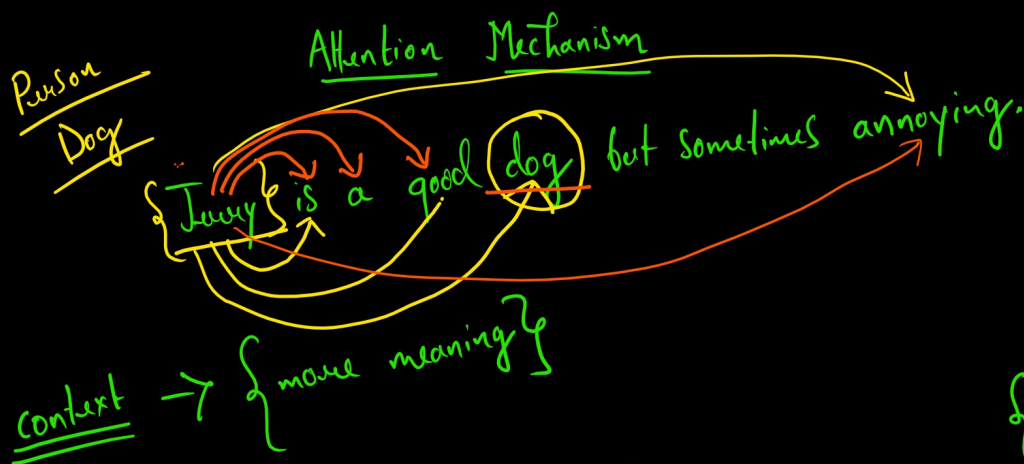


Today's Agenda

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



Word Embeddings

✓ TF-IDF
✓ Word2Vec

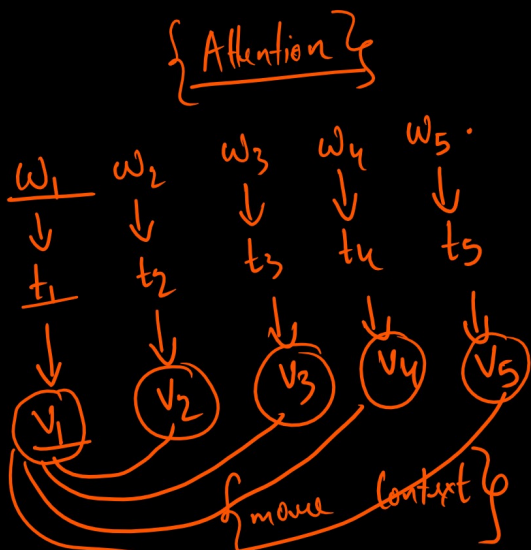
Nearest Neighbour Rule
{n} spans bi-directional

