X5 Tech

SEQ2SEQ
И ATTENTION

Андреева Дарья
Data Scientist, X5 Tech

машинный перевод

1954г. – джорджтаунский эксперимент

до 2010г. – статистические модели

с 2016г. – машинный перевод

X5 Tech

формально

source: $x = (x_1, \ldots, x_n)$

target: $y = (y_1, \ldots, y_m)$

C – корпус из пар (x, y)

$$P_\theta(Y|X) = \prod_{i=1}^{n} p(y_i|X, y_{<i}) \longrightarrow \max$$

$$\hat{y}_i = argmax(p_\theta(y_i|X, \hat{y}_{<i}))$$

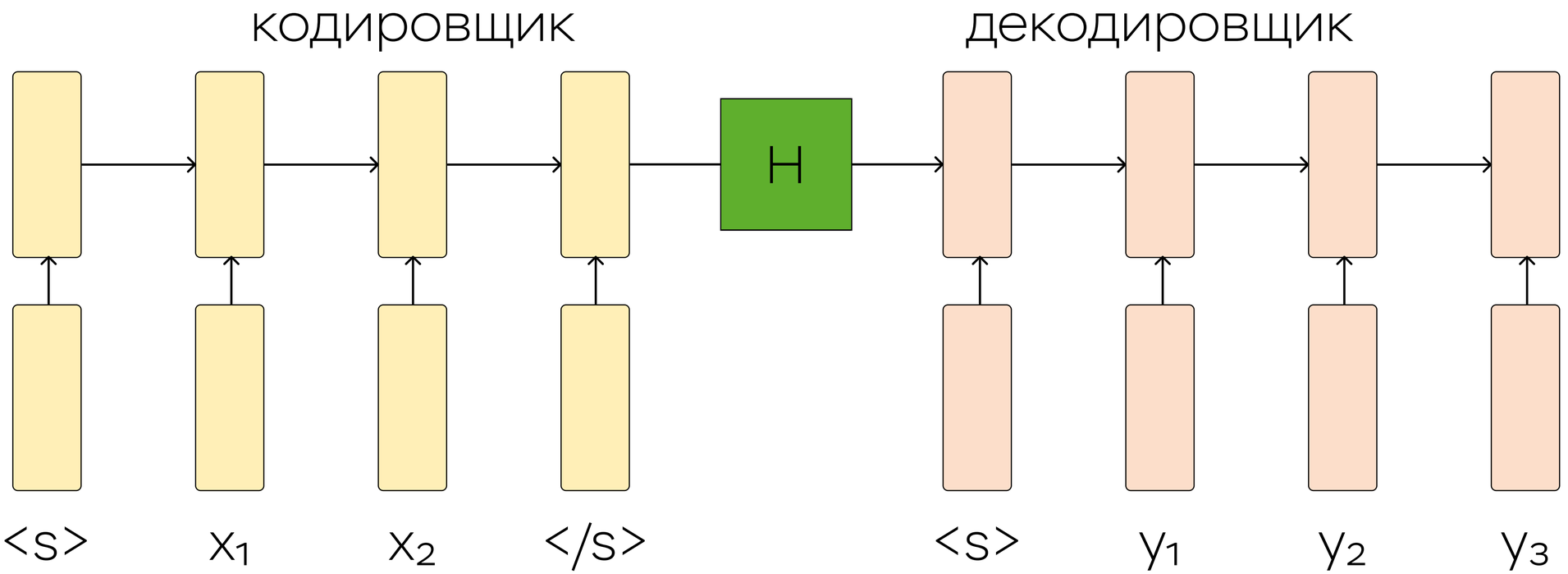
жадно семплируем?

кодировщик-декодировщик

source: $x = (x_1, \ldots, x_n)$

target: $y = (y_1, \ldots, y_m)$

C – корпус из пар (x, y)

