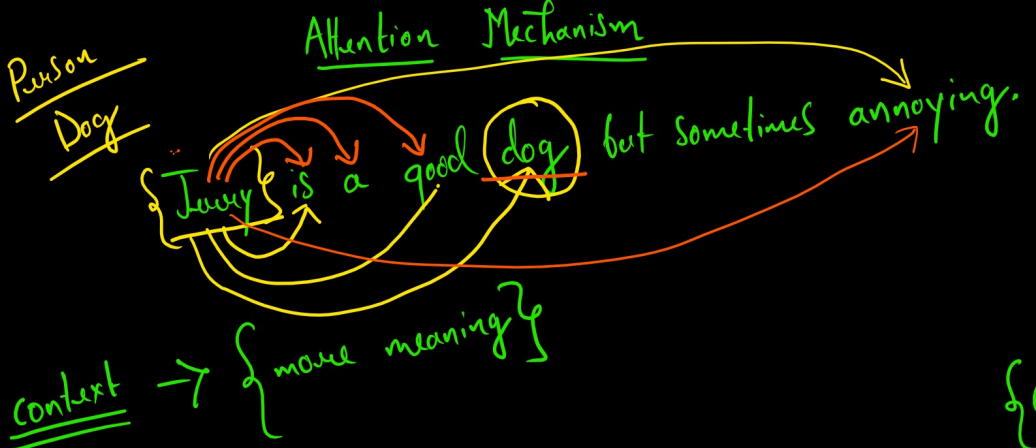
Today's Agenda

1) Attention
2) Self Attention
3) Multi Head Attention
4) Transformers

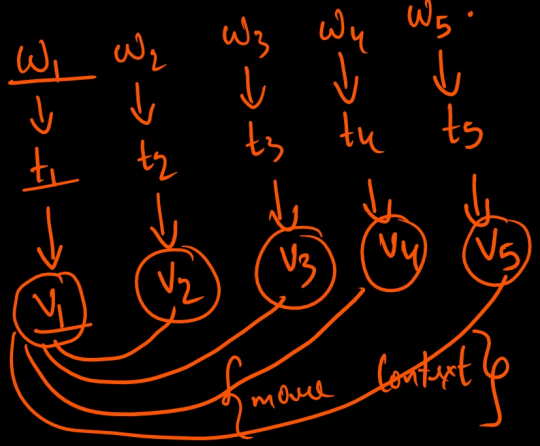


Word Embeddings
- TF-IDF
- Word2Vec

Nearest Neighbour Rule
{1 towards bi-directional}

{context}
↓
Glove
↳ Bidirectional Tagger
1→

{Attention}

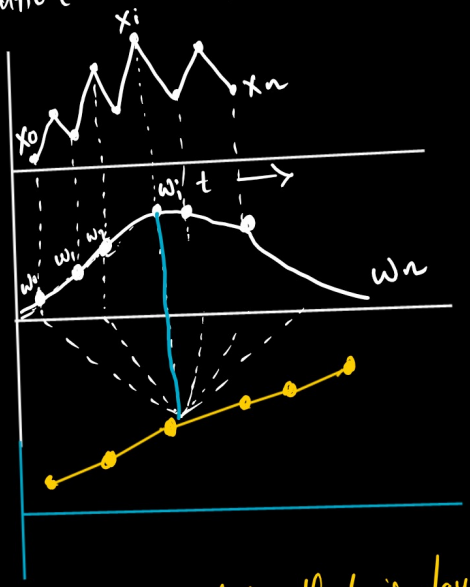


weights
↳ more complexity
↳ Back Prop

Level thinking of attention

weights ?

