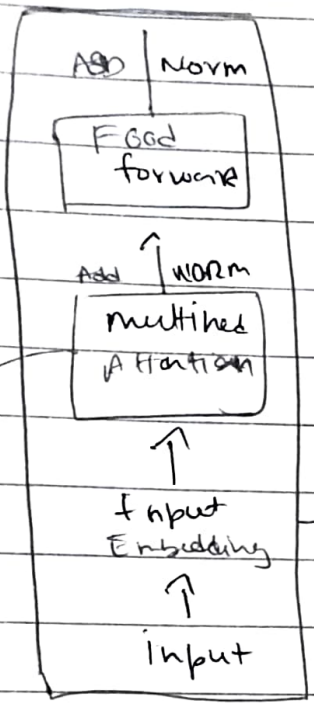
Transformers (LSTMs are DEAD)

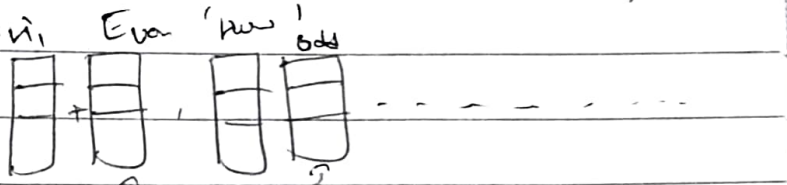
Encoders



Hi How are you

① Input Embedding

② Positional Encoding Odd/Even

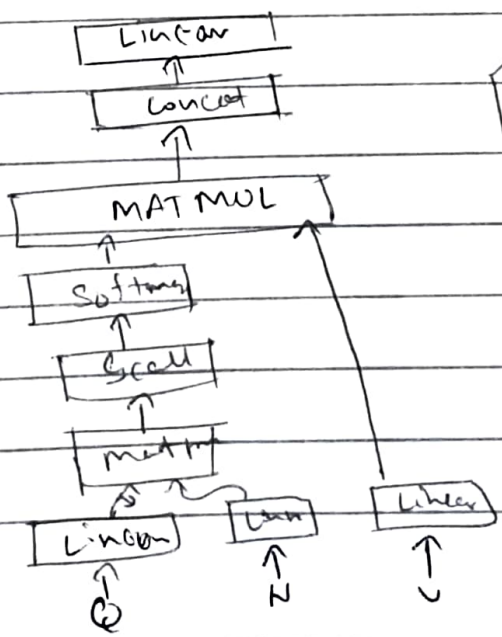


$$Even = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$Odd = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

