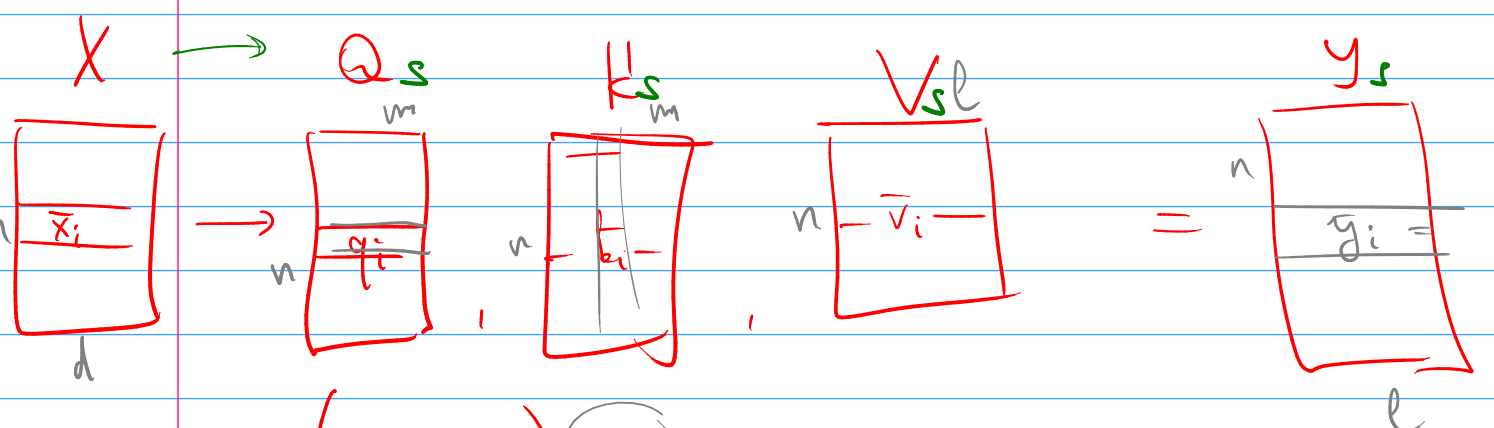


$$x_i \rightarrow \bar{q}_i = \begin{pmatrix} k_{i1} \cdot v_{i1} & \dots & q_i^T k_1 \\ k_{i2} \cdot v_{i2} & \dots & q_i^T k_2 \\ \vdots & \ddots & \vdots \\ k_{in} \cdot v_{in} & \dots & q_i^T k_n \end{pmatrix} \rightarrow y_i = \sum_{j=1}^n \text{softmax} \left( \frac{1}{\sum_m q_i^T k_j} \right) \cdot \bar{v}_j$$