Attention in Time Series



Text

1) filter data that is far away
2) amplify data that is nearby

Noisy Data

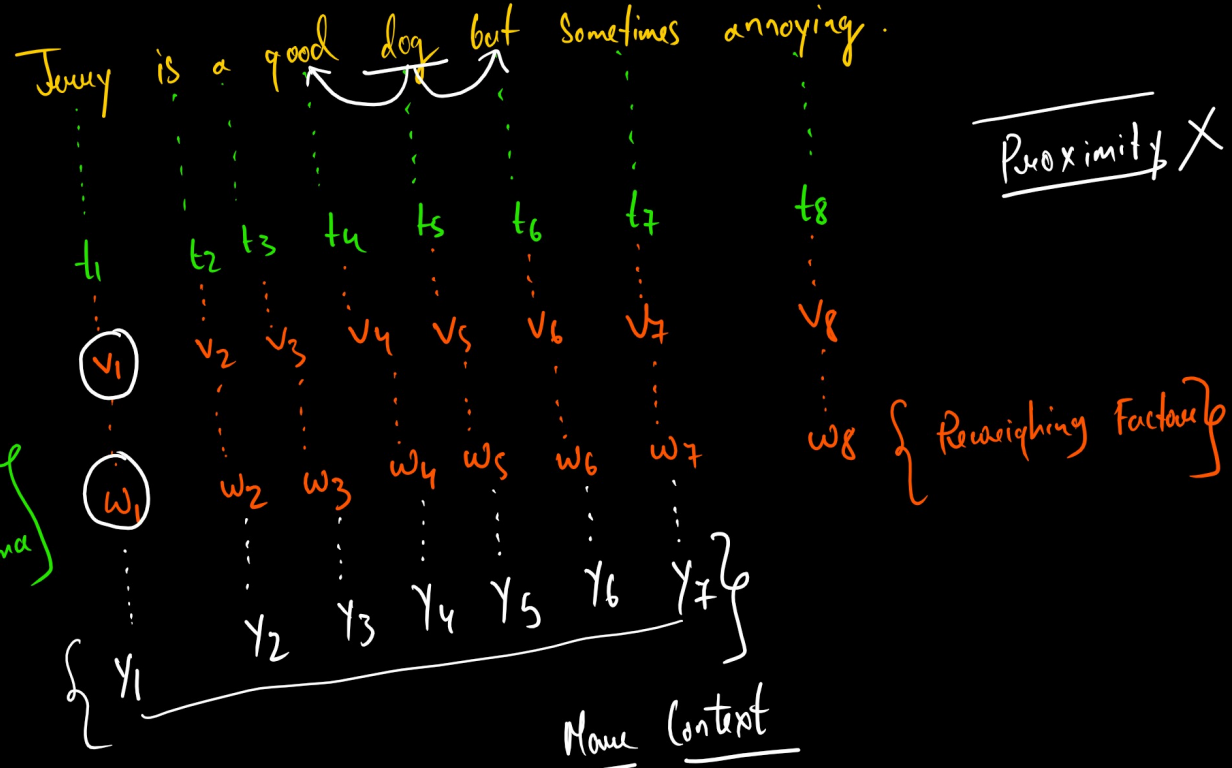
Standa
Normaliz

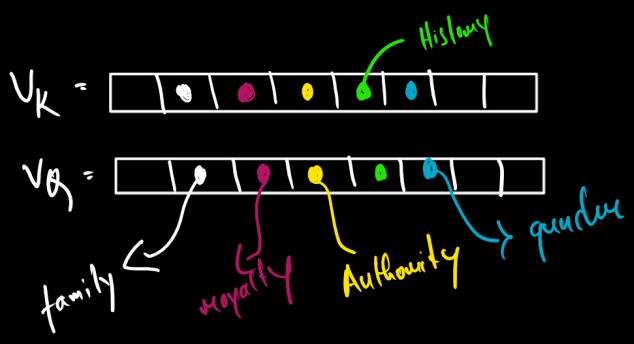
Training/Data Points = $[x_0, x_1, \ldots x_n]$

Reweigh Factor $W_i = [w_{0i} + w_{1i} + w_{2i}]$

Noisy → Normalization → {Reweighing}

Box Cox
log



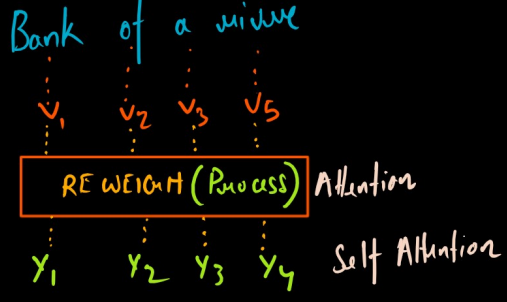


Do W. Embeddings

cluster distance should be less.