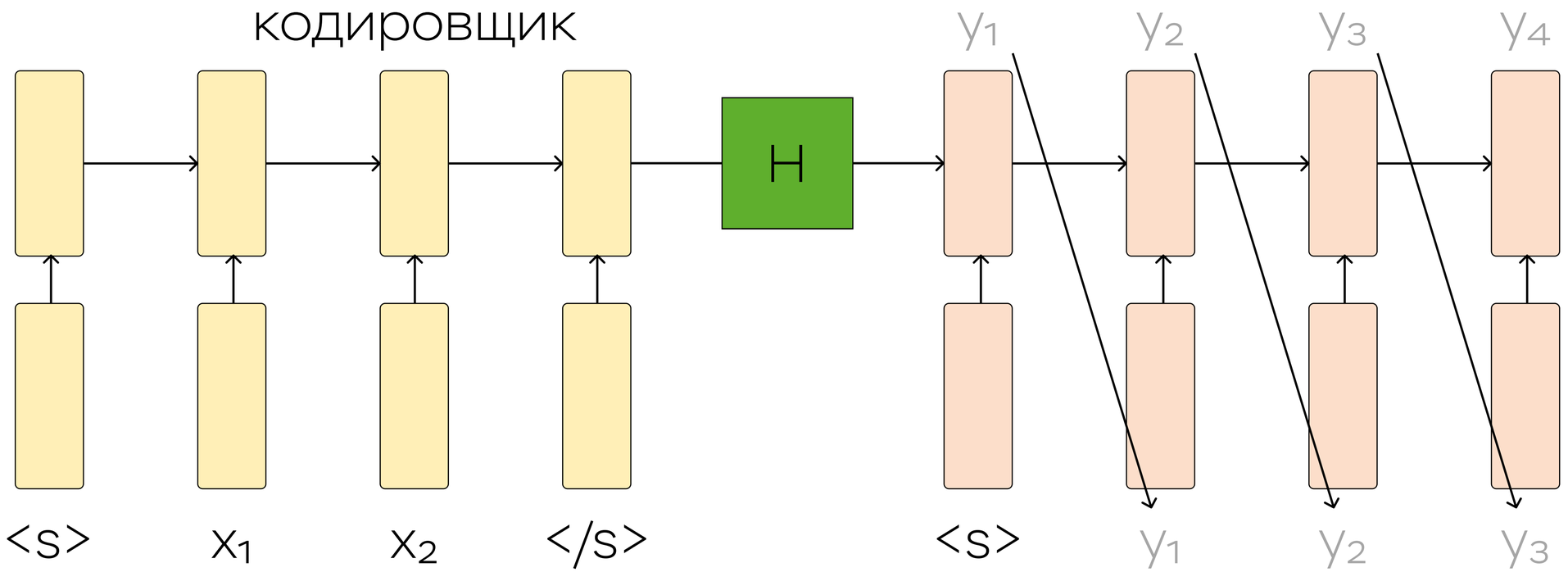


кодировщик-декодировщик

source: $x = (x_1, \ldots, x_n)$

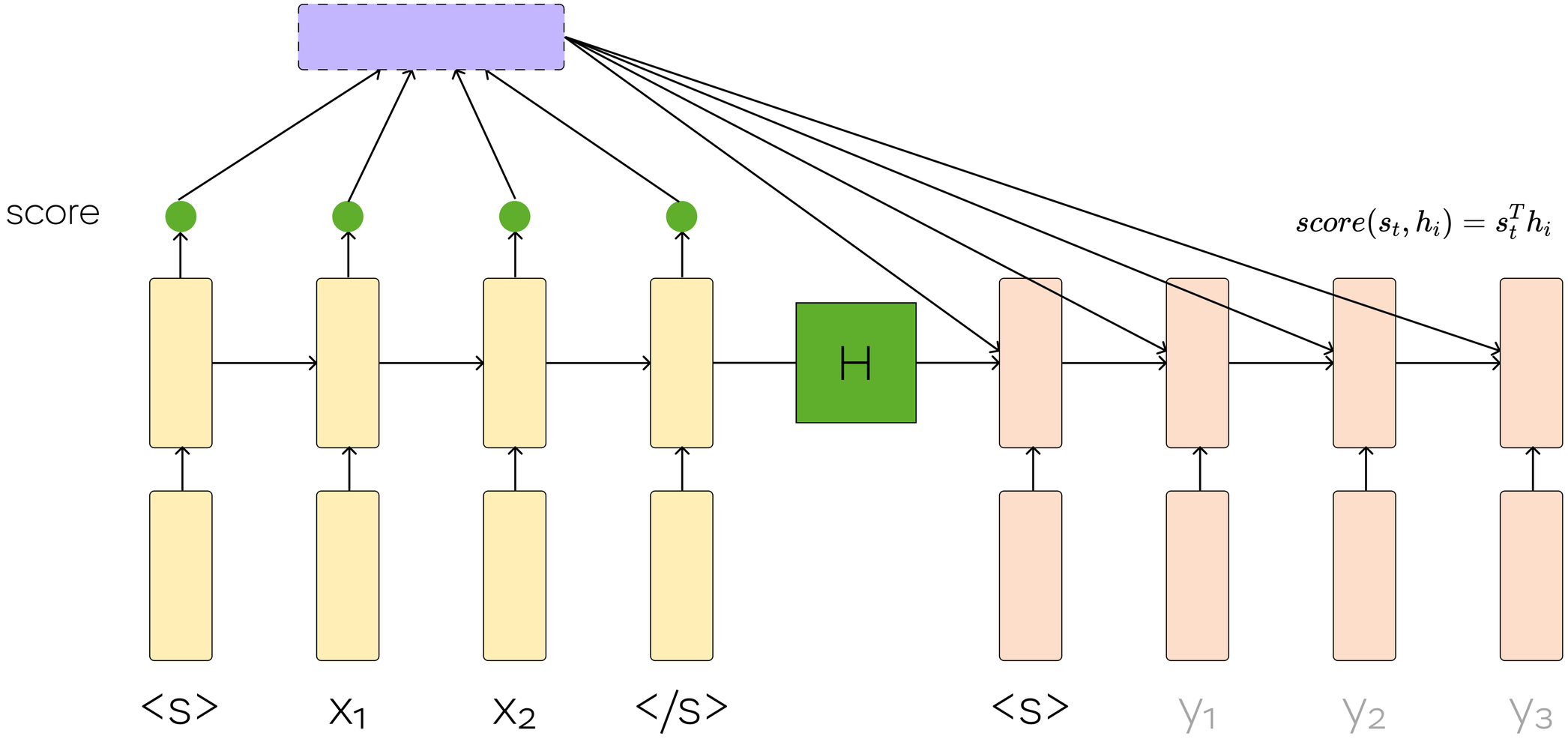
target: $y = (y_1, \ldots, y_m)$

C – корпус из пар (x, y)



