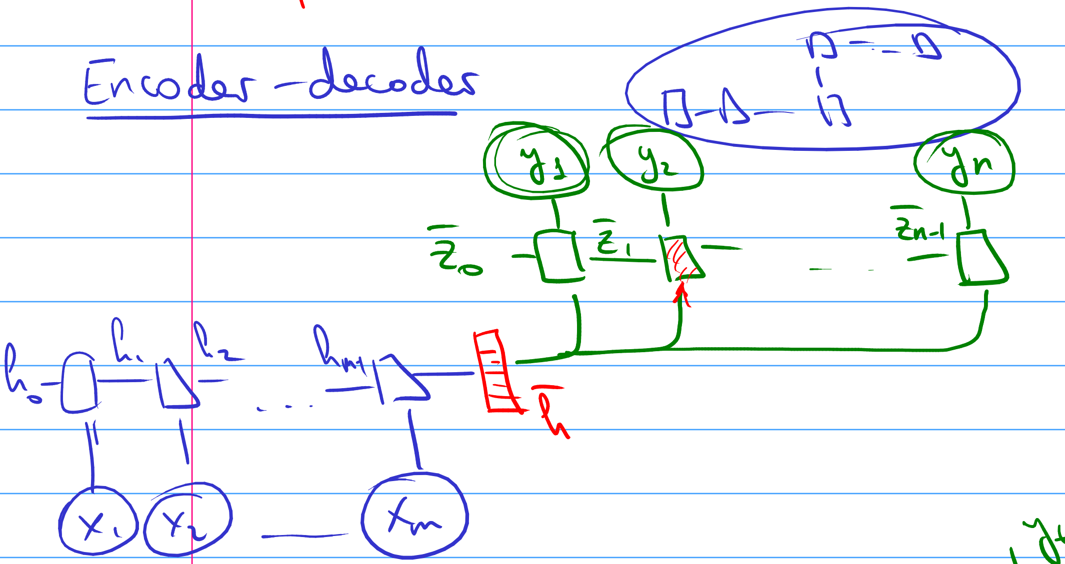
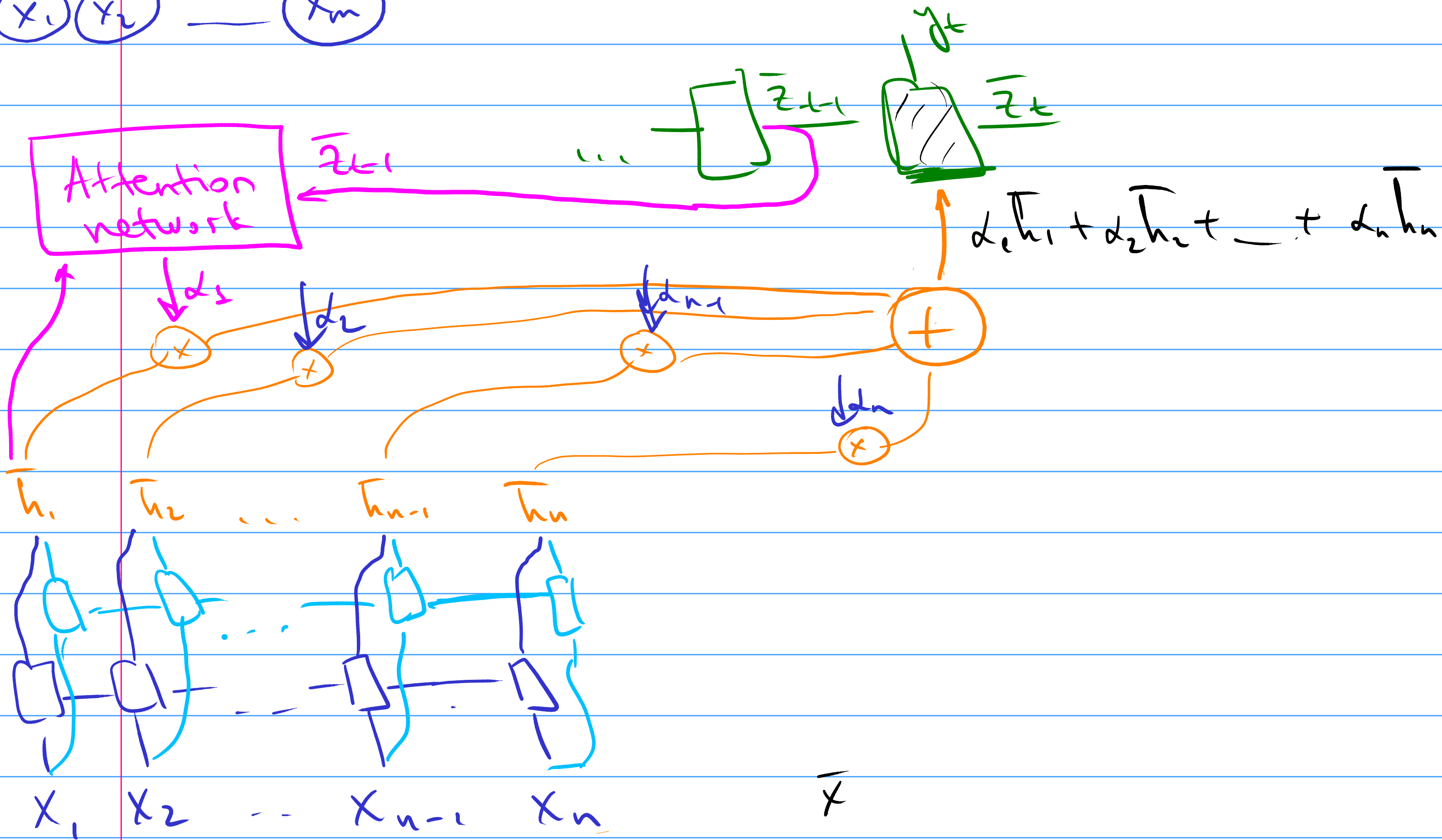
Лисица, преслед. зайчик и
и медведем, бегит по
по лесу - — -

