$\text{Var}(z) > \text{Var}(y) > \text{Var}(x)$

$\text{Var}(z) \simeq 3\,\text{Var}(x)$

If d dim then dm → $d\,\text{Var}(x)$

* Somehow $\begin{bmatrix} [ac+bd] \\ [ae+bf] \\ \vdots \end{bmatrix}$ aise number se ← ise divide karna hai ki every time Var(x) hi aaye

$$[a\ b] \quad \left.\begin{matrix} \dfrac{[c\ d]}{\sqrt{2}} \\ \dfrac{[e\ f]}{\sqrt{2}} \\ \dfrac{[g\ h]}{\sqrt{2}} \end{matrix}\right\} \; 2\,\text{Var}(x)$$

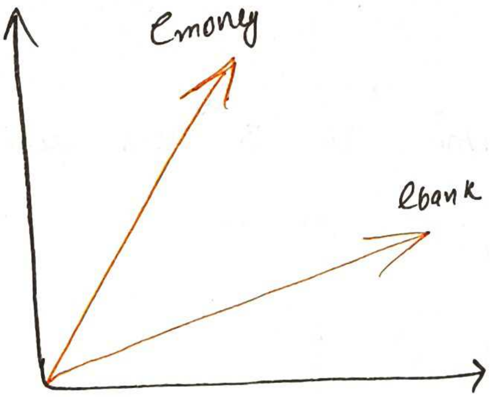
dim

↳ $\sqrt{d_k}$

$\frac{1}{2}\text{Var}(y) \Rightarrow \frac{1}{\cancel{2}}\cancel{2}\,\text{Var}(x) = \text{Var}(x)$

d_k → dimension

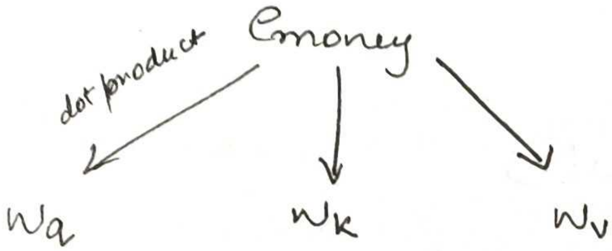
Self Attention Geometric Intuition



money bank

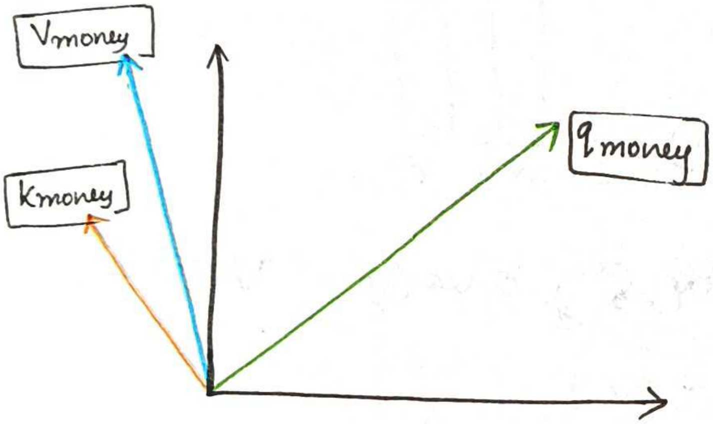
e_{money} → [2, 7]

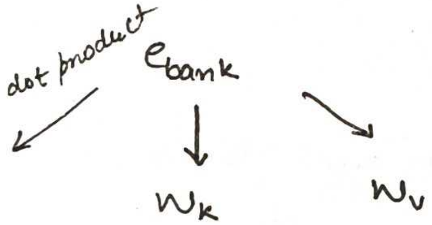
e_{bank} → [9, 3]



$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 3 & 4 \\ 5 & 1 \end{bmatrix} \quad \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}$$