X5 Tech

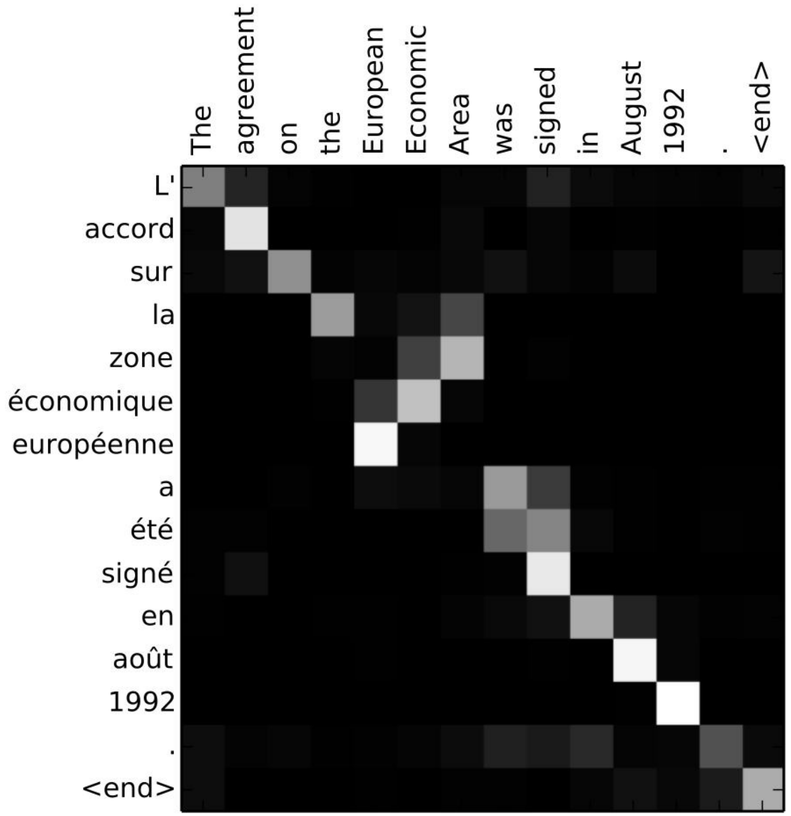
внимание



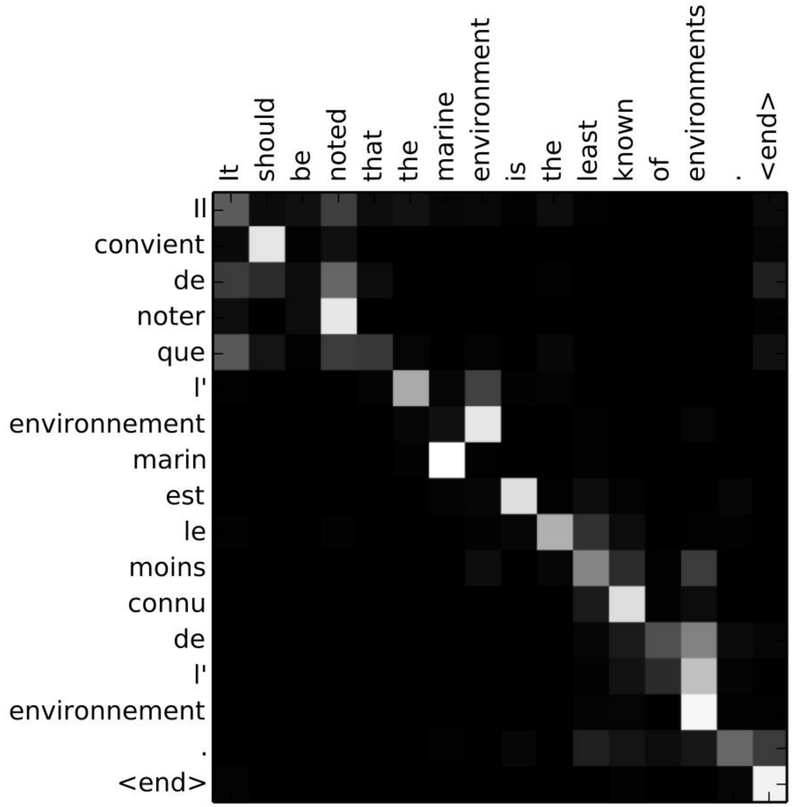
X5 Tech

Neural Machine Translation by Jointly Learning to Align and Translate: https://arxiv.org/abs/1409.0473

внимание



(a)



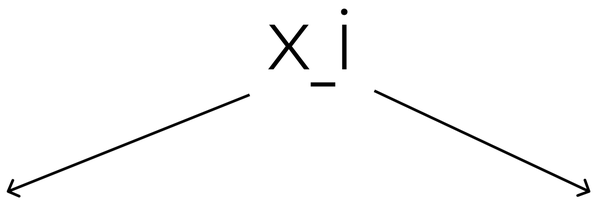
(b)

