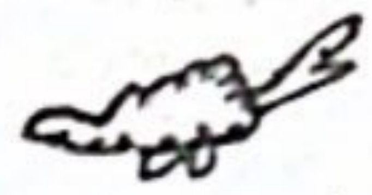
- Now that we know about Wv and value vectors lets go back to attention pattern. we dont care about K, Q anymore as we are done with those

- Now you take the value matrix and multiply each embedding by it to make a set of value vectors, you can think of these value vectors as being associated with their corresponding Keys for each column in this Diagram you multiply each of the value vectors by the corresponding weight in that column then the values for fluffy and blue would be some number the rest would be almost 0 then you sum up the whole column and add it to the original embedding this updates the embedding "giving it new meaning and context" and ofcourse you do this for all columns and update all embeddings "words" $E_1 \to E_8$ would once adding with $\Delta E_1 \to \Delta E_8 = E_1' \to E_8'$. This whole process is <u>One head of Attention</u>

Value matrix W_v	a ⇓ E_1 . . .	creature ↓ E_4	. . . forest ⇓ E_8
$a \to E_1 \xrightarrow{W_v} V_1$	. . .	$0 \cdot V_1$ +	. . .
$\text{fluffy} \to E_2 \xrightarrow{W_v} V_2$	. . .	$0.42 \cdot V_2$ +	. . .
$\text{blue} \to E_3 \xrightarrow{W_v} V_3$	. . .	$0.58 \cdot V_3$ +	. . .
$\text{creature} \to E_4 \xrightarrow{W} V_4$	. . .	$0 \cdot V_4$ ⋮	do for all cols
⋮ V_8		$0 \cdot V_8$ =	

ΔE_4

creature embedding → $E_4 + \Delta E_4 = E_4'$ → new creature embedding ⇓ "fluffy creature" ...

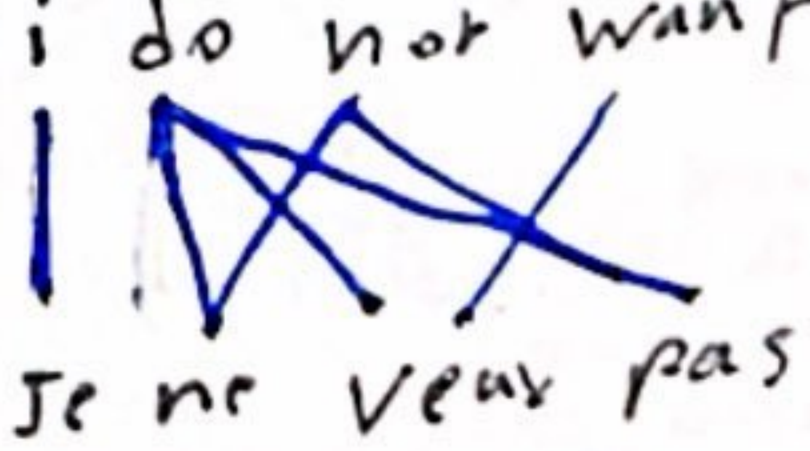


NOTE while we did say W_v is $12{,}288^2$ Dim it better if: # Values in W_v = # Queries in W_Q + # keys in W_K so to achive this W_v would be a product of 2 smaller matricies. the first matrix maps from 12k Dim to 128 Dim the second one maps it back upto 12K Dim this in linear algebra terms is called a <u>Low Rank transformation</u>

- So far we talked about Self Attention: the blue creature...

- cross Attention: is another variation used in other models, cross attention involves models that process two Different types of Data Like translation, text in one lang and text in another lang or audio transcription. its almost the same as SA but the difference is that in key Querie maps the both use Different Data sets ie key might come from one language and the Queries from another

Ex • which words in one lang correspond to which words in another ↖ tells us # **No masking!**


i do not want
Je ne veux pas