2x2 2x2 2x2

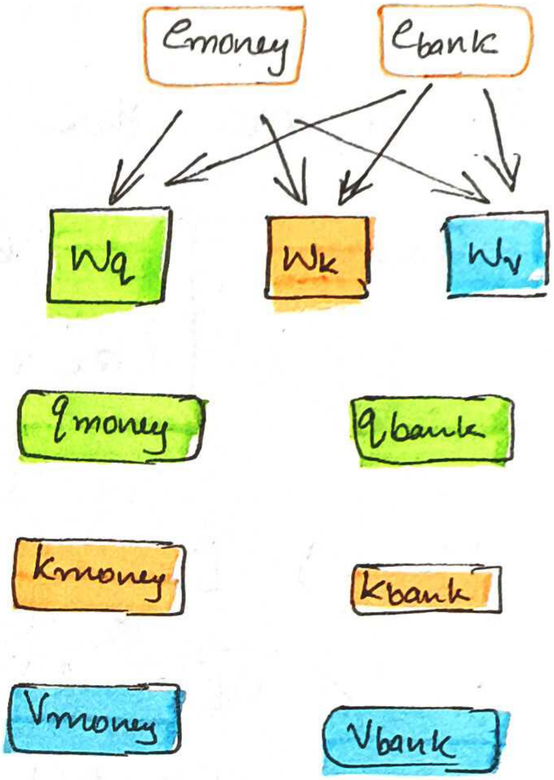


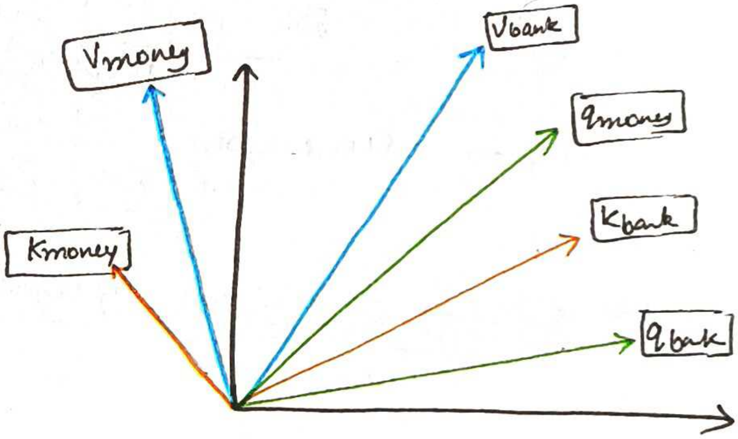


$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 3 & 4 \\ 5 & 1 \end{bmatrix} \quad \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}$$

