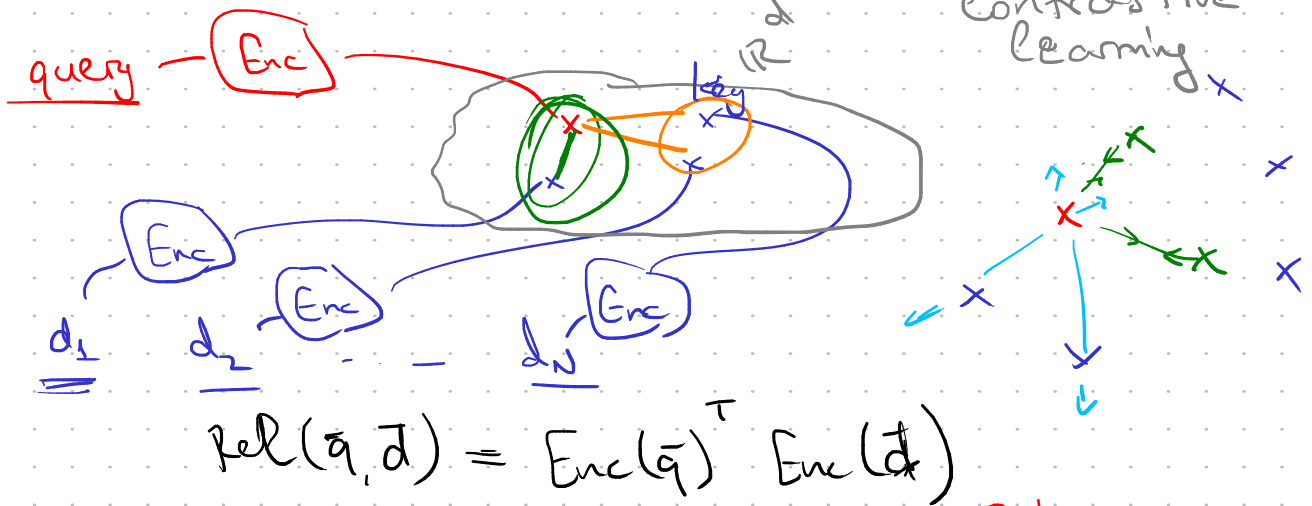
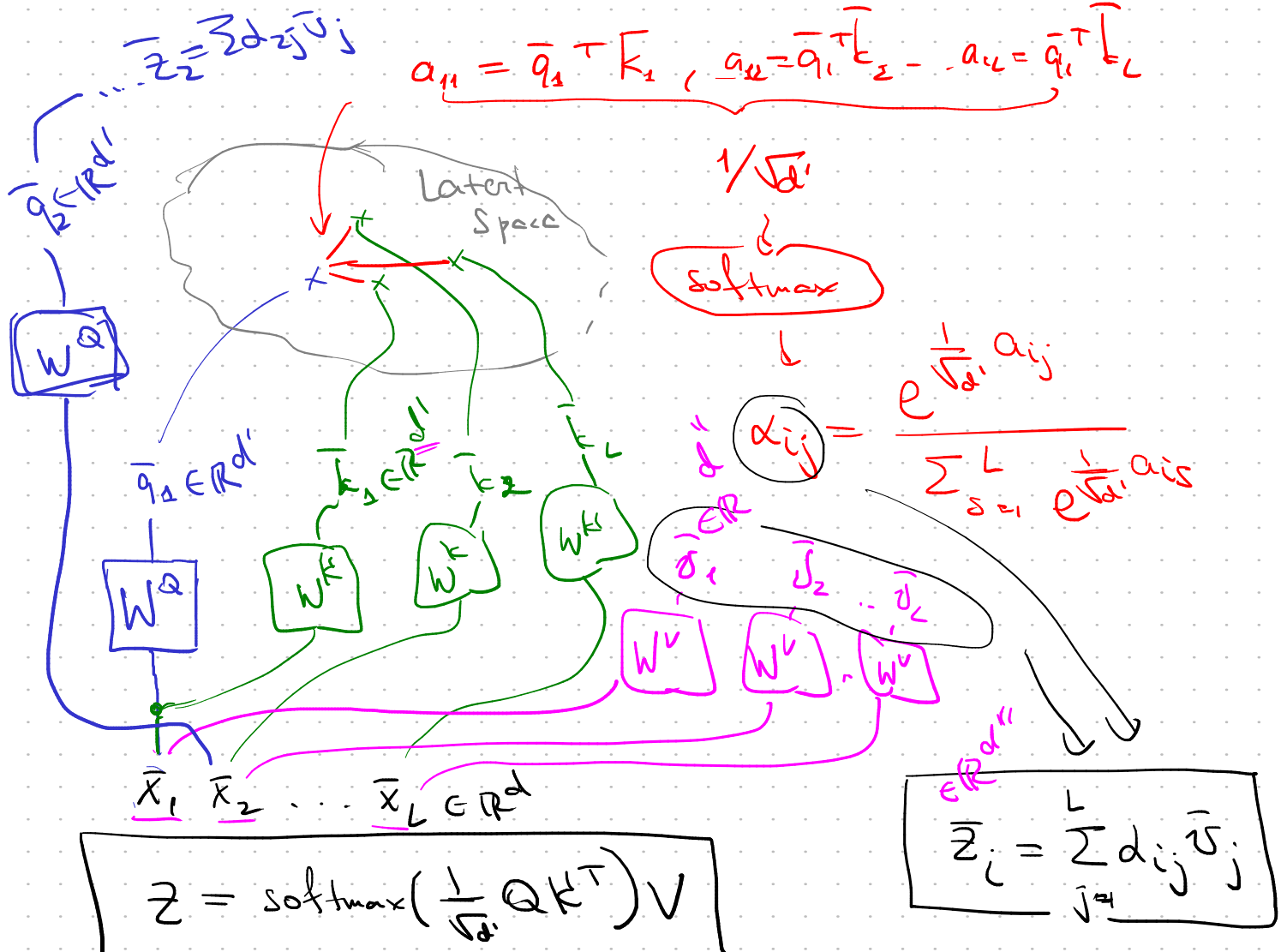