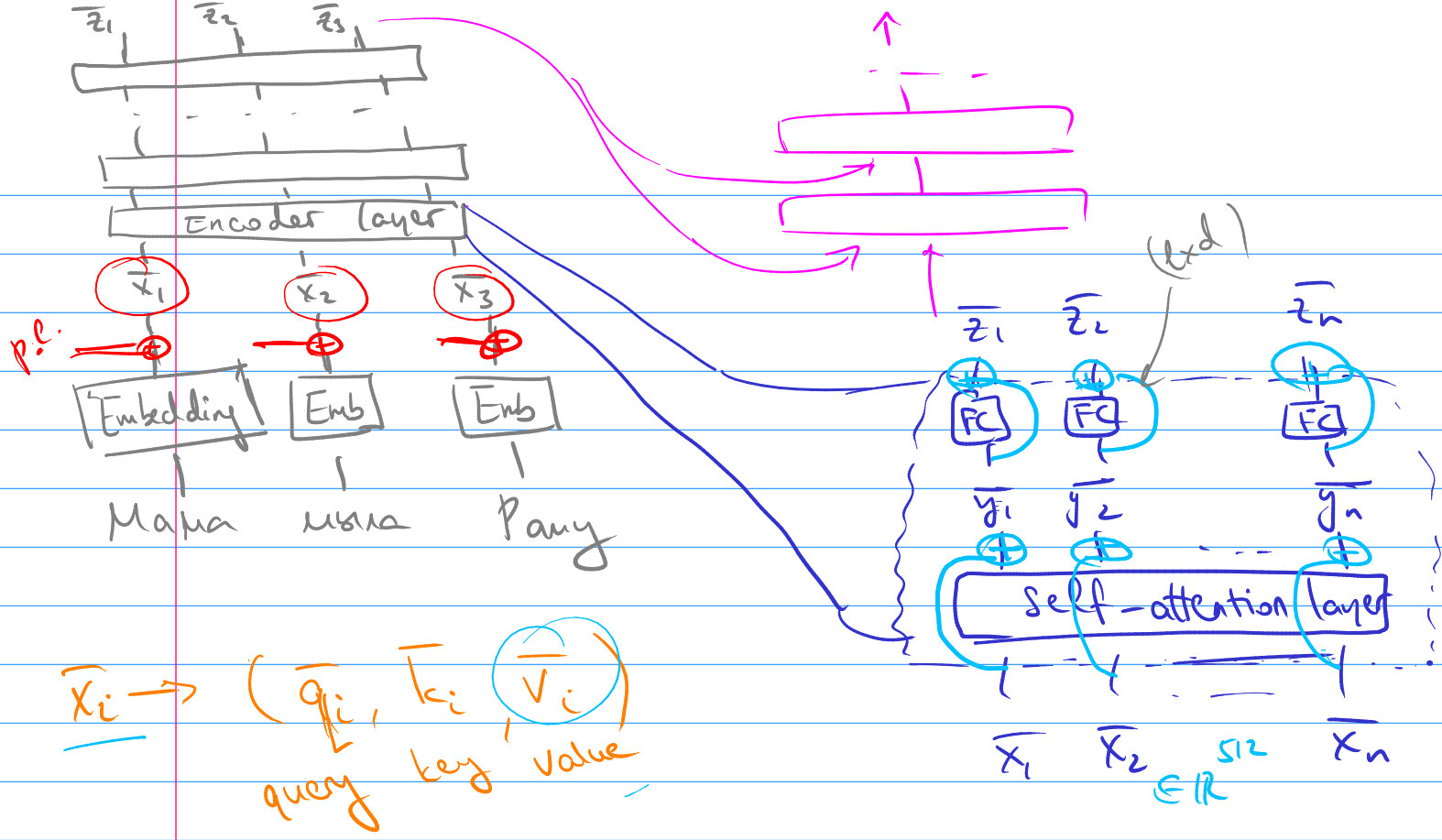
$$p(y_1, \dots, y_n \mid x_1, \dots, x_m) =$$
$$= p(y_1 \mid \bar{x})\, p(y_2 \mid y_1, \bar{x})\, p(y_3 \mid y_1, y_2, \bar{x}) \dots p(y_n \mid \bar{y}_{1..n-1}, \bar{x})$$