* All values are hypothetical

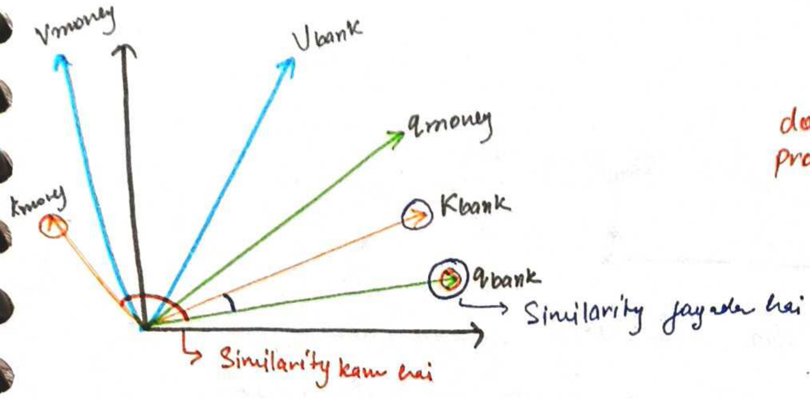
money bank

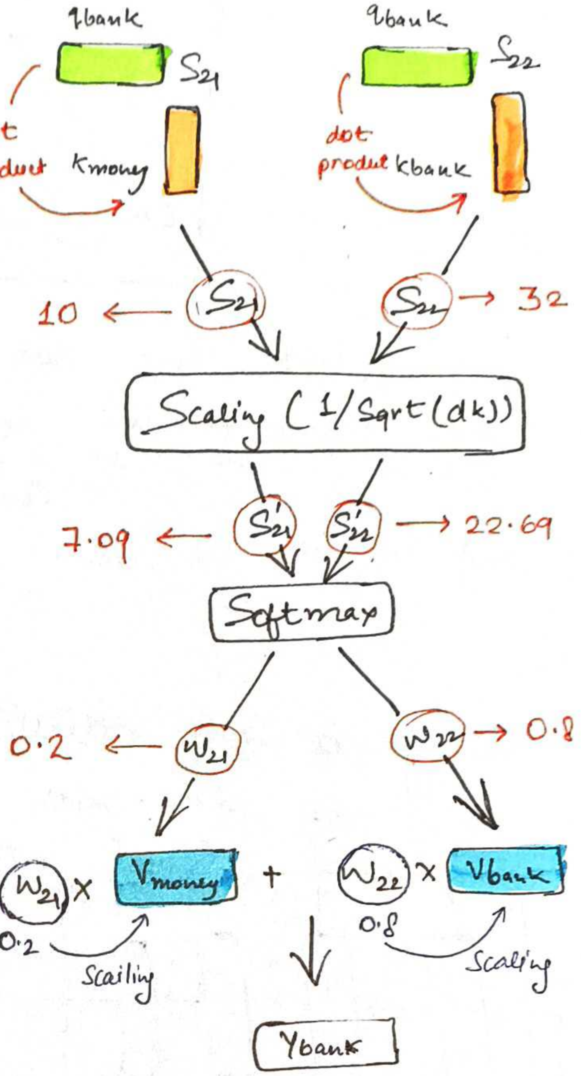




Here we make single vector to 3 new vector

from $e_{money} \xrightarrow{to}$ q_{money}
K_{money}
V_{money}





Dot Product

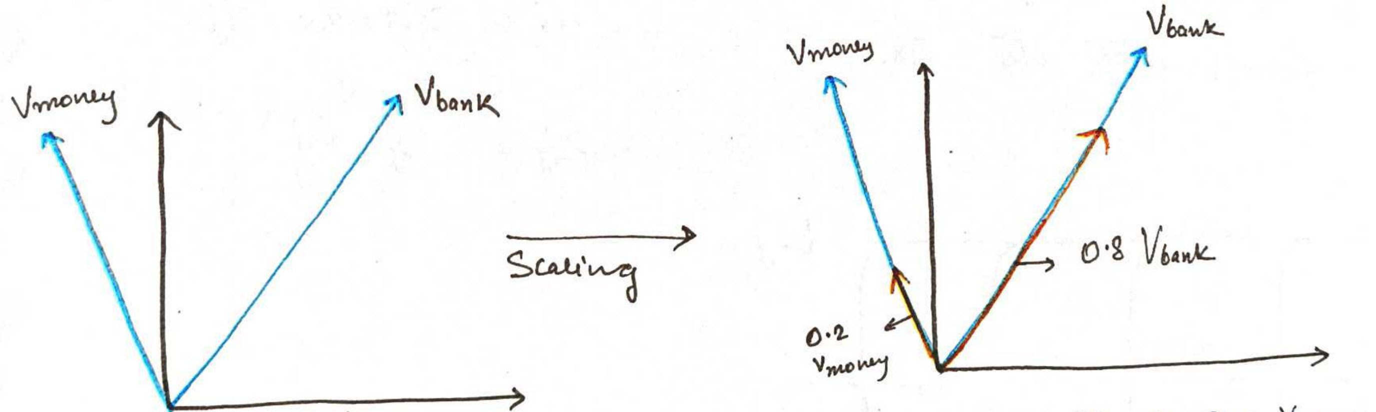
$S_{21} = 10$ $\quad$ $S_{22} = 32$

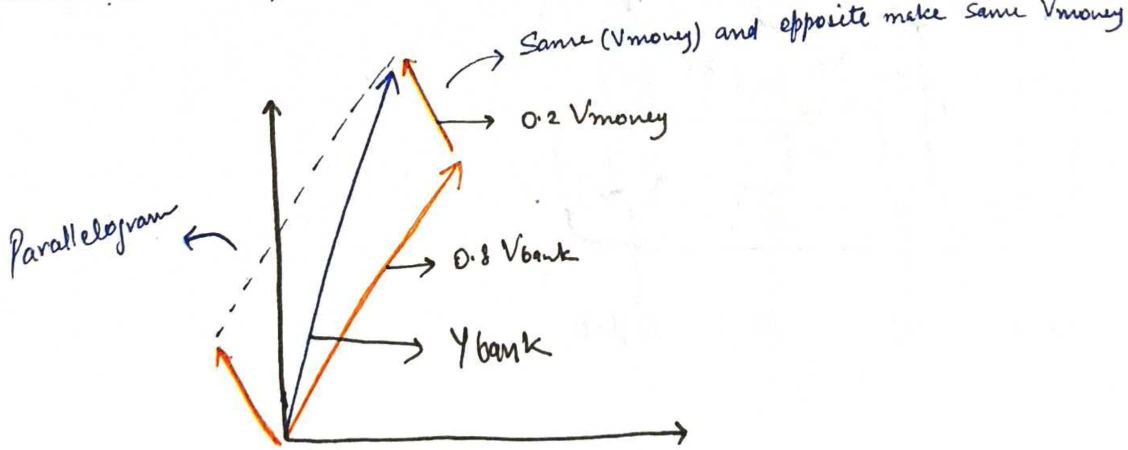
Scaling

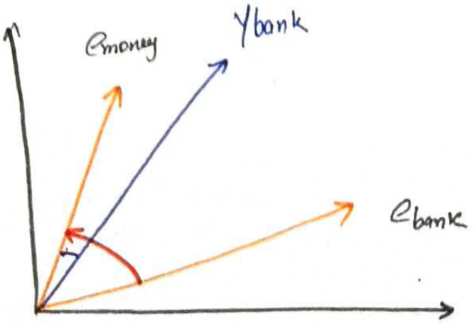
