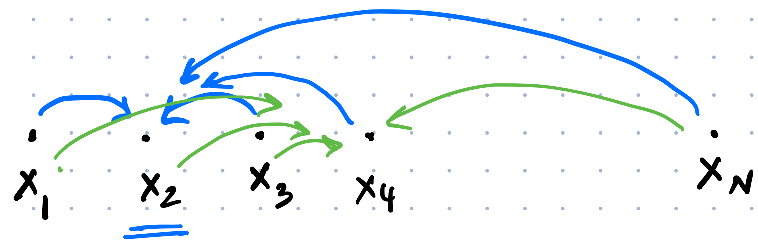
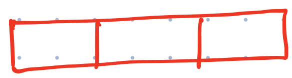
Transformers

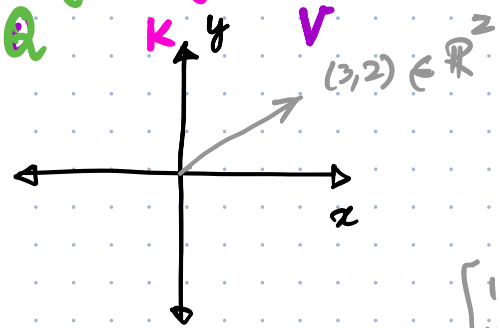
- Self-attention:

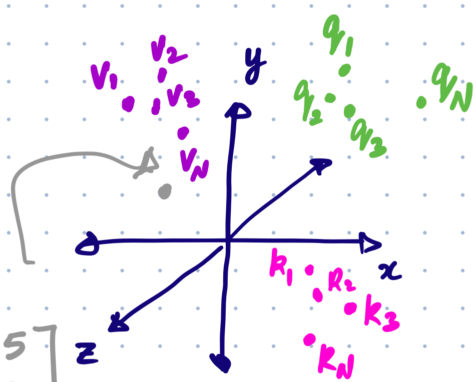




Query, Key, Value

Q K V



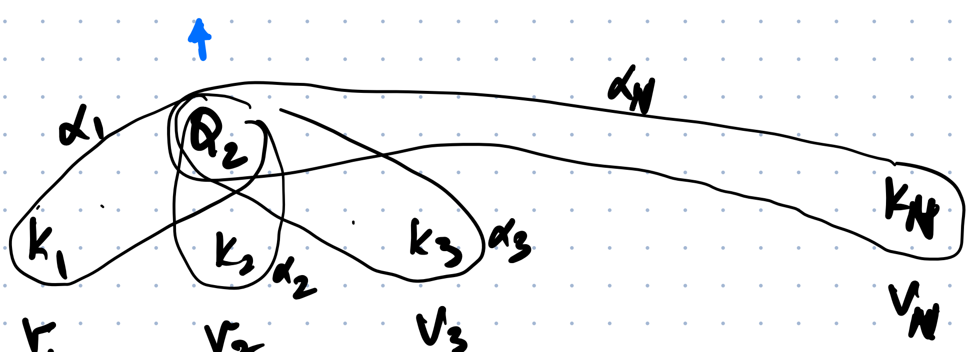


$$\begin{bmatrix}1 & 3\\4 & 5\\6 & 0\end{bmatrix}\begin{bmatrix}3\\2\end{bmatrix}=\begin{bmatrix}5\\22\\18\end{bmatrix}$$

$\in \mathbb{R}^{3\times 2}$

Q. What is the attention on x_2

x_1	x_2	x_3	x_N	$\} \in \mathbb{R}^d$
Q_1	Q_2	Q_3	Q_N	
K_1	K_2	K_3	K_N	$\} \in \mathbb{R}^{d_{model}}$
V_1	V_2	V_3	V_N	



$$x_2' = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \dots + \alpha_N v_N$$

Aside:

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \end{bmatrix} \in \mathbb{R}^{N \times d}$$

$$Q = XQ_W = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \end{bmatrix} \begin{bmatrix} & Q_W & \end{bmatrix} \in \mathbb{R}^{d \times d_{model}}$$

$$= \begin{bmatrix} - & Q_1 & - \\ - & Q_2 & - \\ & \vdots & \\ - & Q_N & - \end{bmatrix} \in \mathbb{R}^{N \times d_{model}}$$

$$K = XK_W = \begin{bmatrix} - & k_1 & - \\ - & k_2 & - \\ & \vdots & \\ - & k_N & - \end{bmatrix} \in \mathbb{R}^{N \times d_{model}}$$

$$V = XV_W = \begin{bmatrix} - & V_1 & - \\ - & V_2 & - \\ & \vdots & \\ - & V_N & - \end{bmatrix} \in \mathbb{R}^{N \times d_{model}}$$