Encoder-decoder



$\bar{y}$





$\bar{x}_i \to (\bar{q}_i, \bar{k}_i, \bar{v}_i)$ — query, key, value

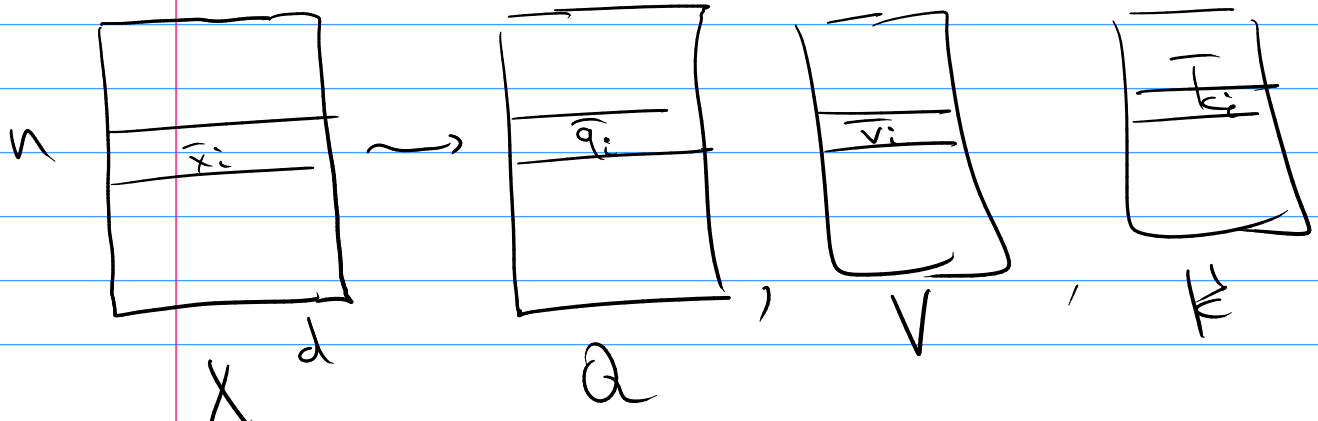
$a_{ij} \approx (\bar{q}_i^T \bar{k}_j)$ → насколько сильно "$\bar{x}_i$ хочет узнать $\bar{x}_j$" — "relevance"