$S'_{21} = \frac{10}{\sqrt{2}} = 7.09$ $\quad$ $S'_{22} = \frac{32}{\sqrt{2}} = 22.69$

Softmax

$W_{21} = 0.2$ $\quad$ $W_{22} = 0.8$

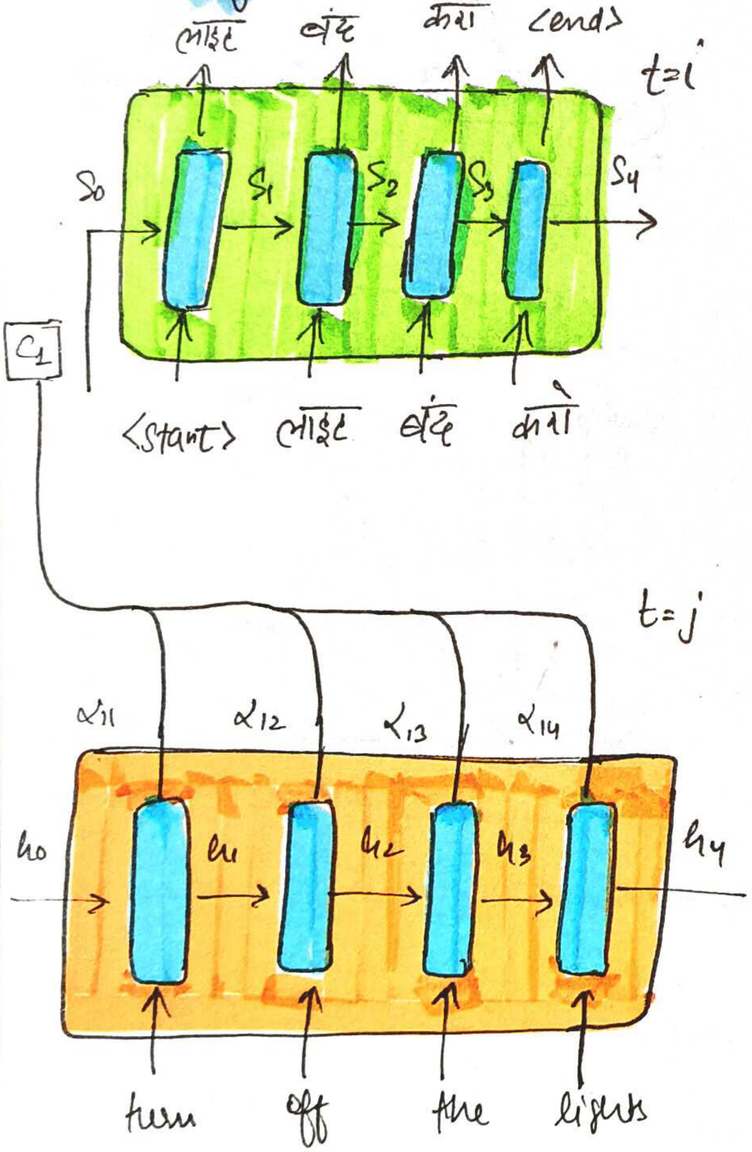






- * Embedding bank bohot dur tha Embedding money starting mai.
- * Now, distance between $\underset{e_{money}}{\underset{\downarrow}{\text{money}}}$ and $\underset{Y_{bank}\ (new)}{\underset{\downarrow}{\text{bank}}}$ is decreased.
- * Self Attentition is content aware.