Attention:

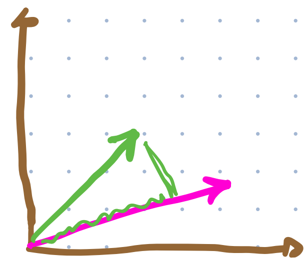
$$QK^T = \begin{bmatrix} - & q_1 & - \\ - & q_2 & - \\ & \vdots & \\ - & q_N & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ k_1 & k_2 & \cdots & k_N \\ | & | & & | \end{bmatrix}$$

$$\in \mathbb{R}^{d_{model} \times d_{model}}$$

Scaled dot-product attention ←

$$= \begin{bmatrix} q_1 \cdot k_1 & q_1 \cdot k_2 & \cdots & q_1 \cdot k_n \\ q_2 \cdot k_1 & q_2 \cdot k_2 & \cdots & q_2 \cdot k_n \\ & & & \\ q_n \cdot k_1 & q_n \cdot k_2 & \cdots & q_n \cdot k_n \end{bmatrix}$$

$$\left(\frac{QK^T}{\sqrt{d_{model}}} \right)$$



↓ let's compute

$$\left(\frac{QK^T}{\sqrt{d_{model}}} \right) V$$

$$\begin{bmatrix} q_1 \cdot k_1 & q_1 \cdot k_2 & & q_1 \cdot k_n \\ & & & \\ & & & \\ q_n \cdot k_1 & & & q_n \cdot k_n \end{bmatrix} \begin{bmatrix} — & V_1 & — \\ — & V_2 & — \\ & \vdots & \\ — & V_N & — \end{bmatrix}$$

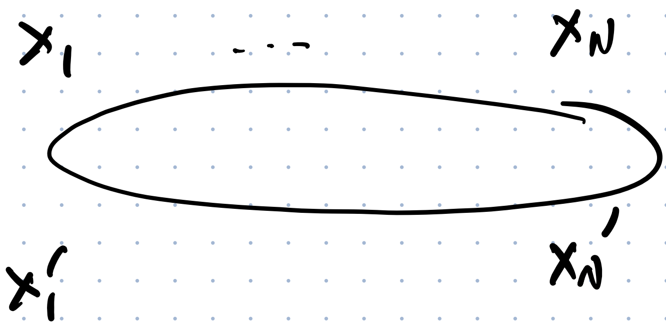
$$\mathbb{R}^{N \times N} \qquad \mathbb{R}^{N \times d_{model}}$$

$$\begin{bmatrix} (q_1 \cdot k_1)v_1 + (q_1 \cdot k_2)v_2 \cdots (q_1 \cdot k_n)v_N \\ \\ \\ \end{bmatrix} = \begin{bmatrix} q_1 \cdot k_1 & q_1 \cdot k_2 & & q_1 \cdot k_n \\ & & & \\ q_n \cdot k_1 & & & q_n \cdot k_n \end{bmatrix} \begin{bmatrix} — v_1 — \\ — v_2 — \\ \vdots \\ — v_N — \end{bmatrix}$$