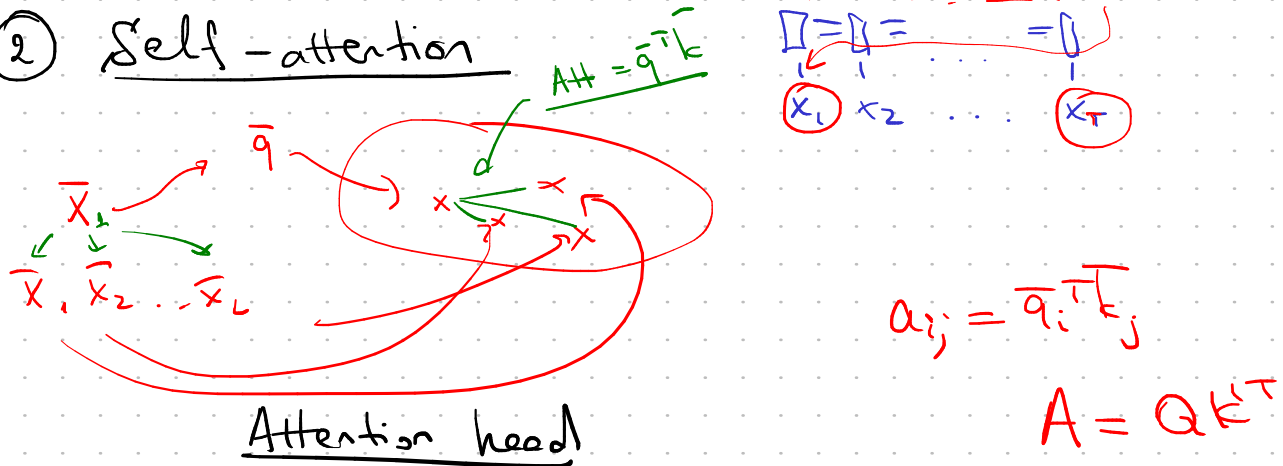
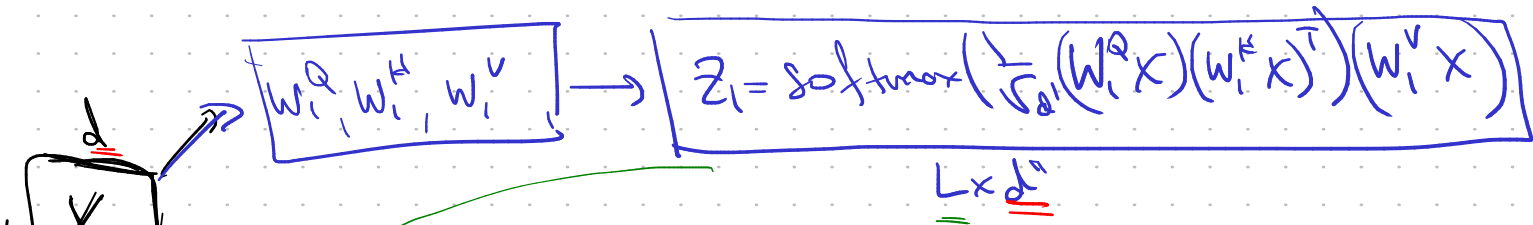