$\bar{x}_1 \to (\bar{q}_1, \bar{k}_1, \bar{v}_1) \to (\bar{q}_i^T \bar{k}_1)$

$\bar{x}_2 \to (\bar{q}_2, \bar{k}_2, \bar{v}_2) \to (\bar{q}_i^T \bar{k}_2)$

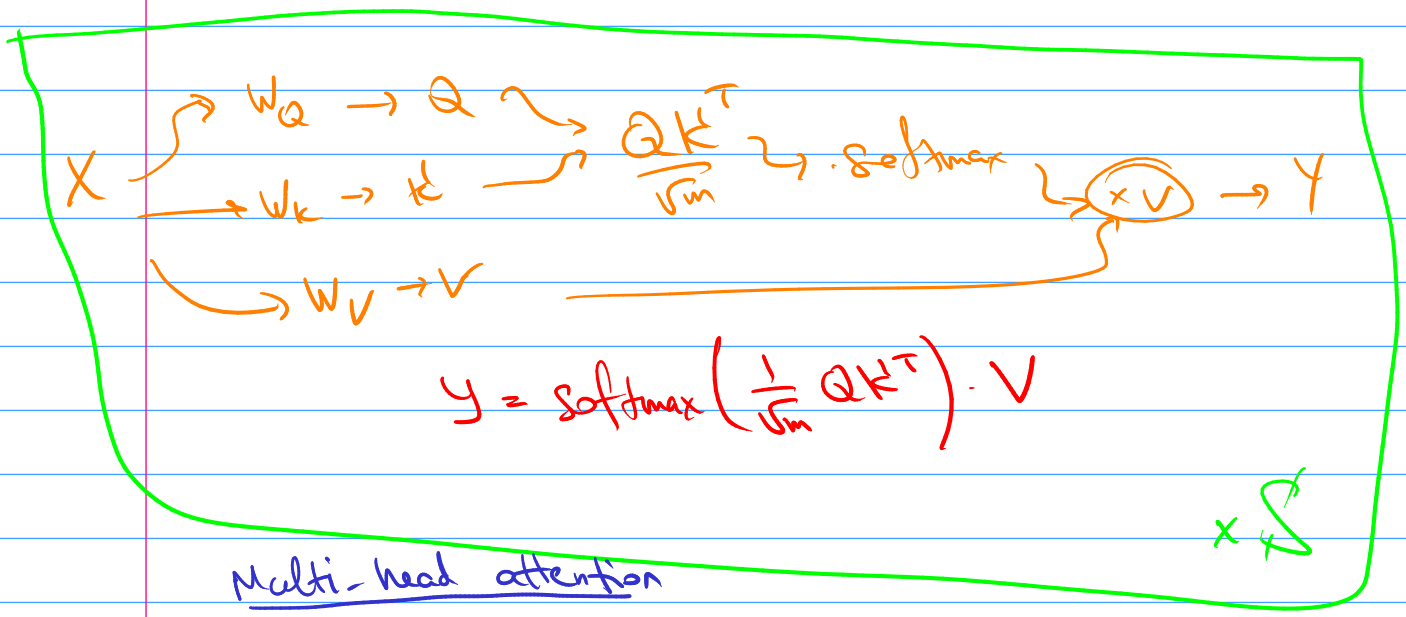
$\bar{x}_n \to (\bar{q}_n, \bar{k}_n, \bar{v}_n) \to (\bar{q}_i^T \bar{k}_n)$

