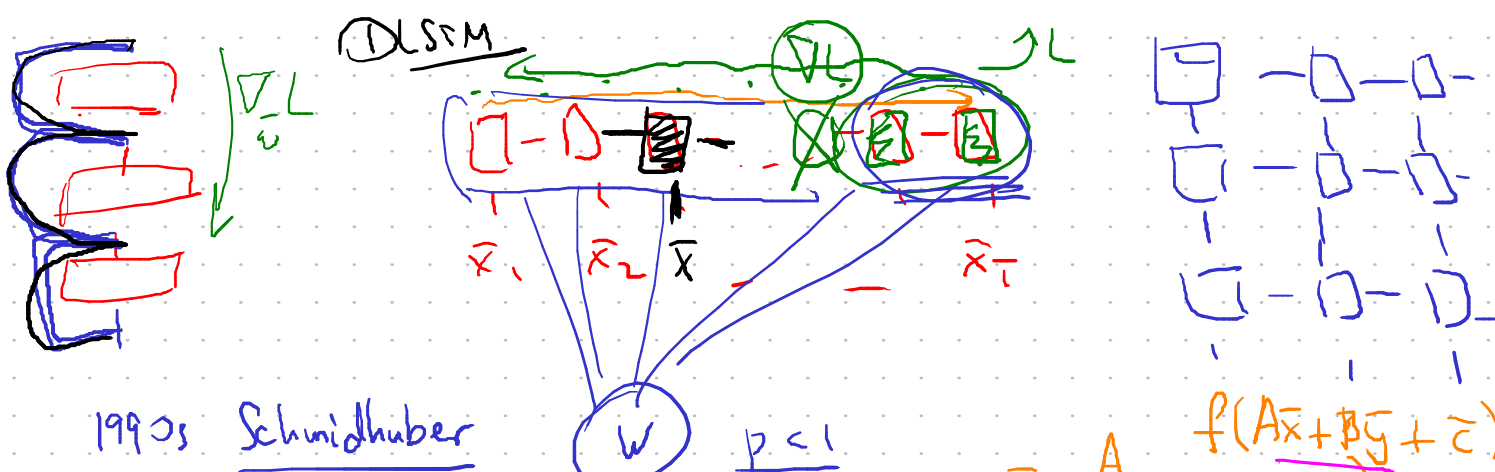
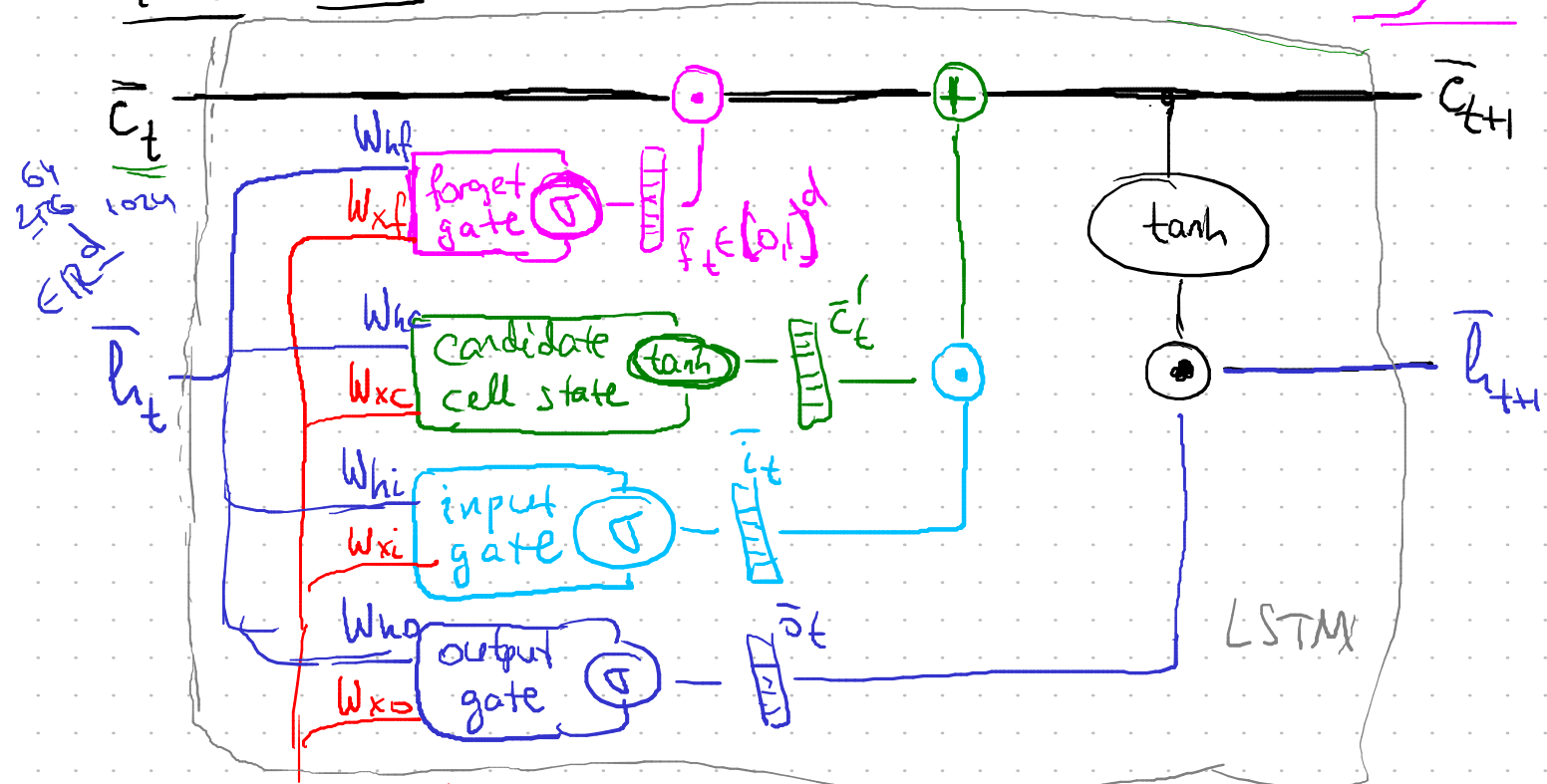
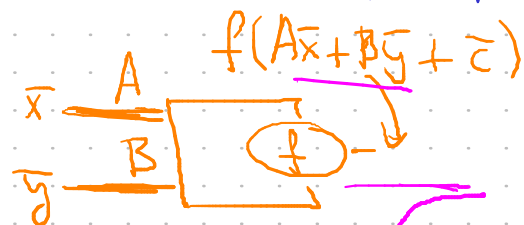