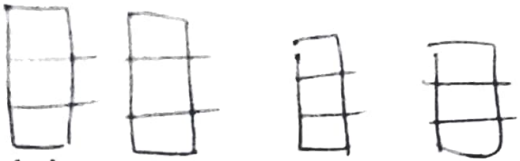
Multihead Attention

-> Self Attention

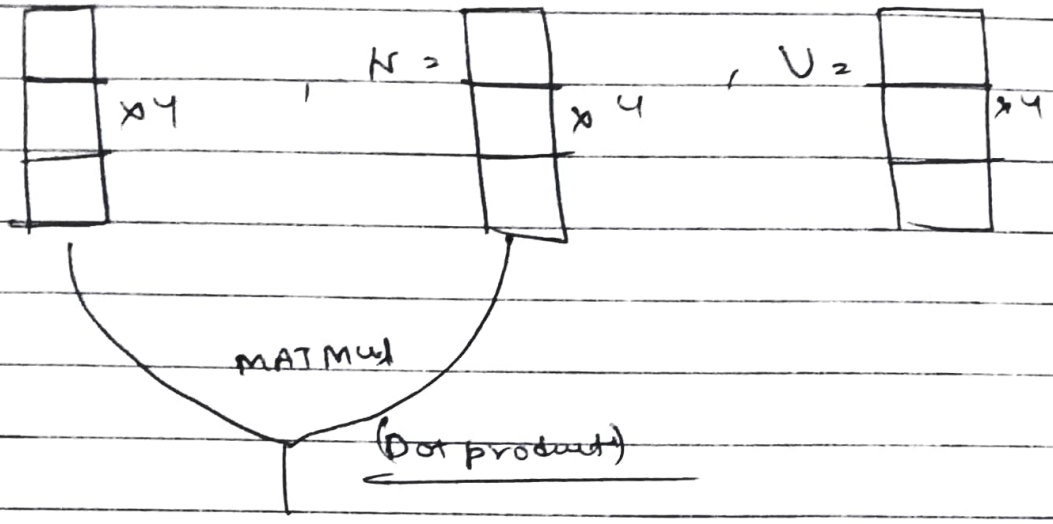


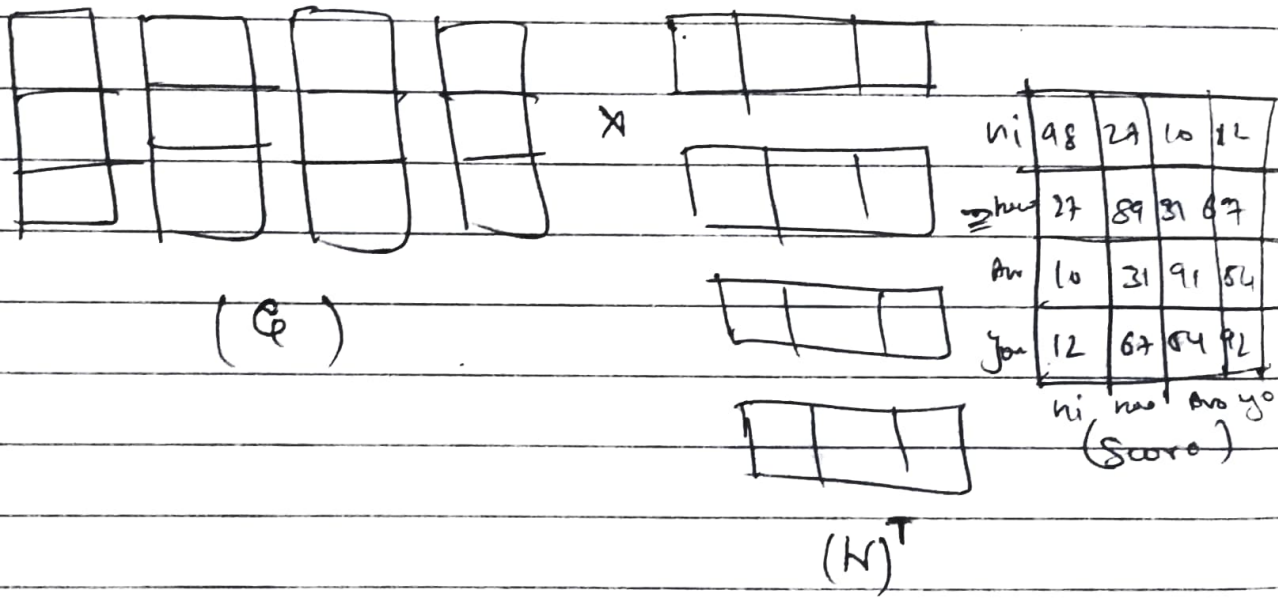
Self Att. helps model to associate each word with every word in the vocabulary Input.

Q, K, V → Youtube example



Page :
Date :

Q = 



=) Basically Score matrix is a kind of confusion matrix which tells the correlation b/w each input vector or word .