Reweighing = $\{ W_{KQ} = V_K \cdot V_Q \}$

Explore this idea



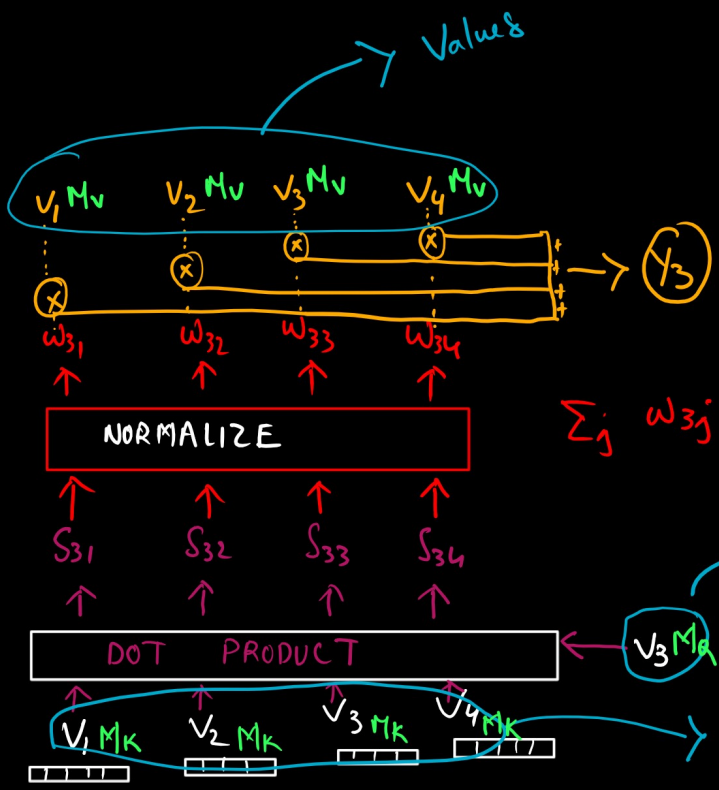
$$Y_1 = w_{11}V_1 + w_{12}V_2 + w_{13}V_3 + w_{14}V_4$$
$$Y_2 = w_{21}V_1 + w_{22}V_2 + w_{23}V_3 + w_{24}V_4$$
$$Y_3 = w_{31}V_1 + w_{32}V_2 + w_{33}V_3 + w_{34}V_4$$
$$Y_4 = w_{41}V_1 + w_{42}V_2 + w_{43}V_3 + w_{44}V_4$$

$V_1 V_1 = w_{11}$
$V_1 V_2 = w_{12}$
$V_1 V_3 = w_{13}$
$V_1 V_4 = w_{14}$
→ Norm → w_{11} w_{12} w_{13} w_{14}

$V_2 V_1 = w_{21}$
$V_2 V_2 = w_{22}$
$V_2 V_3 = w_{23}$
$V_2 V_4 = w_{24}$
→ Norm → w_{21} w_{22} w_{23} w_{24}

* **SELF ATTENTION**

1) No Trained weights
2) Order has no influence
3) Proximity no influence
4) Shape independent