Why is self Attention called "self"?

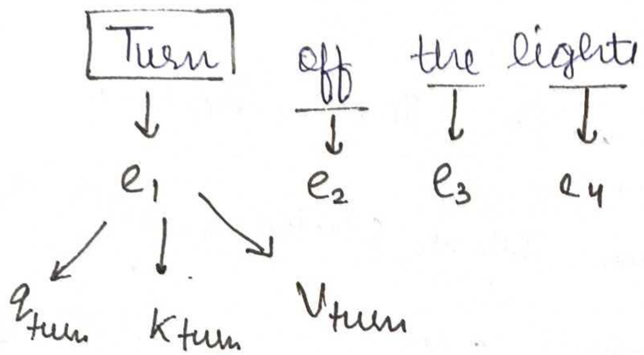


$$c_i = \sum \alpha_{ij} h_j$$

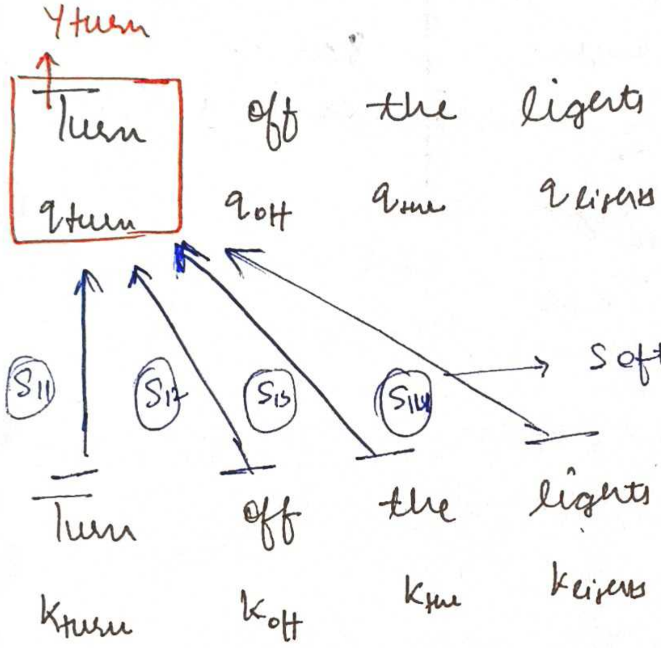
$$\alpha_{ij} = \text{softmax}(e_{ij})$$

$$e_{ij} = s_i^T h_j$$

Self Attention



$$Y_{turn} = W_{11} V_{turn} + W_{12} V_{off} + W_{13} V_{the} + W_{14} V_{lights}$$



Softmax (S_{ij}) ⇒ W_{11} = Softmax (S_{11})

W_{12} = Softmax(S_{12})

* Y_{turn} similar to C_i

* α_{ij} = similar to W_{11} = Softmax(S_{11})

Softmax(e_{ij})

* S_i → query

h_j → key

h_j → value

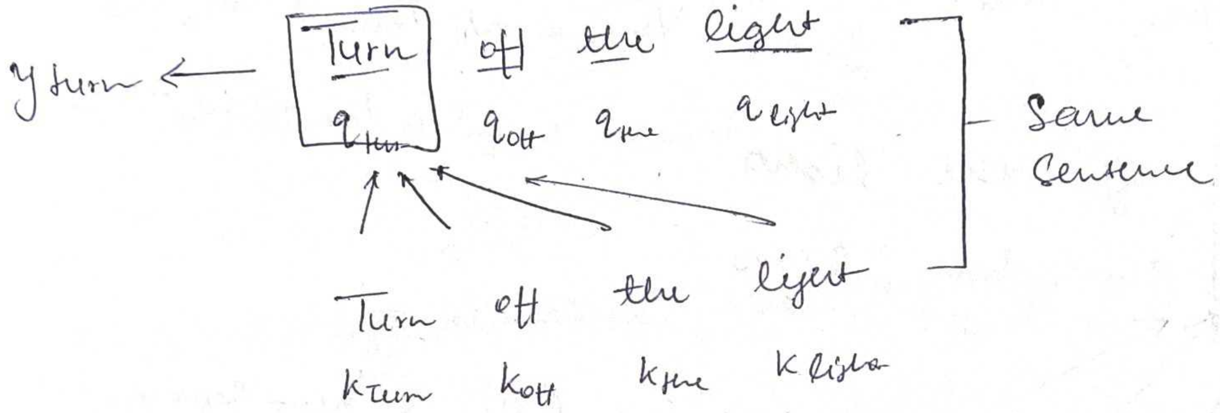
* Because of these similarity we called Attention (Luong Attention similar with Self Attention)