$\bar{x}_i \to (\bar{q}_i, \bar{k}_i, \bar{v}_i) \to \bar{y}_i = \sum_{j=1}^{n} \text{softmax}\left(\frac{1}{\sqrt{m}} \bar{q}_i^T \bar{k}_j\right) \bar{v}_j$



W_Q : $q_i = W_Q x_i$, $Q = W_Q X$

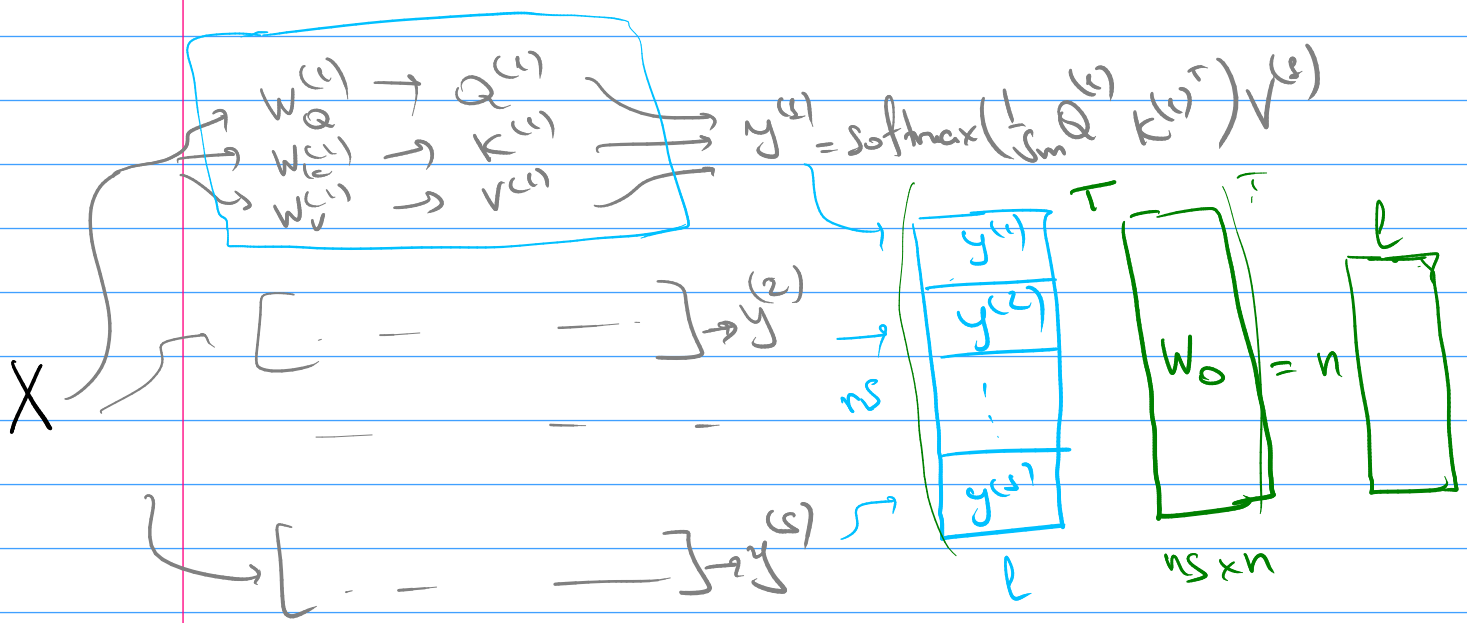
W_K : $k = W_K X$ $\quad$ W_V : $V = W_V X$



$$Y = \text{softmax}\left(\frac{1}{\sqrt{d_m}} QK^T\right) \cdot V$$

Multi-head attention

$$\left(W_Q^{(1)}, W_K^{(1)}, W_V^{(1)}\right), \ldots, \left(W_Q^{(8)}, W_K^{(8)}, W_V^{(8)}\right) \ldots$$