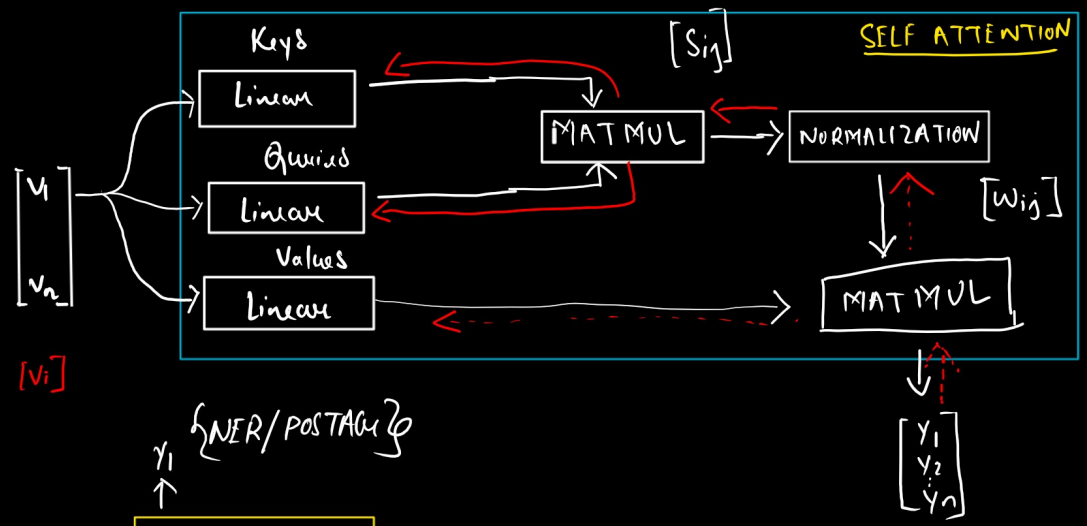
Weights

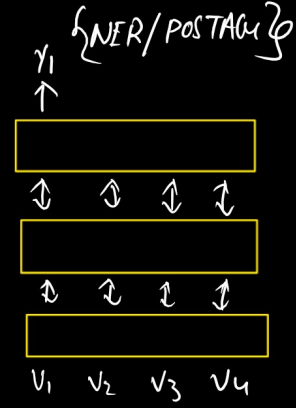
[Matrix] 3x3 3x5

$\sum_j w_{3j} = 1$

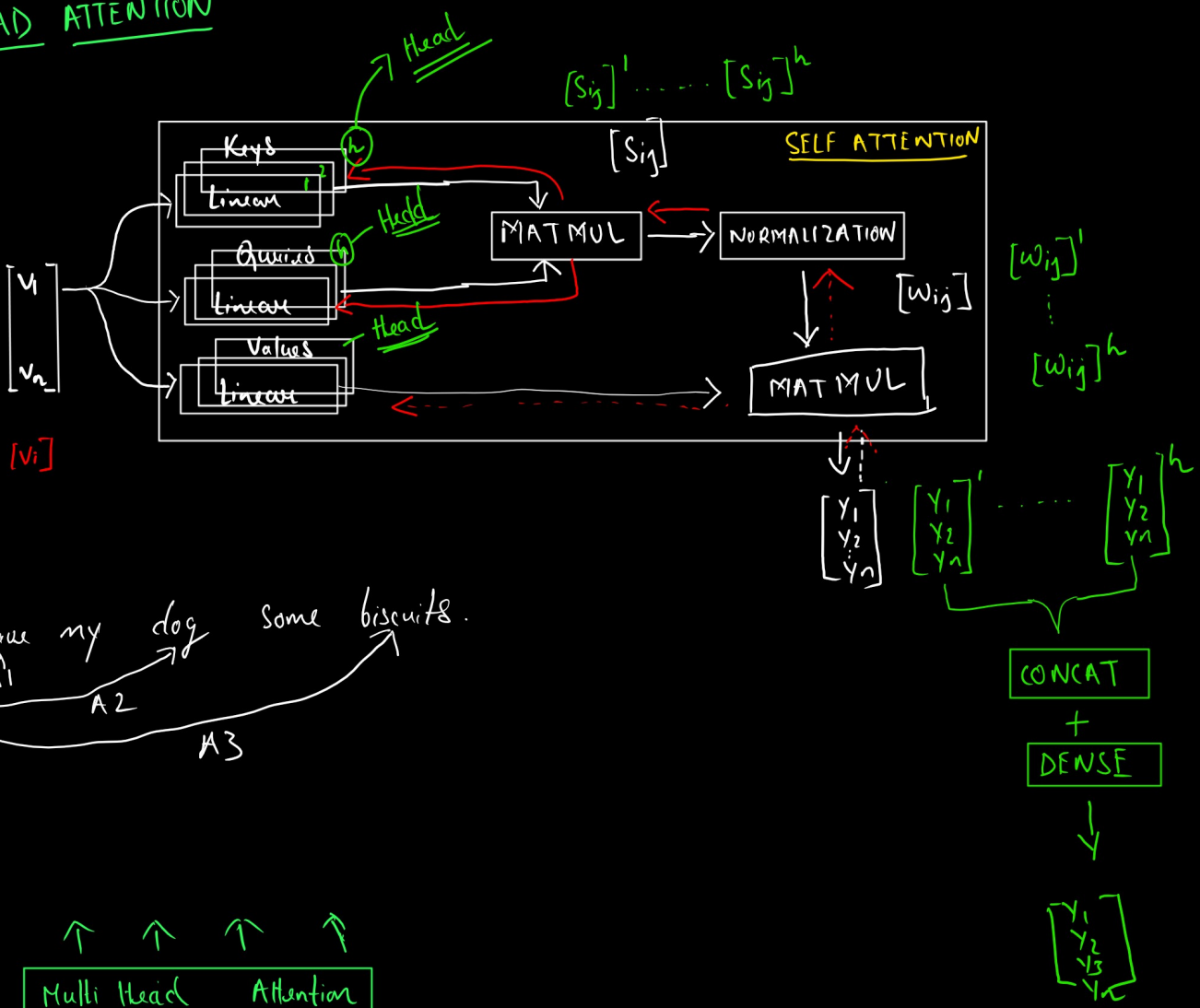
V_i M

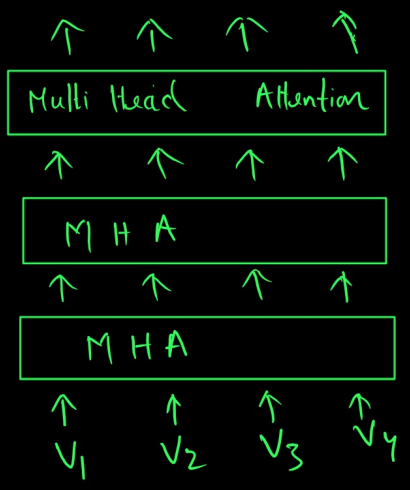
[1xK] [KxK] = [1xK]

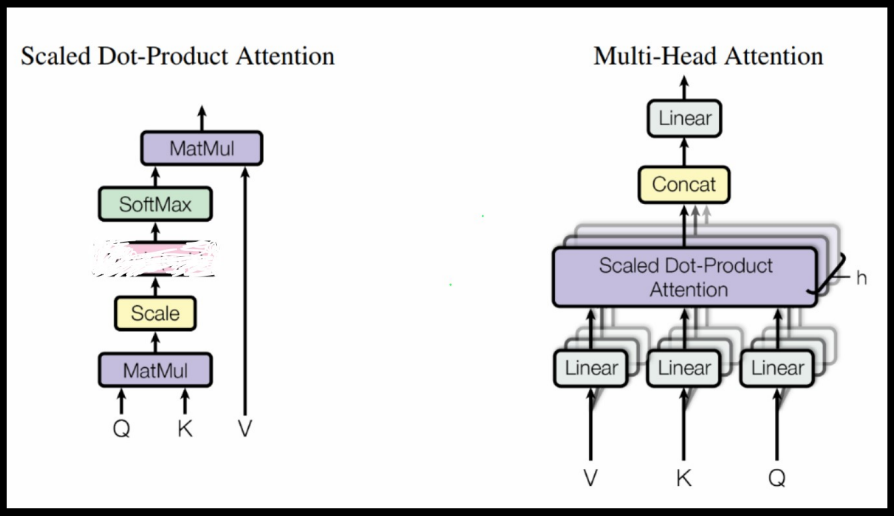