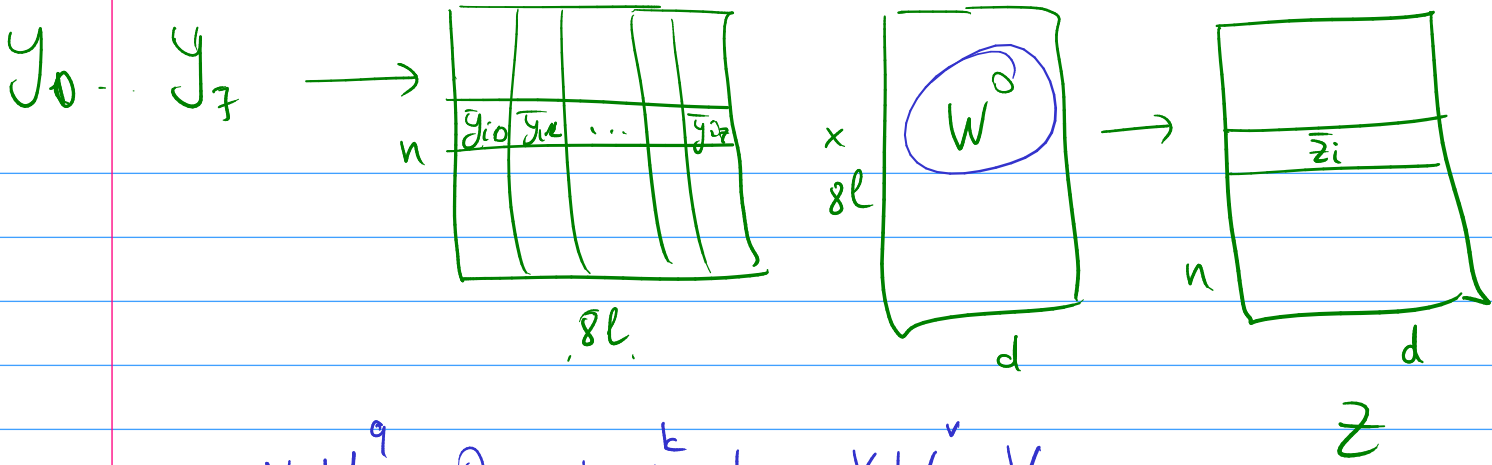
$s = 0 \dots 7$

$a_{ij}$  - self-attention weights



$$\text{softmax} \left( \frac{1}{\sum_m} Q_s K_s^T \right) \cdot V_s = Y_s$$

Dimensions:  $n \times n$  for  $Q_s K_s^T$ ,  $n \times l$  for  $V_s$ , and  $n$  for  $Y_s$ .