1990s Schmidhuber

LSTM long short-term memory

1995 2000

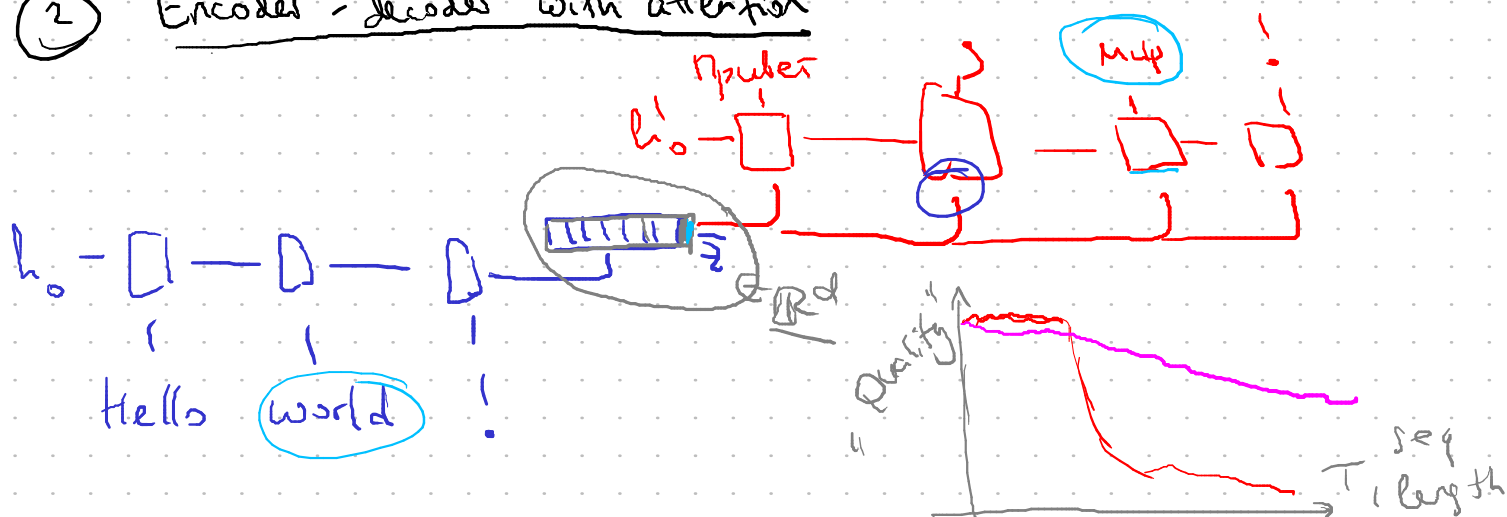


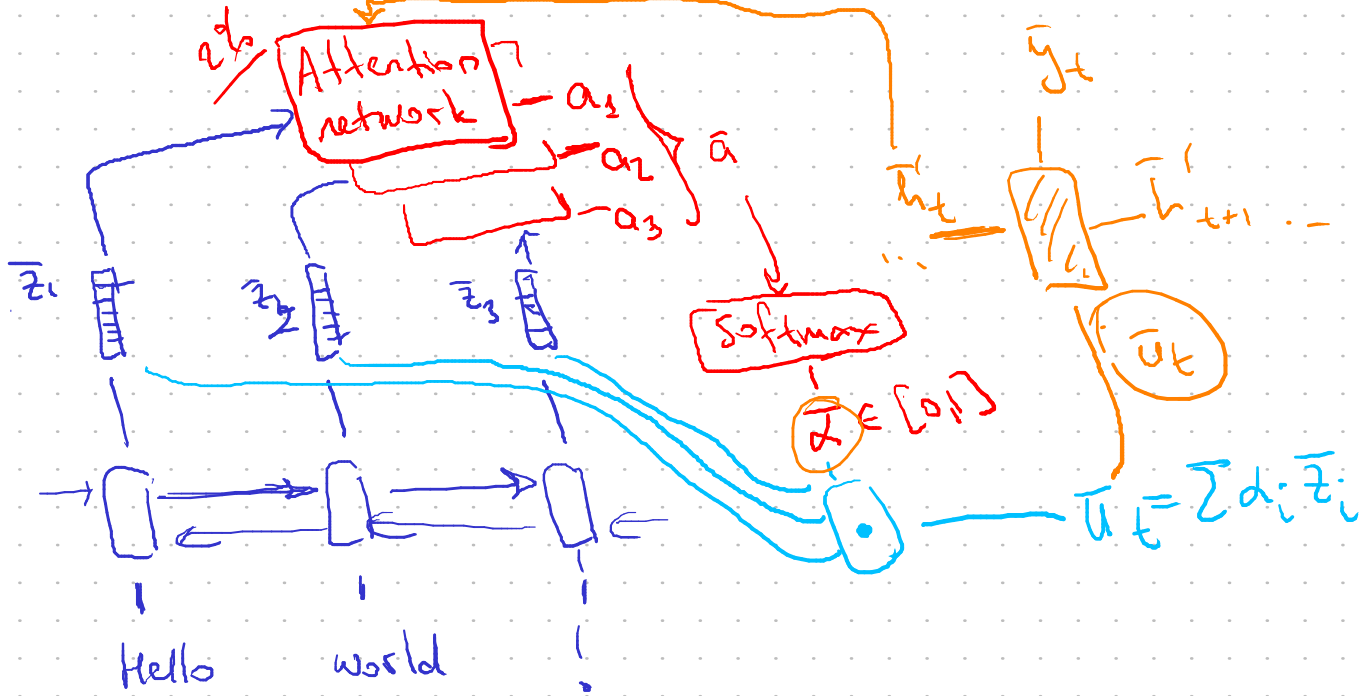
