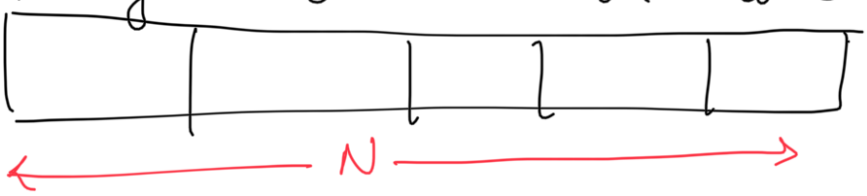
Key, Query in self attention

Sentence → Mary had a little lamb



$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V$$

Query, Key, Value

↳ self attn

embedinds of mary

d_q → what we want

d_K → what other embedins respond

d_V → what we understood from it.

$$d_v = d_q = d_K$$

We can pack all values in

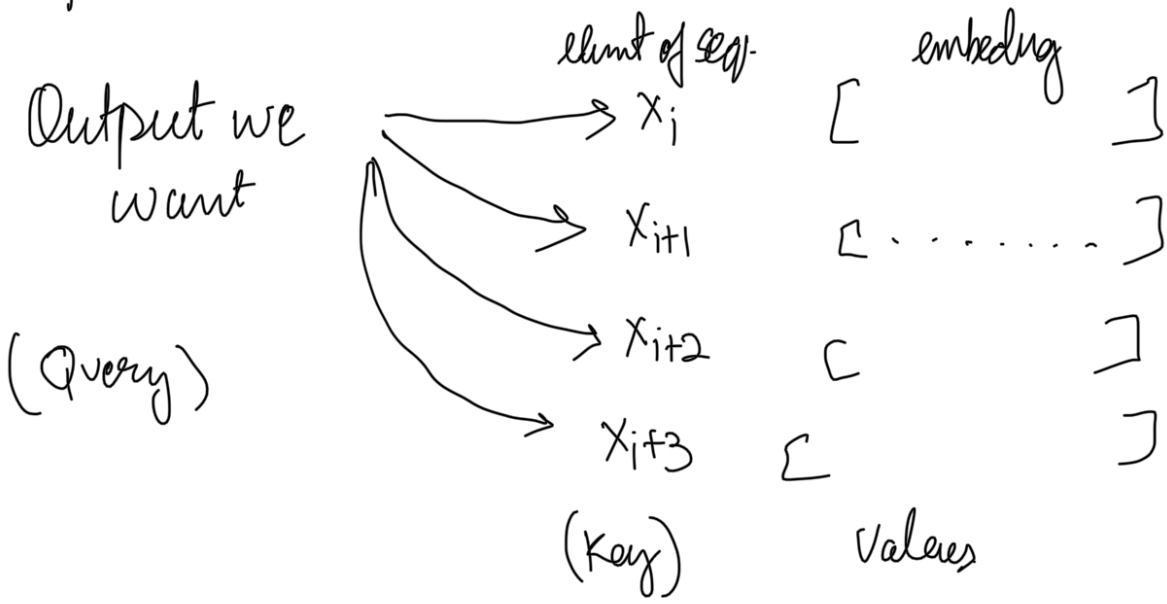
$$\overset{\uparrow}{\underset{\downarrow}{N}} \begin{bmatrix} N \times d_q \end{bmatrix} \times \begin{bmatrix} d_K \times N \end{bmatrix} = \begin{bmatrix} N \times N \end{bmatrix}$$

d_q

so as compute everything at once

$$[N \times N] \times [N \times d_v] = \text{result}$$

logic



Pick the one we want

masked self attn continued

$$z_{13} = \frac{K_1' \, q_3}{\sqrt{d}}$$ (How? look in the PAPER)

$$\begin{bmatrix} w_{13} \\ w_{23} \\ w_{33} \\ w_{43} \\ w_{53} \end{bmatrix} = \begin{bmatrix} \exp(z_{13}) \\ \exp(z_{23}) \\ \exp(z_{33}) \\ 0 \\ 0 \end{bmatrix} \Rightarrow \underbrace{\begin{bmatrix} w_{13} \\ w_{23} \\ w_{33} \\ 0 \\ 0 \end{bmatrix} \frac{1}{\sum_{i=1}^{3} w_{i3}}}_{\text{nothing but softmax}}$$

$$= \begin{bmatrix} \text{Softmax}(z_{13}, z_{23}, z_{33}) \\ 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow y_3 = \sum_{i=1}^{5} v_i w_{i3}$$

Since $w_{43}, w_{53} = 0$ it doesn't matter