① Information retrieval

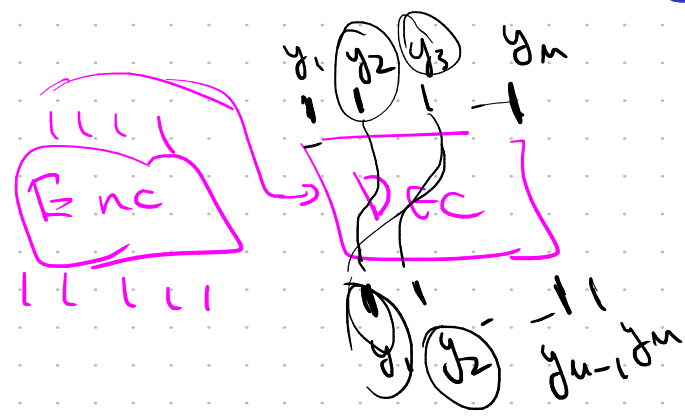
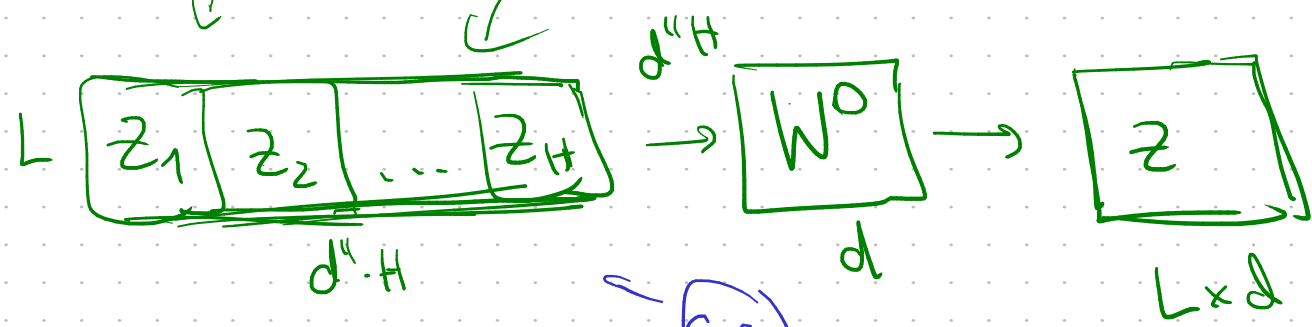