$$\bar{x}_t \in \mathbb{R}^{d'}$$

$$\frac{W_{xf}, W_{xf}}{dx dx'}, \dots, \frac{W_{xo}, W_{xo}}{dx dx'}$$

AlphaStar

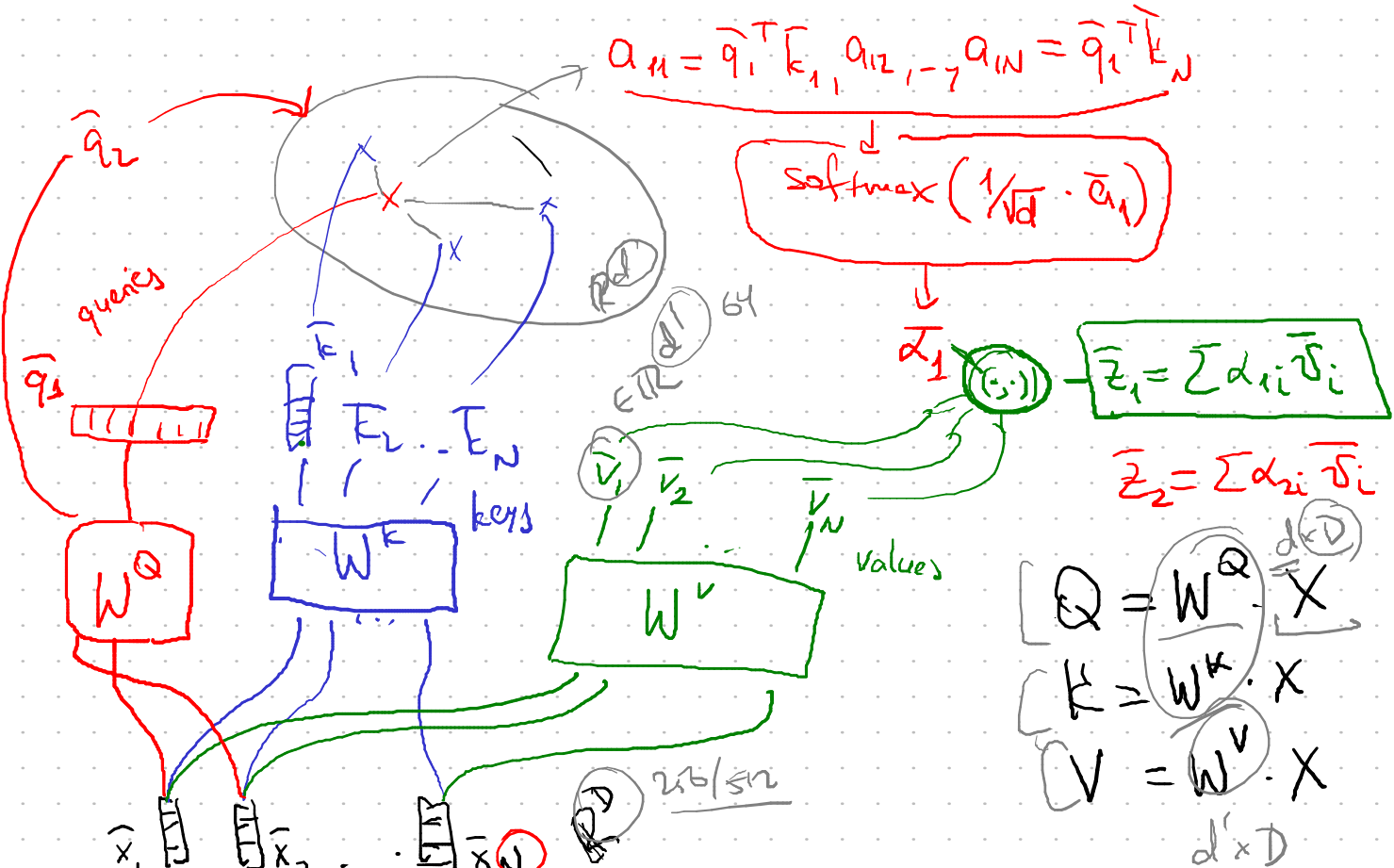
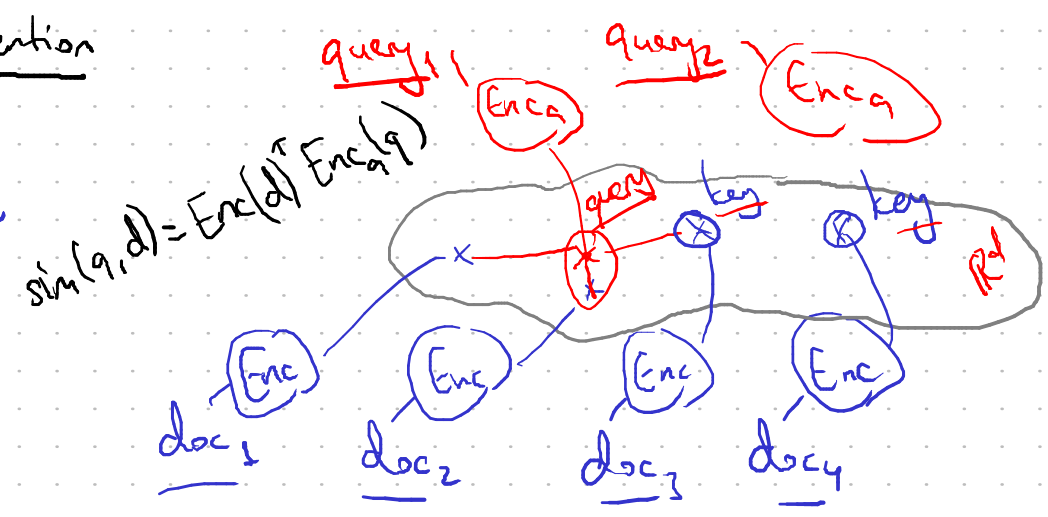
② Encoder - decodes with attention