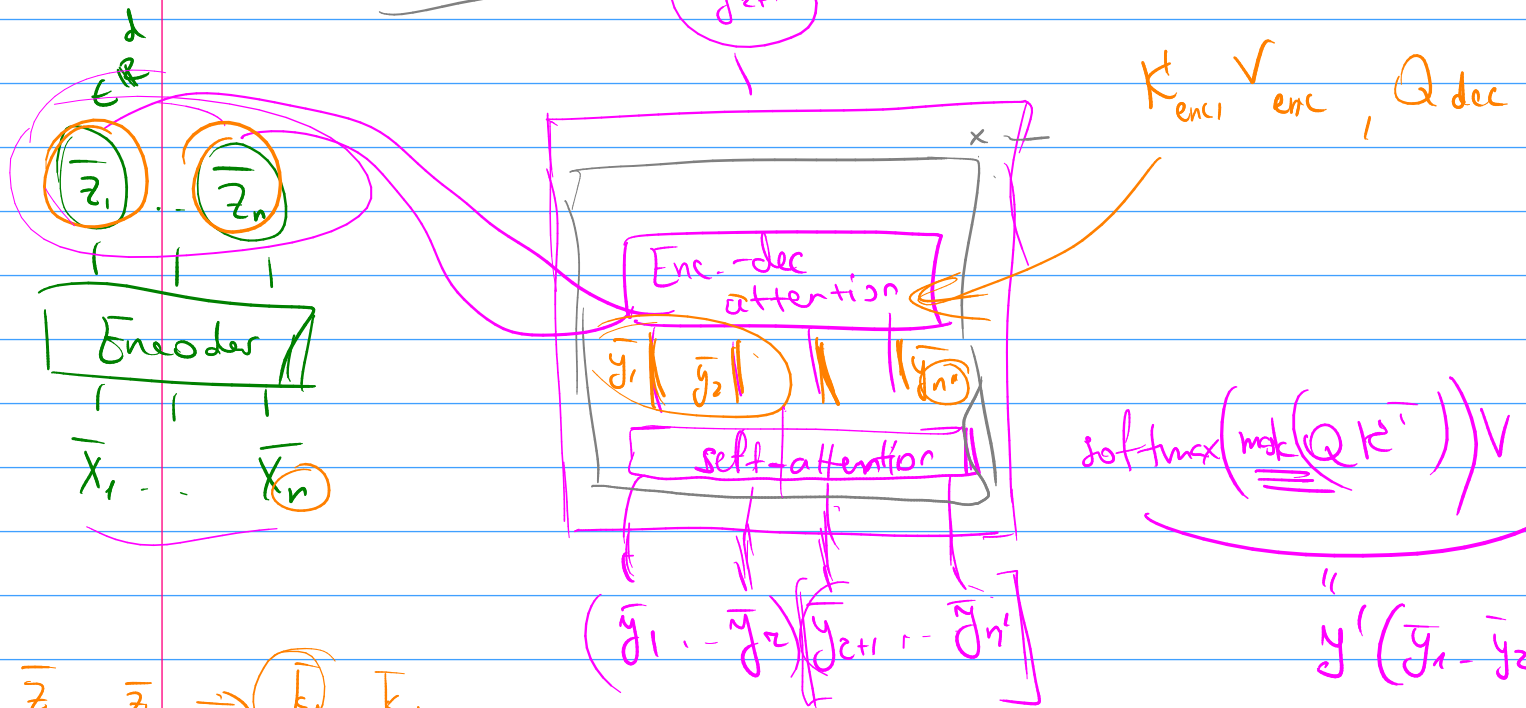


$XW_0^q = Q_0, XW_0^k = K_0, XW_0^v = V_0$

$X \rightarrow XW_T^q = Q_T, XW_T^v = V_T$

Beca:  $W_s^q, W_s^k, W_s^v, s=0, \dots, T-1, W^0$   
 $d \times m, d \times m, d \times l, sL \times d$

$d \times (2m+l) \times T$



$\bar{z}_1, \dots, \bar{z}_n \rightarrow \begin{pmatrix} \bar{k}_1 \\ \vdots \\ \bar{k}_n \end{pmatrix}, \begin{pmatrix} \bar{v}_1 \\ \vdots \\ \bar{v}_n \end{pmatrix}$   
 $\bar{y}_1, \dots, \bar{y}_n \rightarrow \begin{pmatrix} \bar{q}_1 \\ \vdots \\ \bar{q}_n \end{pmatrix}, \dots, \bar{q}_n$

$\text{softmax}(\frac{1}{\sqrt{e}} \text{mask}(QK^T))V - n \times l$   
 $n \times m, m \times n, n \times l$   
 $n \times n$

→ 0  
→ 0  
→ 0

$C_{12}$   $C_{13}$   $C_{14}$   $C_{15}$   $C_{16}$   $C_{17}$   $C_{18}$   $C_{19}$   $C_{20}$   $C_{21}$   $C_{22}$   $C_{23}$   $C_{24}$   $C_{25}$   $C_{26}$   $C_{27}$   $C_{28}$   $C_{29}$   $C_{30}$   $C_{31}$   $C_{32}$   $C_{33}$   $C_{34}$   $C_{35}$   $C_{36}$   $C_{37}$   $C_{38}$   $C_{39}$   $C_{40}$   $C_{41}$   $C_{42}$   $C_{43}$   $C_{44}$   $C_{45}$   $C_{46}$   $C_{47}$   $C_{48}$   $C_{49}$   $C_{50}$   $C_{51}$   $C_{52}$   $C_{53}$   $C_{54}$   $C_{55}$   $C_{56}$   $C_{57}$   $C_{58}$   $C_{59}$   $C_{60}$   $C_{61}$   $C_{62}$   $C_{63}$   $C_{64}$   $C_{65}$   $C_{66}$   $C_{67}$   $C_{68}$   $C_{69}$   $C_{70}$   $C_{71}$   $C_{72}$   $C_{73}$   $C_{74}$   $C_{75}$   $C_{76}$   $C_{77}$   $C_{78}$   $C_{79}$   $C_{80}$   $C_{81}$   $C_{82}$   $C_{83}$   $C_{84}$   $C_{85}$   $C_{86}$   $C_{87}$   $C_{88}$   $C_{89}$   $C_{90}$   $C_{91}$   $C_{92}$   $C_{93}$   $C_{94}$   $C_{95}$   $C_{96}$   $C_{97}$   $C_{98}$   $C_{99}$   $C_{100}$

n