Why "Self" called in Self Attention?

In Luong Attention calculate between two different Sequences.

In Self Attention calculate in a Single Sequence

like calculate similarity between same sentence



Problem with Self Attention

[The man saw the astronomer with a telescope]

Meaning:

1. बंदे ने दूरबीन पर एक एस्ट्रोनॉमर देखा।

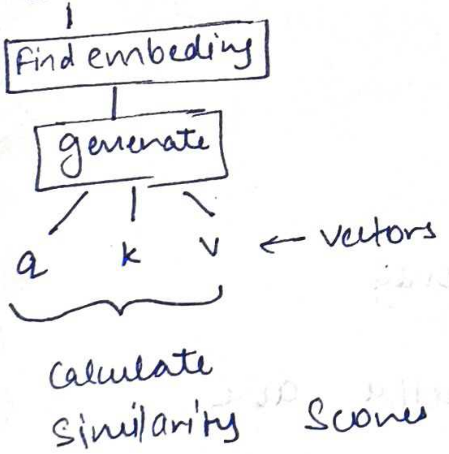
2. बंदे ने एस्ट्रोनॉमर को देखा उसके पास दूरबीन था

So, this Sentence is Ambiguity. 2 meaning of single sentence.

But Self Attention capture only one meaning from both meaning (1 and 2).

* How self Attention works?

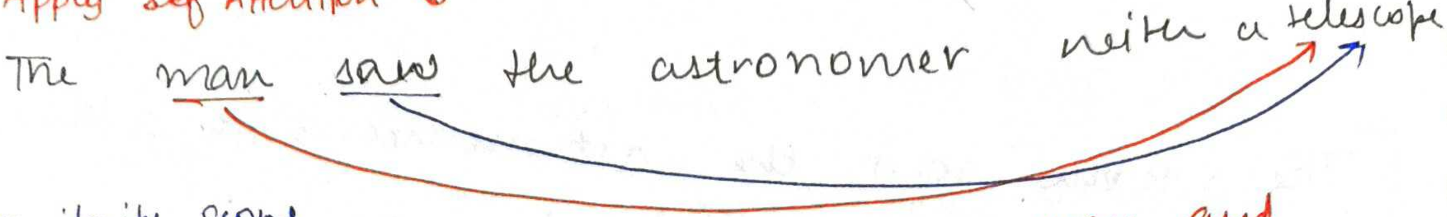
The man saw the astronomers with a telescope



* Same process for all words.

There is a chance to capture first meaning.

Apply self Attention ⤵



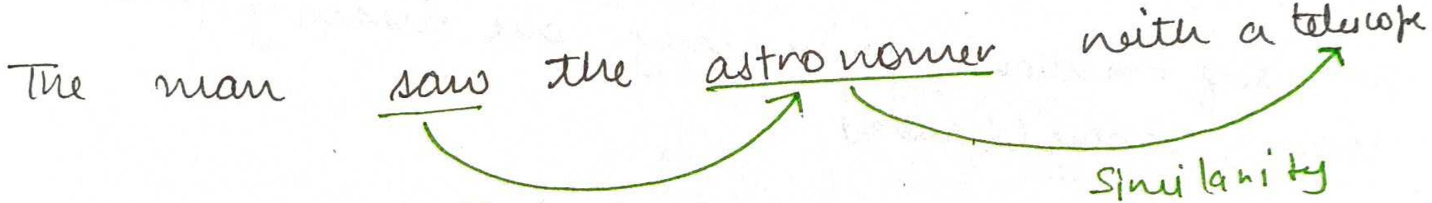
The man saw the astronomer with a telescope

Similarity score Saw and telescope high

Similarity score betn man and telescope is very high.