{context}

✓ Glove

↳ Bidirectional Tagger

{n} →

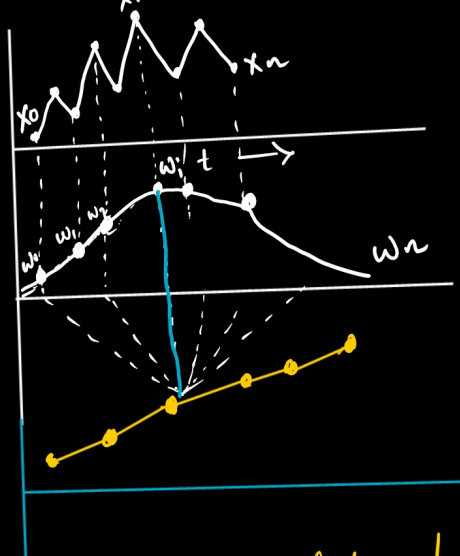


weights
more complexity
↳ Back Prop

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

Standard Normal

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

Reweight Factor $w_i = [w_{0i} + w_{1i} + w_{2i}]$

Noisy

Box Cox

Normalization

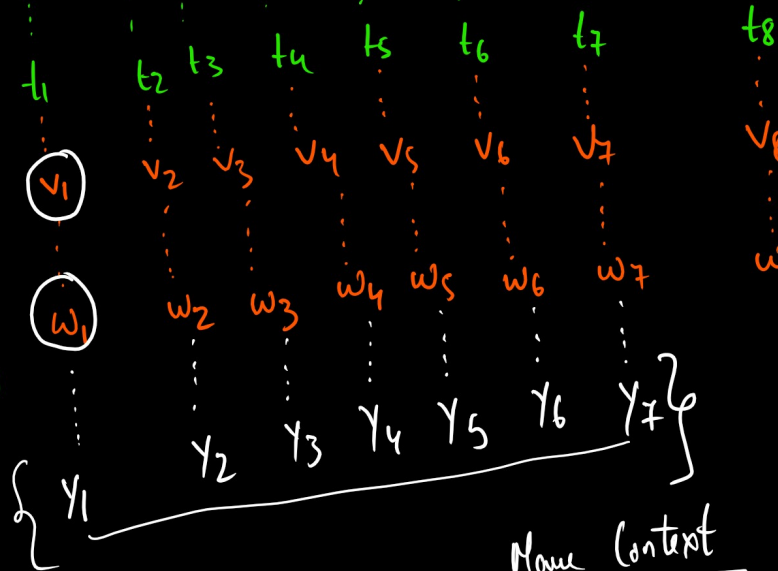
log

Rewighting

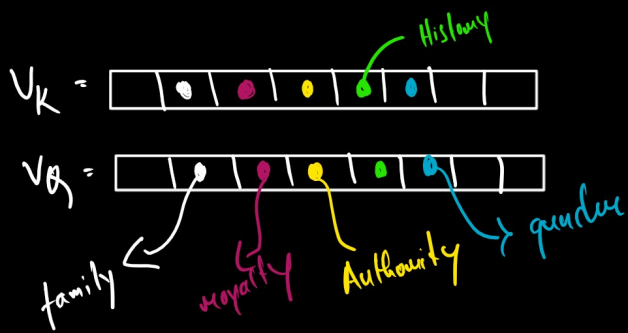
Jimmy is a good dog but sometimes annoying.

Proximity X

tokens
lower case
stopwords
stemming, lemmatization



Name Context



Son
 daughter
 land
 Authority
 own
 Kingdom
 Army

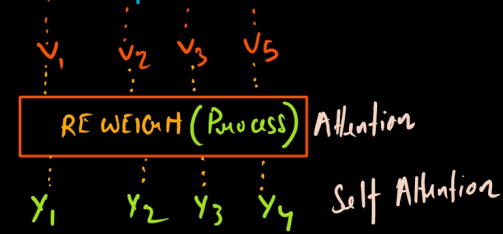
dog
 cat
 giraffe
 computer
 medicines

W Embeddings

cluster distance should be less.

Reweighting = { $W_{KQ} = V_K \cdot V_Q$ }
 Explore this idea

Bank of a mirror



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

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

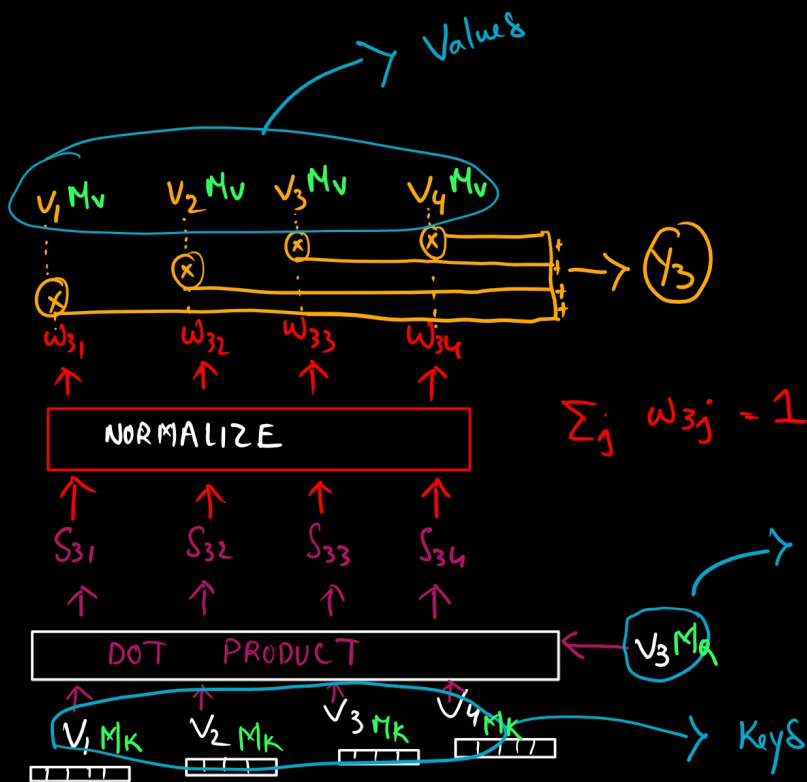
Norm
 \rightarrow
 w_{11}
 w_{12}
 w_{13}
 w_{14}