W_O — $n_s \times n$

W_Q, W_K — $(d \times m) s$

W_V — $(d \times l) s$

ns × l

$s(n^2 + d_m + d_l)$ heads

s = 8, d_m = 512, d_l = 64, 64

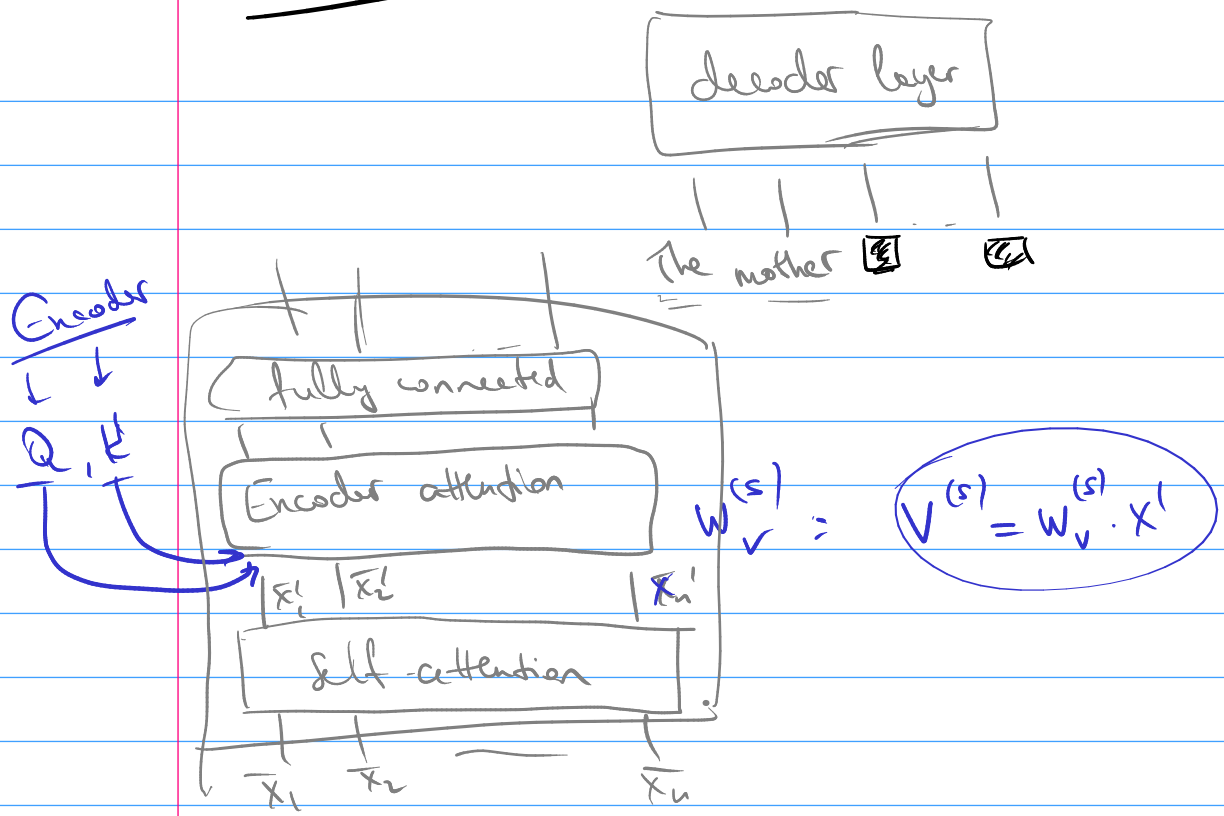
positional embedding → + Emb

MAMA

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

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

Decoder



$W_v^{(s)}$: $V^{(s)} = W_v^{(s)} \cdot X^1$

Transformer

Enc + Dec

- GPT - Dec — language models
- BERT — Enc