MULTIHEAD ATTENTION







Decoder
1) unidirection
2) Auto regressive

[[ ]

Bank of a River
[ ] 1x768

Auto regressive

CV
[ ]
[ ]

⊗ → hyper parameter

Encoder

Fusion of English + French

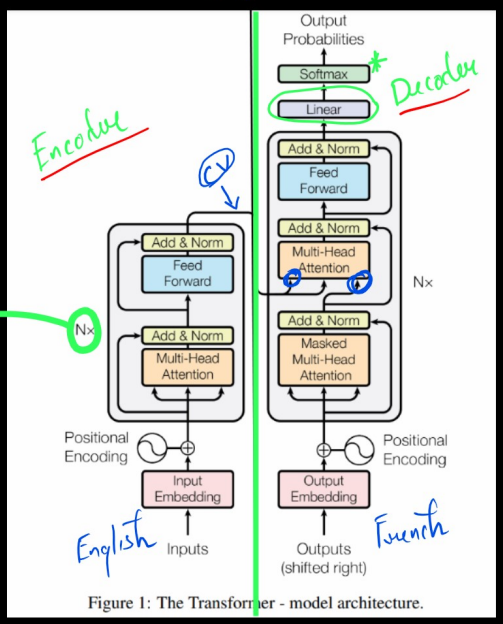


Figure 1: The Transformer - model architecture.

My name

[ ]

[ ]