③ Self-attention

Information retrieval



$$N \times N \rightarrow Q \cdot K^T = \left(\frac{1}{\sqrt{d}} \bar{q}_i^T \bar{k}_j \right)$$

Quadratic complexity

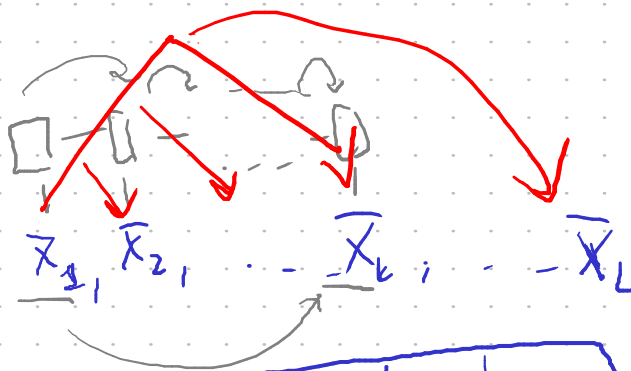
$$Att = \text{softmax} \left(\frac{1}{\sqrt{d}} Q K^T \right)$$

$$Z^{(i)} = \text{softmax} \left(\frac{1}{\sqrt{d}} Q^{(i)} K^T \right) \cdot V^{(i)}$$