higher the score in score matrix higher is the focus

Page
Date

=) Now scores are scaled by dividing them with $\sqrt{\text{dimension of key}}$

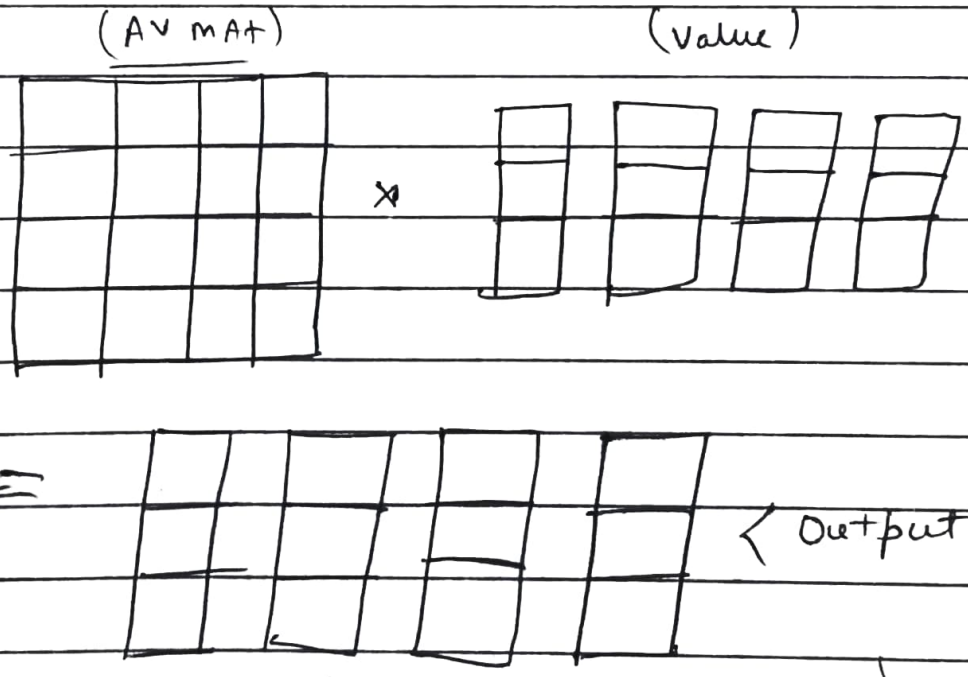
=) Apply softmax to scaled scores.

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

By softmax high scores are relatively probabilized higher and low scores are diminished.

=) now <u>score matrix</u> becomes <u>Attention value matrix</u> <u>(AV MAt)</u>

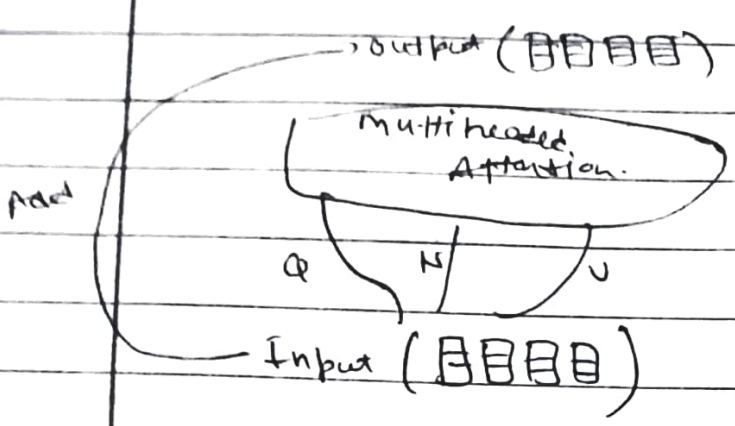
=) Multiply AV matrix with value.



{ This vector gives the metric of how each word should attend to others words. }

Page :
Date :

4) 4) Residual Connect / Layer Nor and Point Food forward.



Now Input and output are added and passed through norm layer to get values easy to be calculated.

Now it is passed in a linear layer with Relu activation and At last is normalised.

x —— x End of the Encoder layer —— x —— x

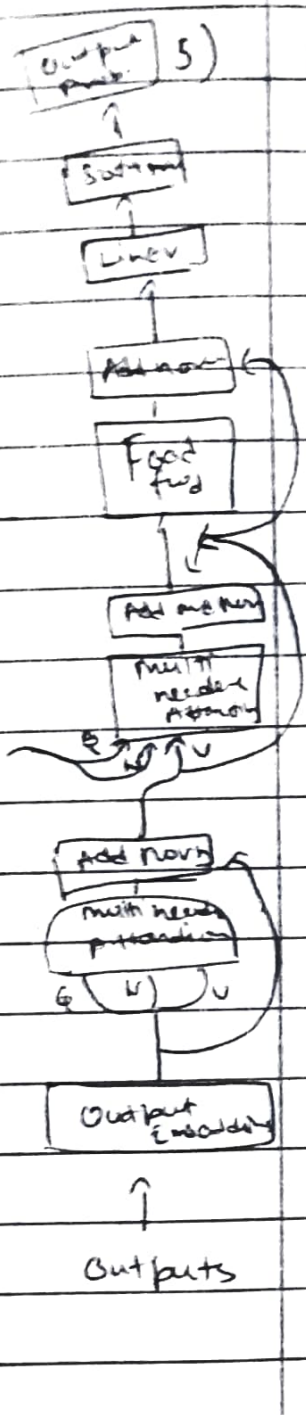
Note :- We can use Encoder much more times to further encode the information and thus it learns further more about the features and gets better.