X5 Tech

ATTENTION IS ALL YOU NEED

Attention is all you need: https://arxiv.org/abs/1706.03762

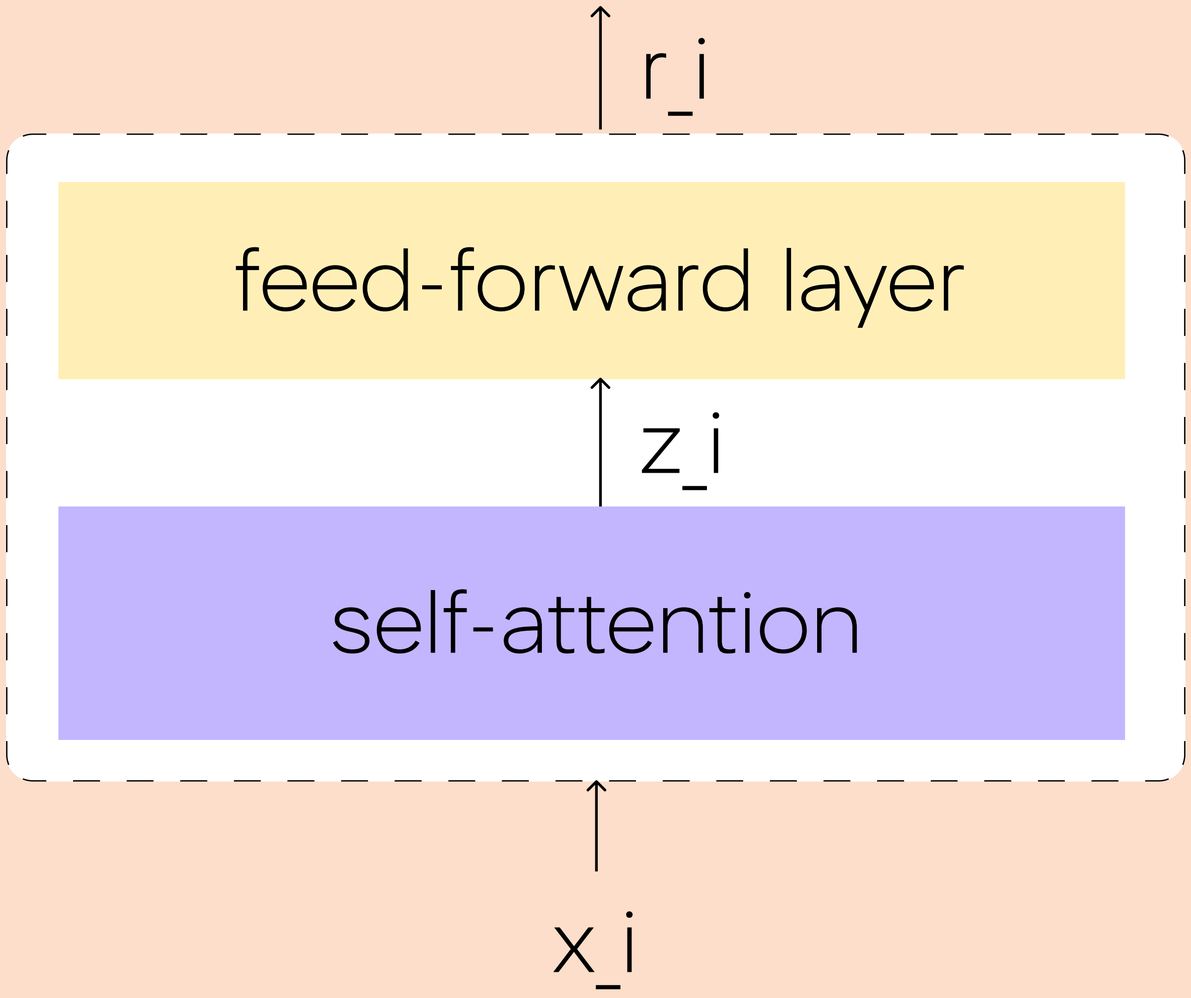
encoder



embedding

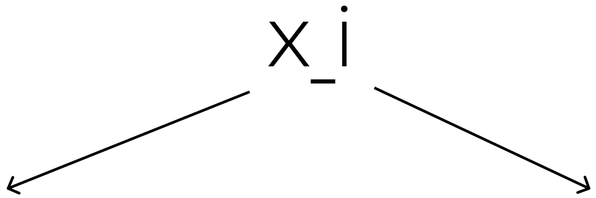
$q_i : W_Q * x\_i$
$k_i : W_K * x\_i$
$v_i : W_V * x\_i$

attention:



X5 Tech

encoder



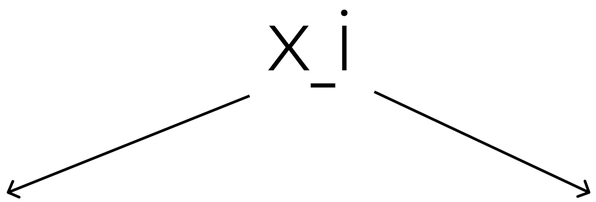
embedding

$q_i : W_Q * x\_i$
$k_i : W_K * x\_i$
$v_i : W_V * x\_i$

attention:

X5 Tech

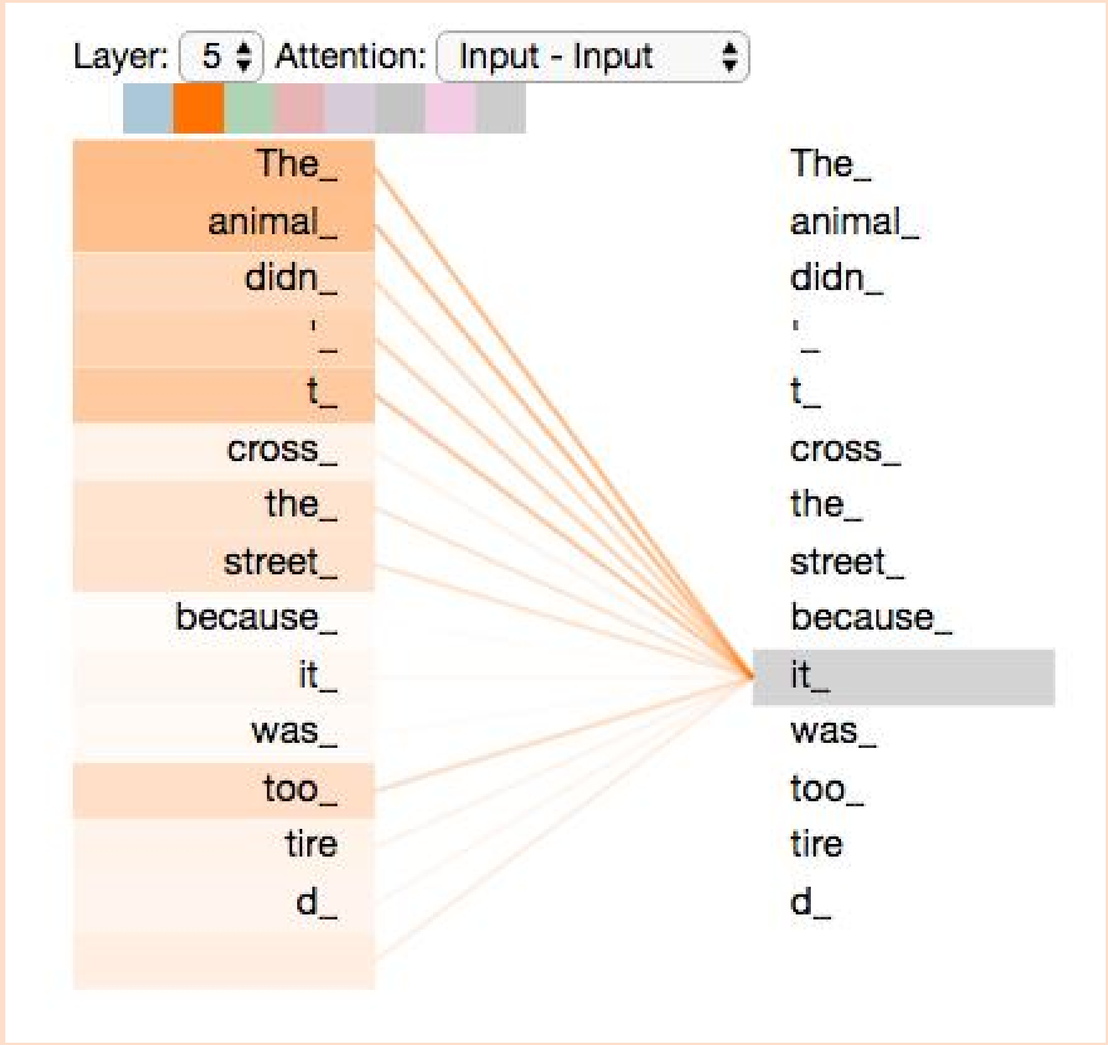
encoder



embedding

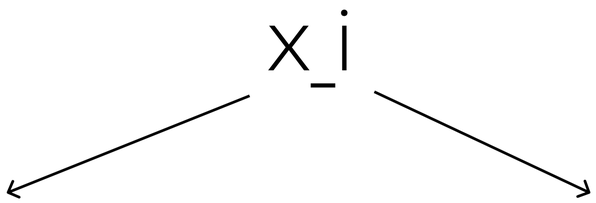
$q_i : W_Q * x\_i$
$k_i : W_K * x\_i$
$v_i : W_V * x\_i$

attention:



X5 Tech

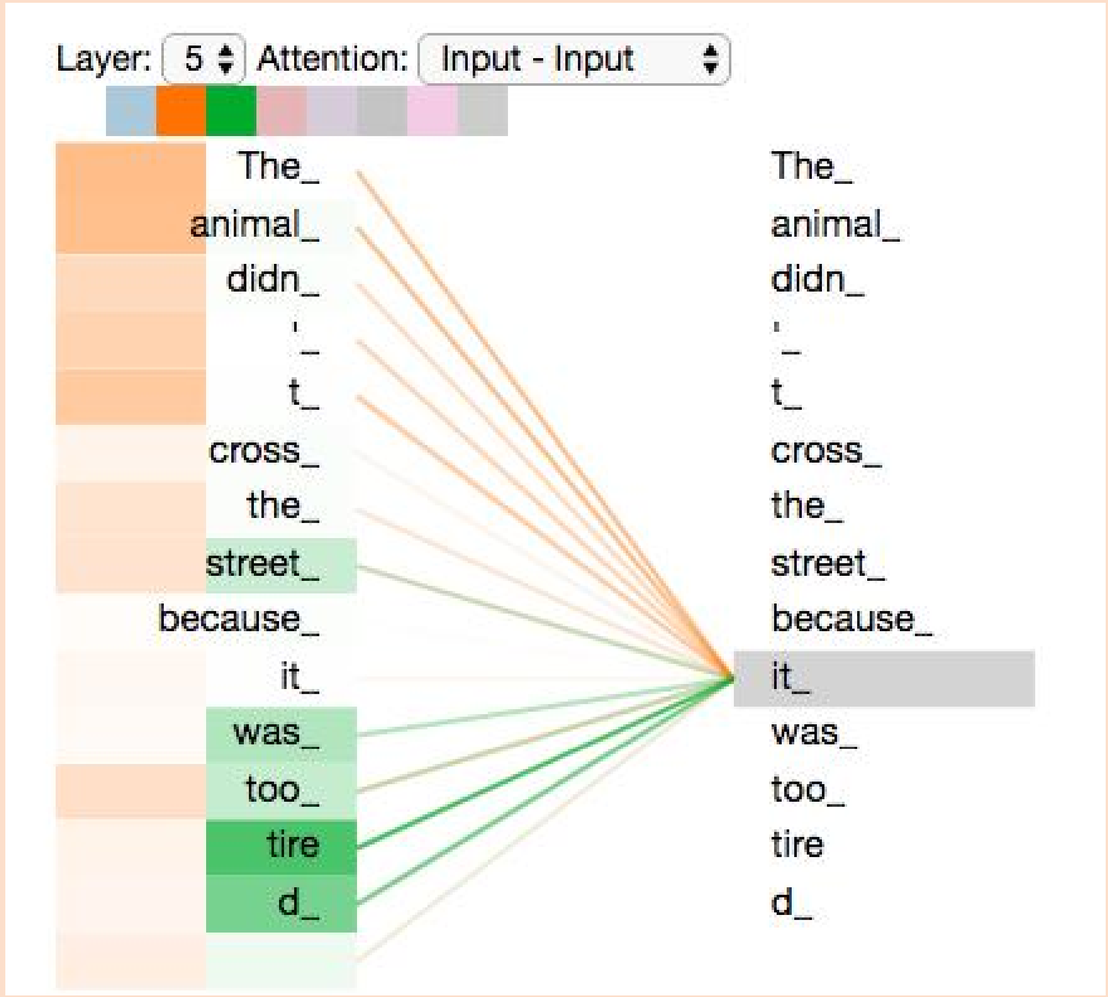
encoder



embedding

$q_i : W_Q * x\_i$
$k_i : W_K * x\_i$
$v_i : W_V * x\_i$

attention:



X5 Tech

encoder

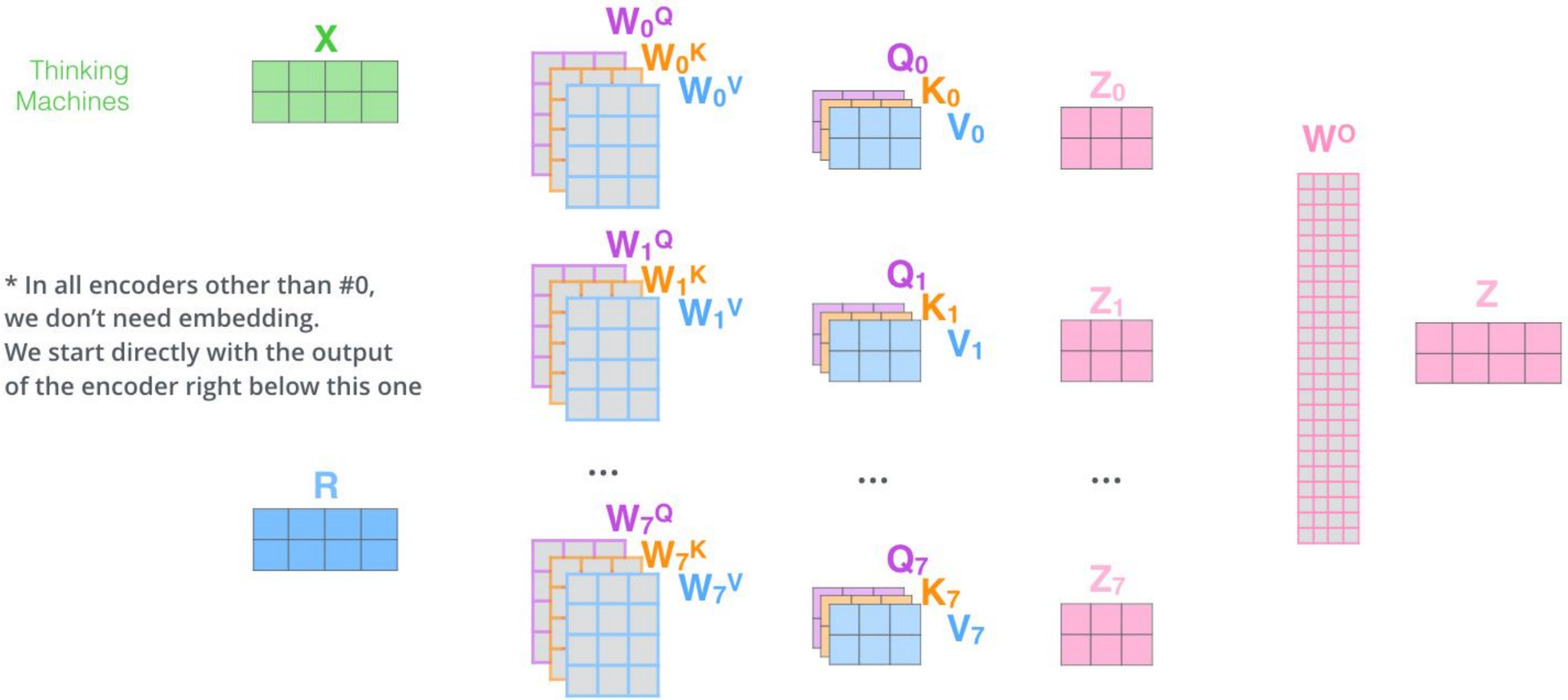
1) This is our input sentence*

2) We embed each word*

3) Split into 8 heads. We multiply X or R with weight matrices

4) Calculate attention using the resulting Q/K/V matrices