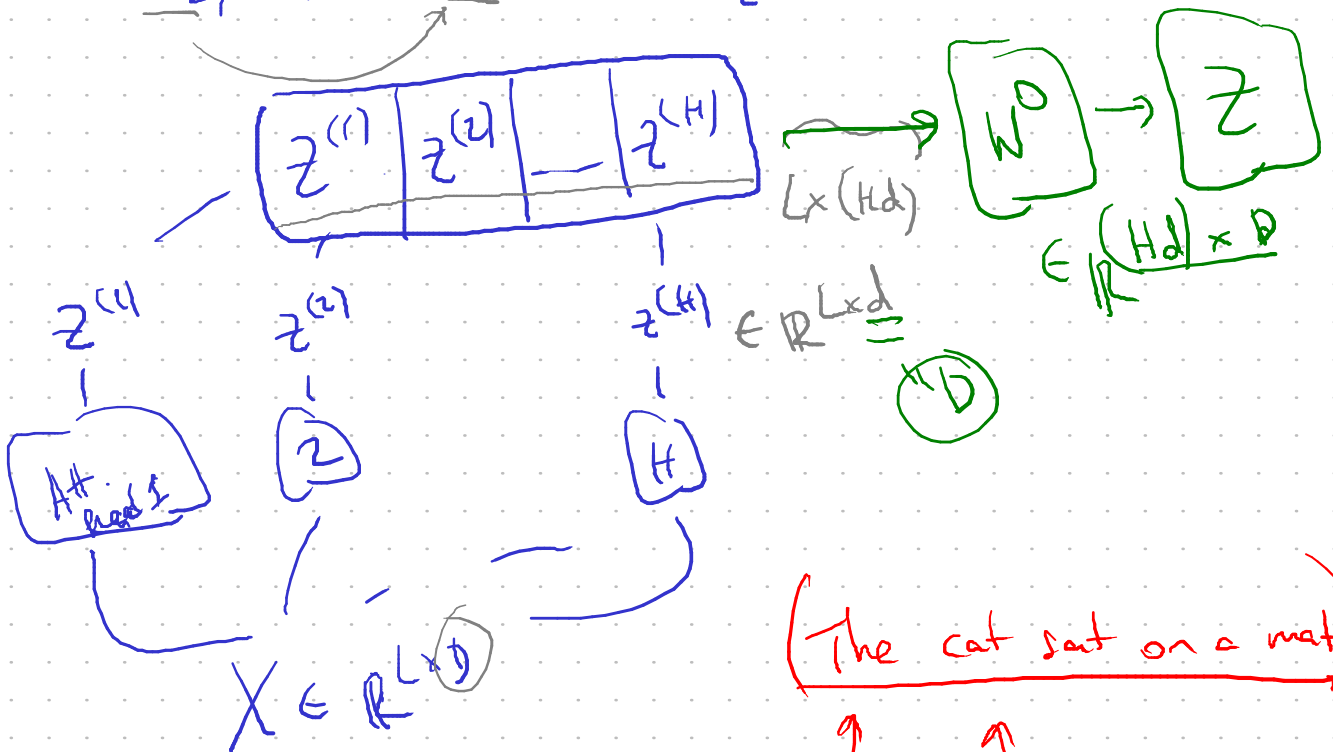
$$\bar{q}^T \bar{k} \quad \mathbb{R}^{d=64}$$

