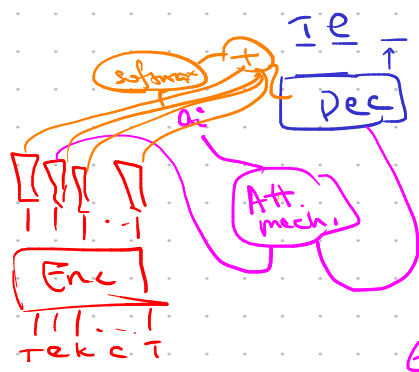
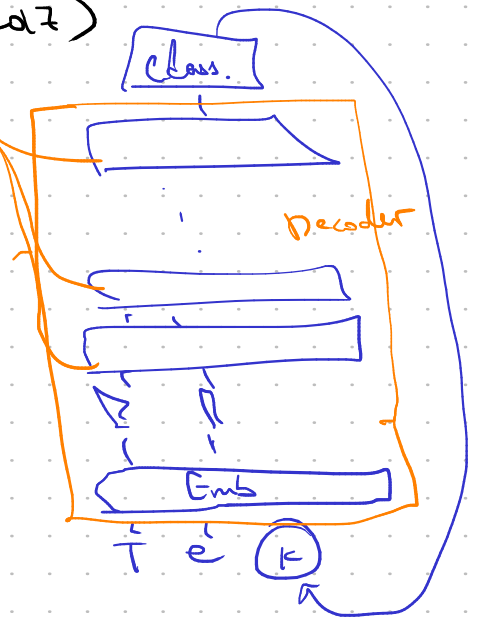
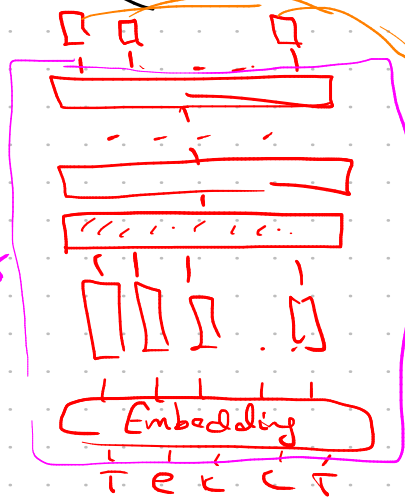


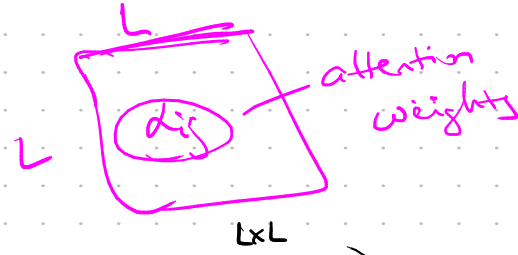
(Vaswani et al., 2017)



Encoder



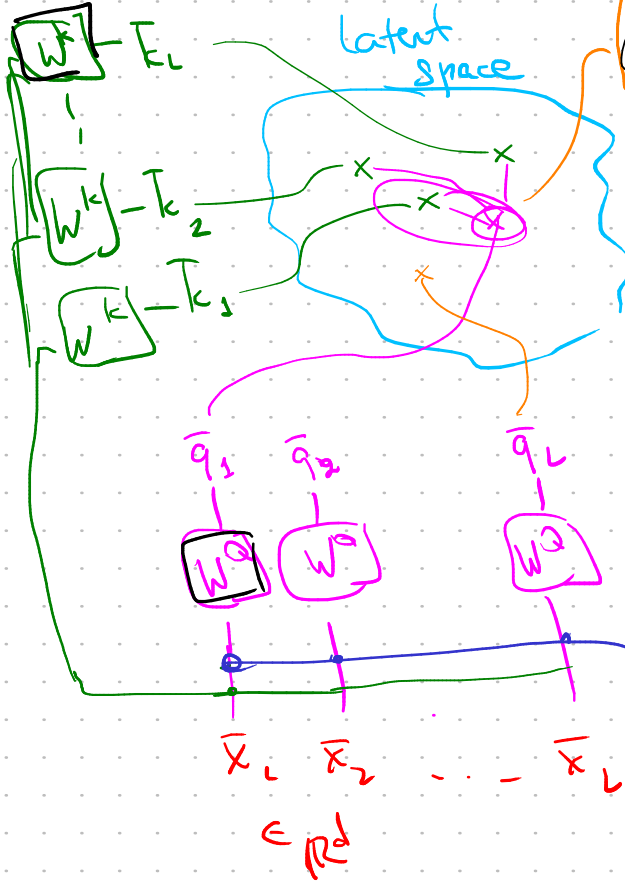
Decoder



$O(L^2)$

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

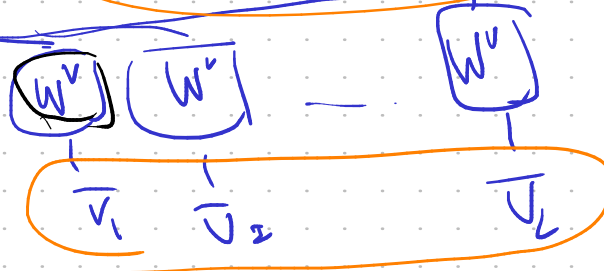
$\bar{q}_i = W^Q x_i$
 $\bar{k}_i = W^K x_i$
 $\in \mathbb{R}^{d'}$



$a_{ij} = \bar{q}_i^T \bar{k}_j$, $a_{11}, \dots, a_{LL}$

$\frac{1}{\sqrt{d_k}}$
 softmax
 $a_{ij} = \frac{e^{\frac{1}{\sqrt{d_k}} a_{ij}}}{\sum_{s=1}^L e^{\frac{1}{\sqrt{d_k}} a_{is}}}$

$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1L} \\ a_{21} & a_{22} & \dots & a_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ a_{L1} & a_{L2} & \dots & a_{LL} \end{pmatrix}$



$W^K, W^Q - d \times d'$
 $W^V - d \times d''$

$\bar{z}_i = \sum_j a_{ij} \bar{v}_j$



Enc-dec. attr.:

$$Q = W^Q X$$

$$\hat{z}_1, \hat{z}_2$$

RoPE

$$K = W^K Z$$

$$V = W^V Z$$



discriminative

$$(\bar{x}, y)$$

$$p(y|\bar{x})$$

generative

$$\bar{x}$$

$$p(\bar{x}) \sim \bar{x}$$

$$p(\bar{x}|E)$$

$$\bar{x} \sim p(\bar{x})$$

$$\bar{z} \rightarrow \bar{x} \sim p(\bar{x})$$

$$p(x_1 \dots x_n) = p(x_1) p(x_2|x_1) p(x_3|x_1, x_2) \dots p(x_k|x_1 \dots x_{k-1}) \dots$$

$$\text{Markov chain: } p(x_k|x_1 \dots x_{k-1}) = p(x_k|x_{k-1})$$

$$\dots p(x_n|x_{n-1})$$

$$\text{RNN: } p(x_k|x_1 \dots x_{k-1}) = p(x_k|x_{k-1}, \underline{h_{k-1}})$$

$$= f(x_{k-1}, h_{k-2})$$