② Self-attention

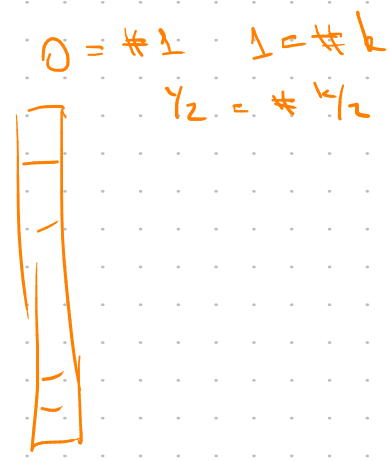
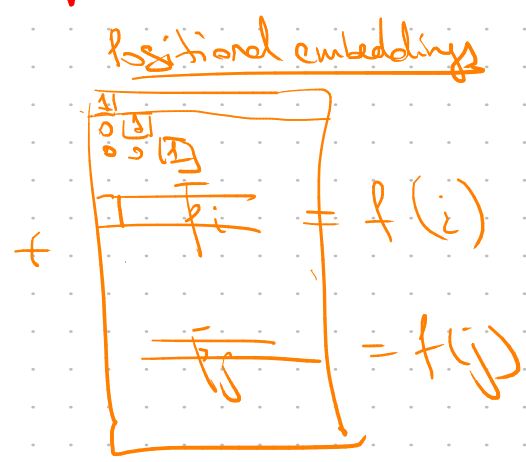
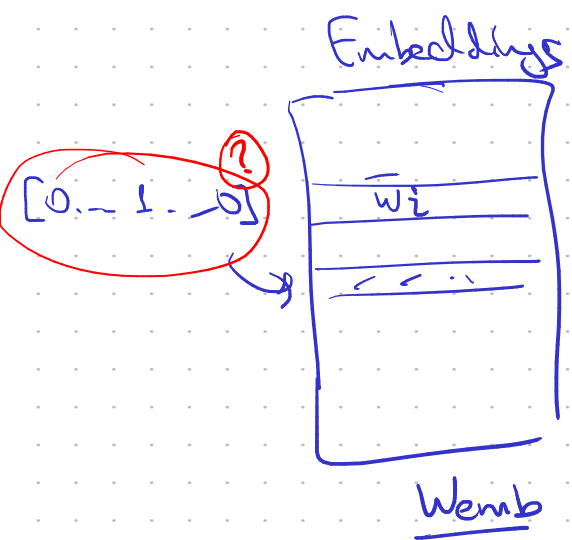
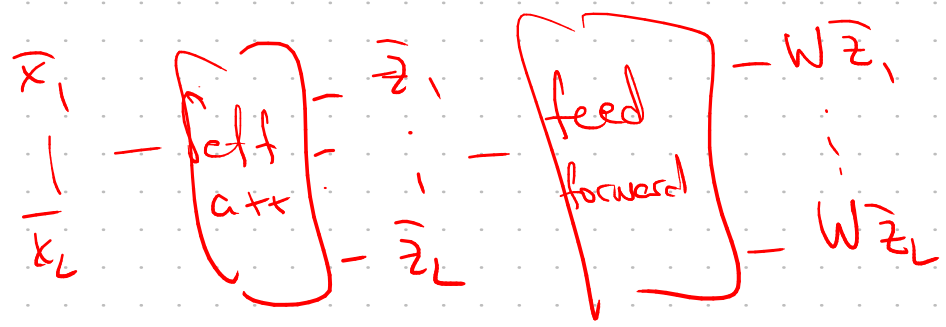
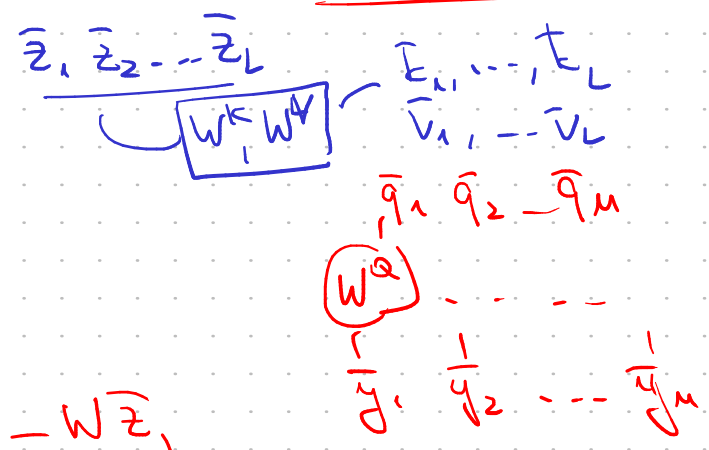




Multi-head attention



Enc-dec att.



RoPE embeddings

rotary

pos

$$0 < \varepsilon < \frac{\pi}{2L}$$

$$\text{RoPE}(\bar{x}, m) = \bar{x} \cdot e^{m i \varepsilon}$$

pair
matrix
tokens

$$f(\bar{q}, m)^T f(\bar{k}, n) = g(\bar{q}, \bar{k}, m-n)$$

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & \ddots \end{bmatrix}$$

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

$$\begin{pmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ & & \ddots \end{bmatrix} & \begin{bmatrix} & 1 & \\ & & 1 \\ & & & \ddots \end{bmatrix} \\ & \begin{bmatrix} & & \\ & & \\ & & \ddots \end{bmatrix} \end{pmatrix}$$

$$\begin{aligned} \langle \bar{q} \cdot e^{m i \varepsilon}, \bar{k} \cdot e^{n i \varepsilon} \rangle &= \\ &= \bar{q}^T \bar{k} e^{m i \varepsilon} \overline{e^{n i \varepsilon}} = \\ &= \bar{q}^T \bar{k} e^{(m-n) i \varepsilon} \end{aligned}$$

Restricted self-att.

$$\bar{x}_i \quad \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_i & \dots & \bar{x}_n \end{bmatrix}$$

$$\bar{x}_L$$

$$A$$