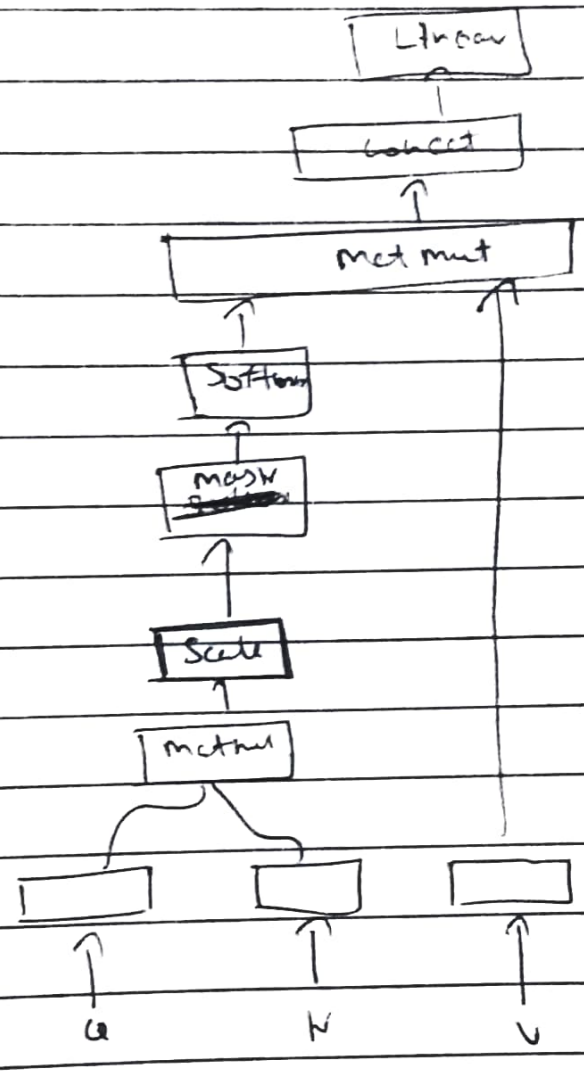
<DECODERS>

Page
Date

5) Output Embedding and Position Encoding

Multi-headed Attention in Decoders





NOTE:

The mask layer is here because Decoder is generating a sequence so it is putting the word at the end of the output what is the significance of the future coming value acc. to the matrix.

So every word has access to scores of itself and words before it.

Masking ex

Page :
Date :

0.1	0.6	0.3
0.3	0.7	0.7
0.4	0.4	0.2

+

0	-Inf	-Inf
0	0	-Inf
0	0	0

=

0.1	-Inf	-Inf
0.3	0.3	-Inf
0.4	0.4	0.2

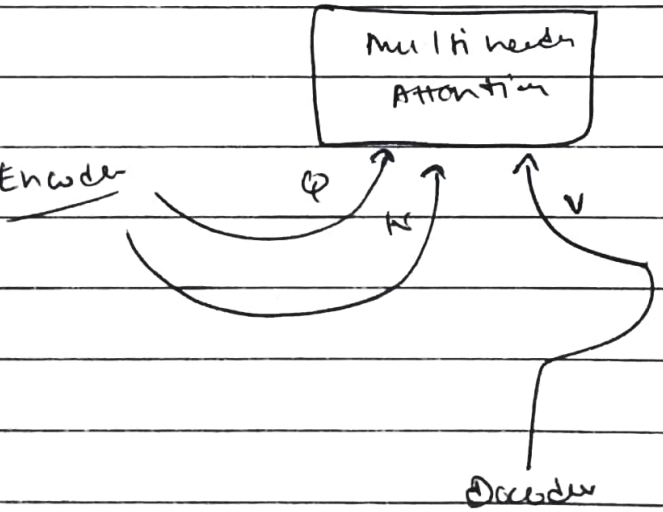
Softmax

Now taking the softmax of Masked Matrix will give us.

0.1	0	0
0.8	0.13	0
0.1	0.2	0.12

Remember it

6.) Second Multi-Headed Attention { check b/w encoder & its and decoder vcl.



there the Decoder and Encoder provided weights are matched and new output matrix is given and in Last Linear and Softmax function is used to get the out probabilities of every word.