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

Norm
 \rightarrow
 w_{21}
 w_{22}
 w_{23}
 w_{24}

* SELF ATTENTION

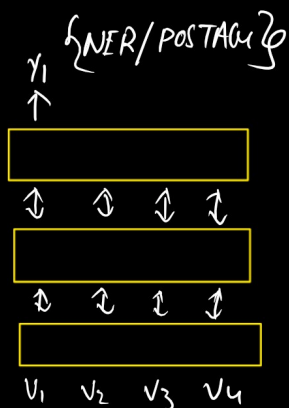
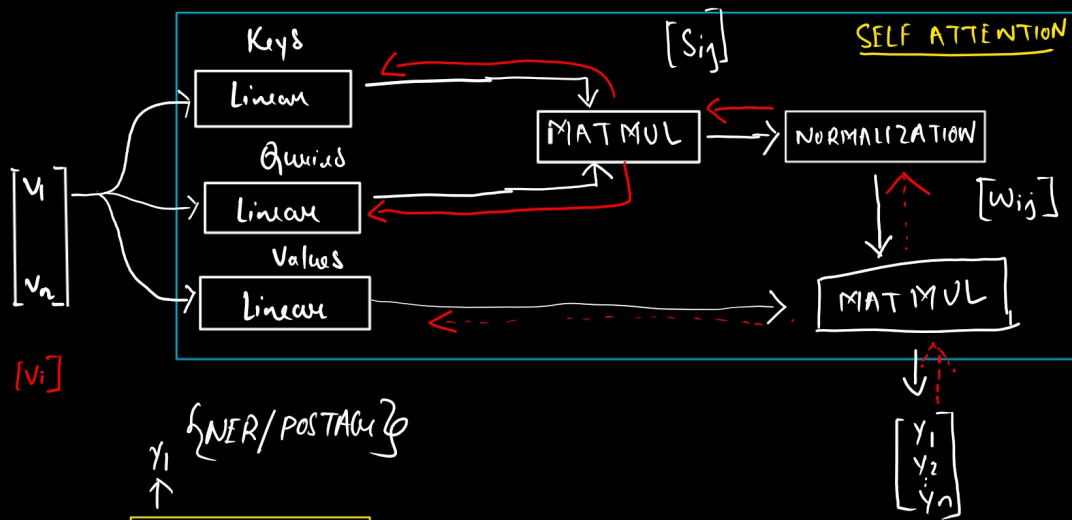
- 1) No Trained weights
- 2) Order has no influence
- 3) Proximity no influence
- 4) Shape independent



Weights

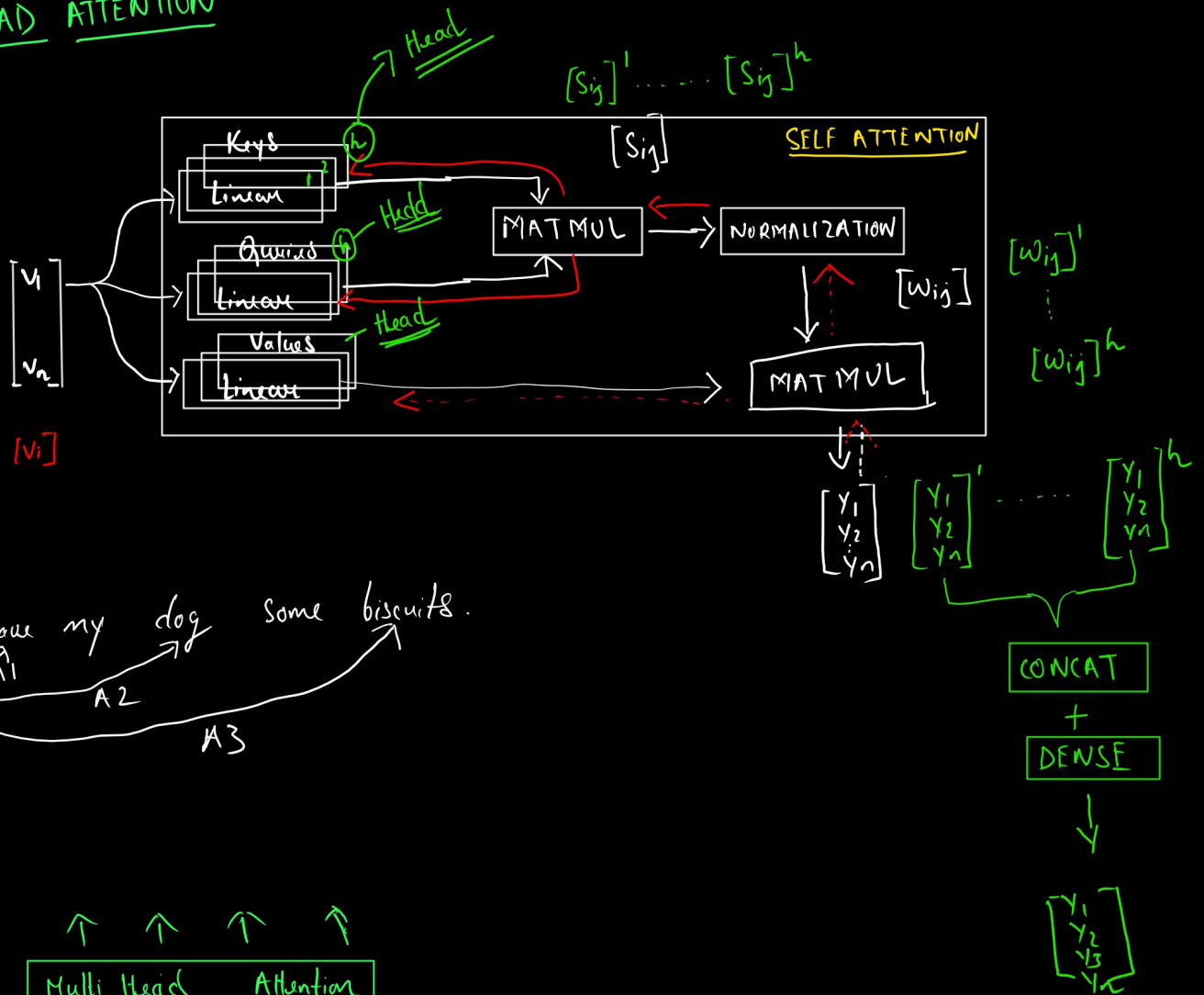
$$[Matrix] \begin{matrix} 3 \times 3 \\ 3 \times 5 \end{matrix}$$