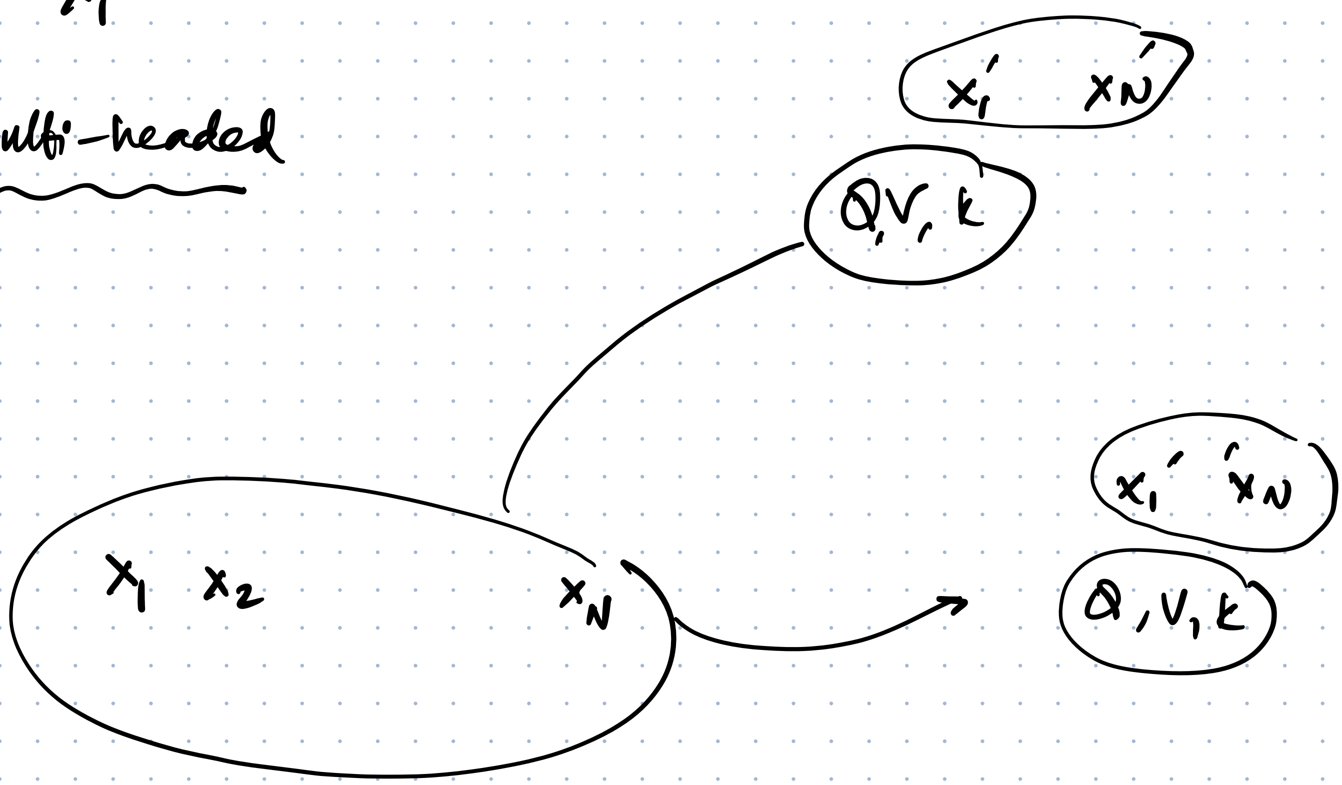
$\downarrow$

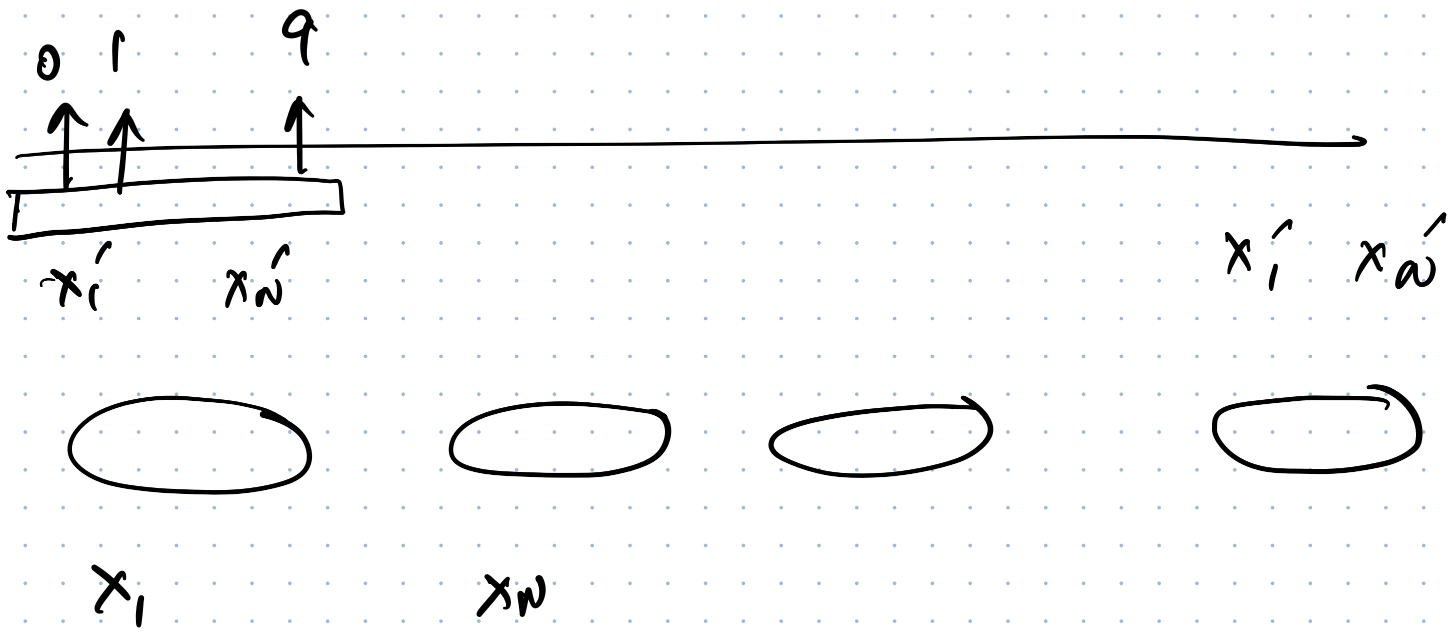
$N \times d_{model}$ $\qquad \mathbb{R}^{N \times N} \qquad \mathbb{R}^{N \times d_{model}}$