$$q_i \approx \pm 1 \quad k_j \approx \pm 1$$

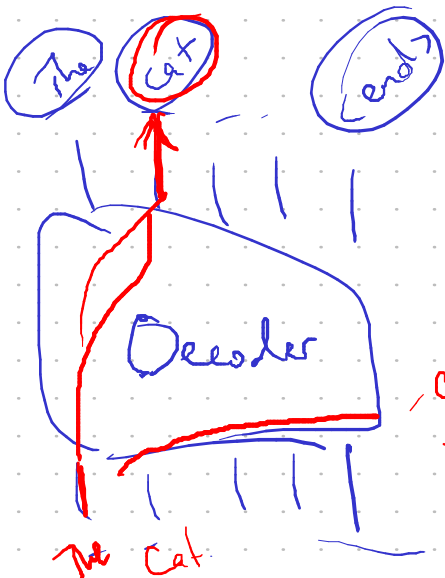
$$\frac{d}{2} \quad \begin{pmatrix} +1 \\ -1 \end{pmatrix} \approx \begin{pmatrix} +\sqrt{d} \\ -\sqrt{d} \end{pmatrix}$$



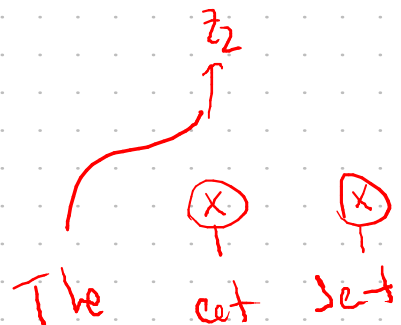
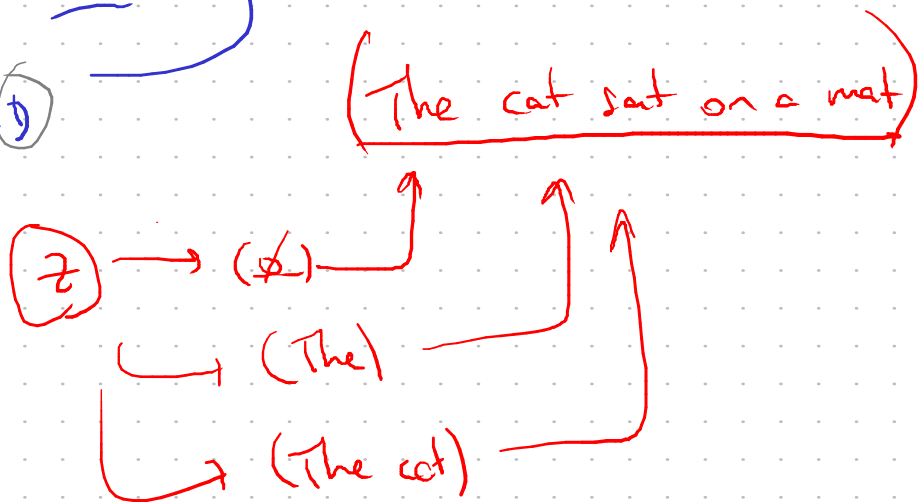
- 1 step from V_i to V_j (+)
- cost: $O(L^2)$ compl. (-)



(... t-1 -> t)



- causal mask



Enc-dec
att. :

prefix X (decoder) $\rightarrow W^Q \rightarrow Q$
 encoder's $Z \rightarrow W^K, W^V \rightarrow K, V$

$$1 \rightarrow \bar{e}_1$$

$$2 \rightarrow \bar{e}_2$$

...

$$\bar{e}_i - \bar{e}_j \approx \bar{e}_{i+k} - \bar{e}_{j+k}$$

RoPE

rotary
position

$$\bar{X} = \begin{pmatrix} \boxed{} & x_{11} \\ \boxed{} & x_{12} \\ \boxed{} & x_{1n} \\ \vdots & \vdots \\ \boxed{} & x_{in} \end{pmatrix}$$

$$\begin{pmatrix} \bar{e}_i \\ \bar{e}_j \\ \vdots \\ \bar{e}_n \end{pmatrix}$$

$$\begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}$$

$$\theta \begin{matrix} \nearrow k\theta \searrow (k+1)\theta \\ \nwarrow l\theta \nearrow \end{matrix}$$