$$V_i M \quad [1 \times K] \quad [K \times K] = [1 \times K]$$



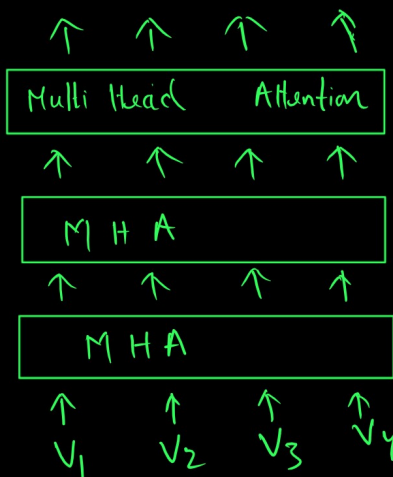
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

MULTIHEAD ATTENTION

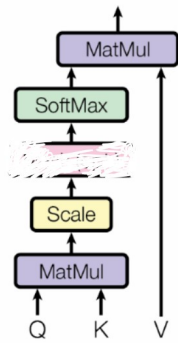


I gave my dog some biscuits.

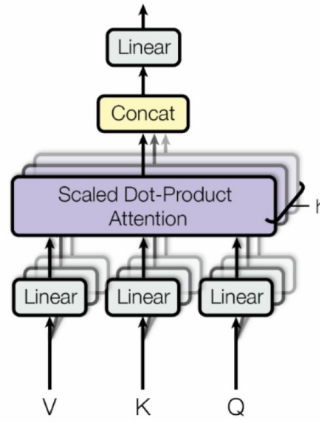
A1: I
A2: gave
A3: my dog some biscuits.



Scaled Dot-Product Attention



Multi-Head Attention



Decoder
1) unidirectional
2) Auto regressive

Bank of a River
[]
[]
1x768

Auto regressive

[]

(X) → hyper parameter

Encoder

Fusion of English + French

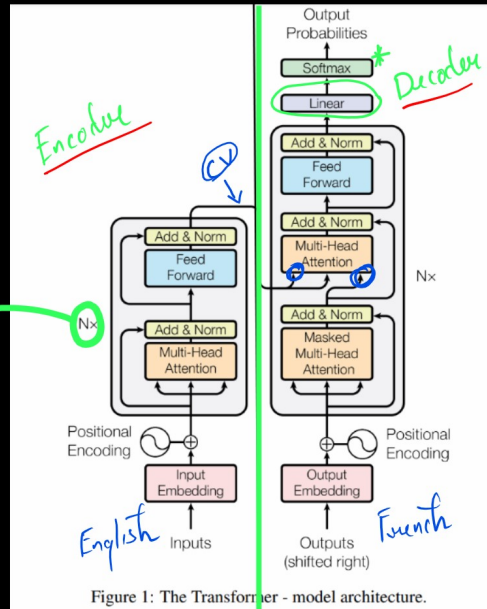


Figure 1: The Transformer - model architecture.

My name

[]
[]