So, Meaning is : Man telescope लगा के देख रहा है

Apply self Attention ⤵



The man saw the astronomer with a telescope

Similarity score betn saw and astronomers is very

Similarity Score betn astronomers and telescope is very high

So Meaning: Man ने Astronomers को देखा And uske हाथ मे telescope tha.

Problem : Self Attention figure out only Single perspective. If Multiple perspective are present then self Attention cannot capture.

In NLP, there are multiple Scenerio to capture multiple perspective.

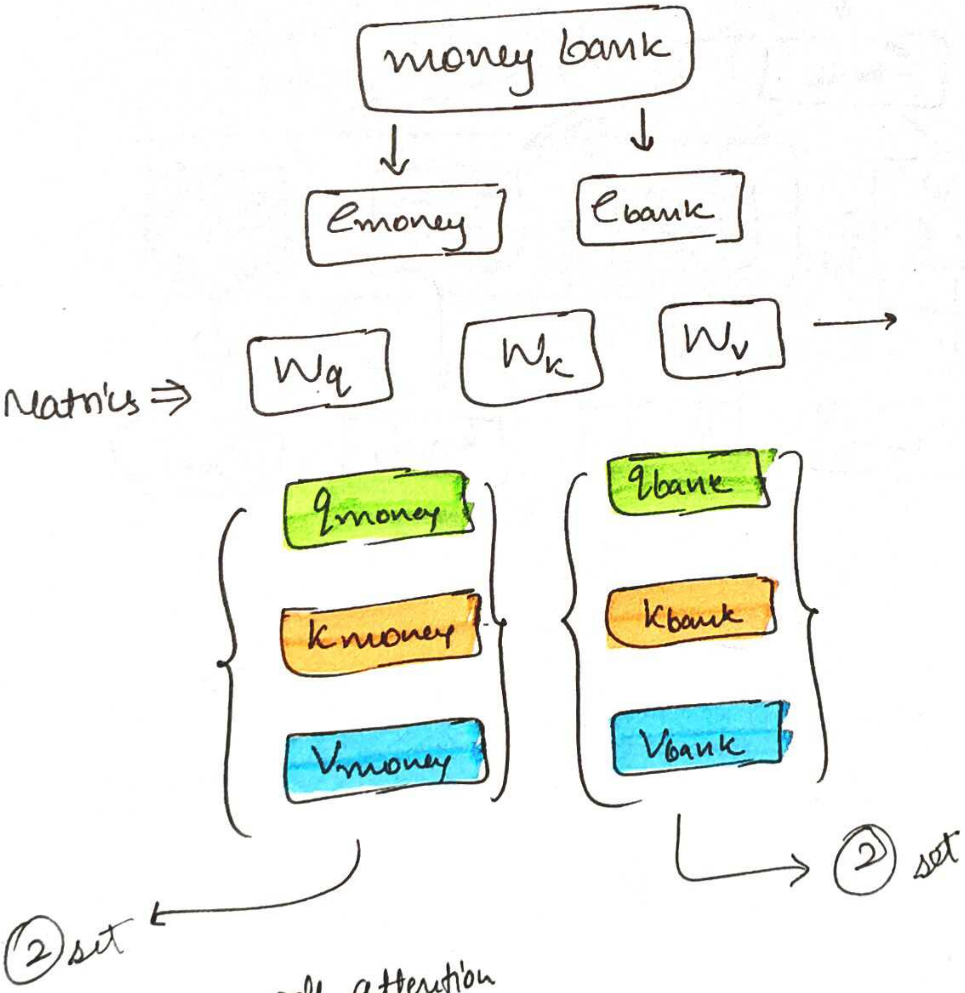
example → Document Summorization tool.

Multi-head Attention

The man saw the astronomer with a telescope

↳ 2 Meaning

So, we use two self Attention.



W_q^1 W_k^1 W_v^1

W_q^2 W_k^2 W_v^2

If self Attention is three

W_q^3 W_k^3 W_v^3

because two self attention