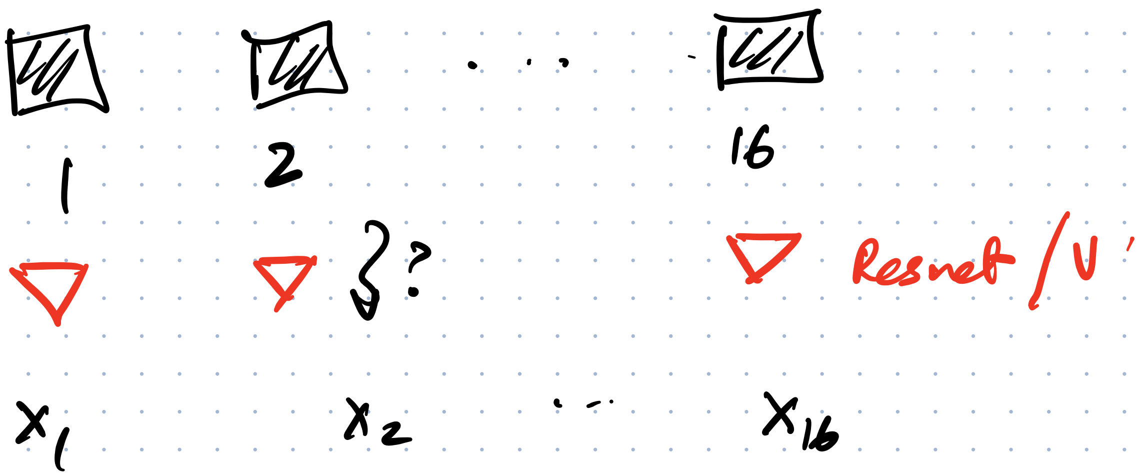
$X \longrightarrow X'$



Multi-headed







This is an amazing course.

↓ word embedding

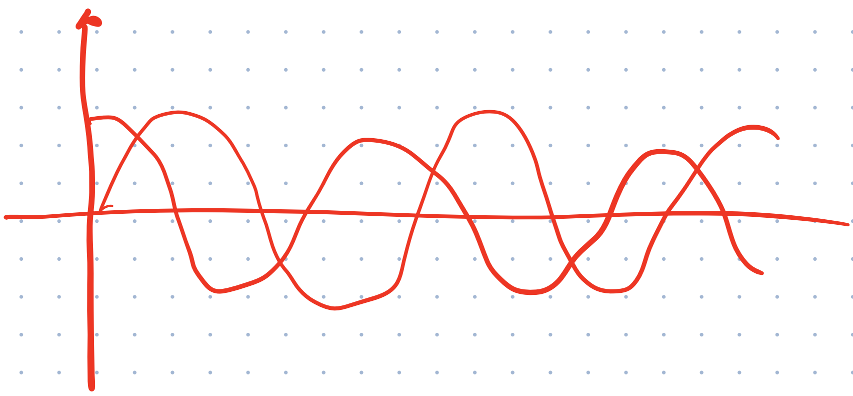
king → v_1

man → v_2

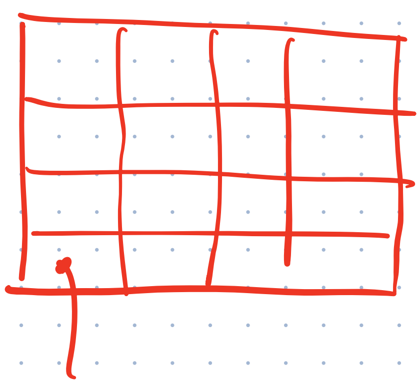
woman → v_3

queen → v_4

$$v_1 - v_2 + v_3 = v_4$$



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



i,j

$$X_1 = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 7 \\ 0 \\ 3 \end{bmatrix} \quad \begin{bmatrix} 4 \\ 1 \\ 4 \end{bmatrix}$$

i: 0, 1, 2

pos: 0, 1, 2

dmodel = 3

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

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$x_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad X_2 = \begin{bmatrix} \sin(1/1) = 0.0174 \\ \cos(1/465) = 0.99 \\ \sin\left(1/10000^{4/3}\right) = 0.00 \ldots 8 \ldots \end{bmatrix}$$