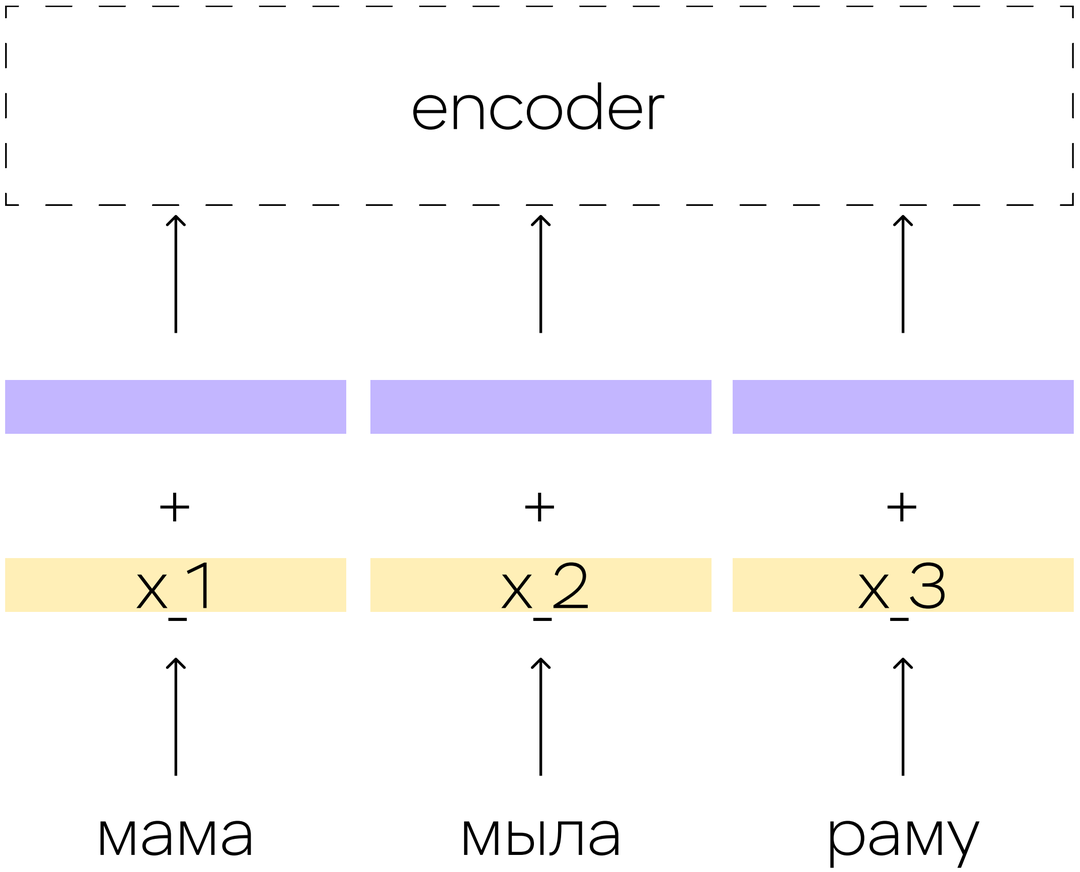
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

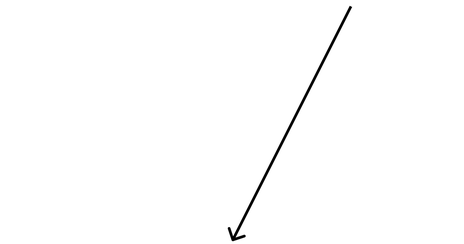


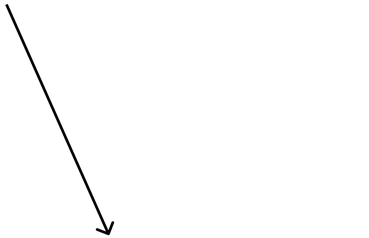
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

X5 Tech

позиционный эмбеддинг

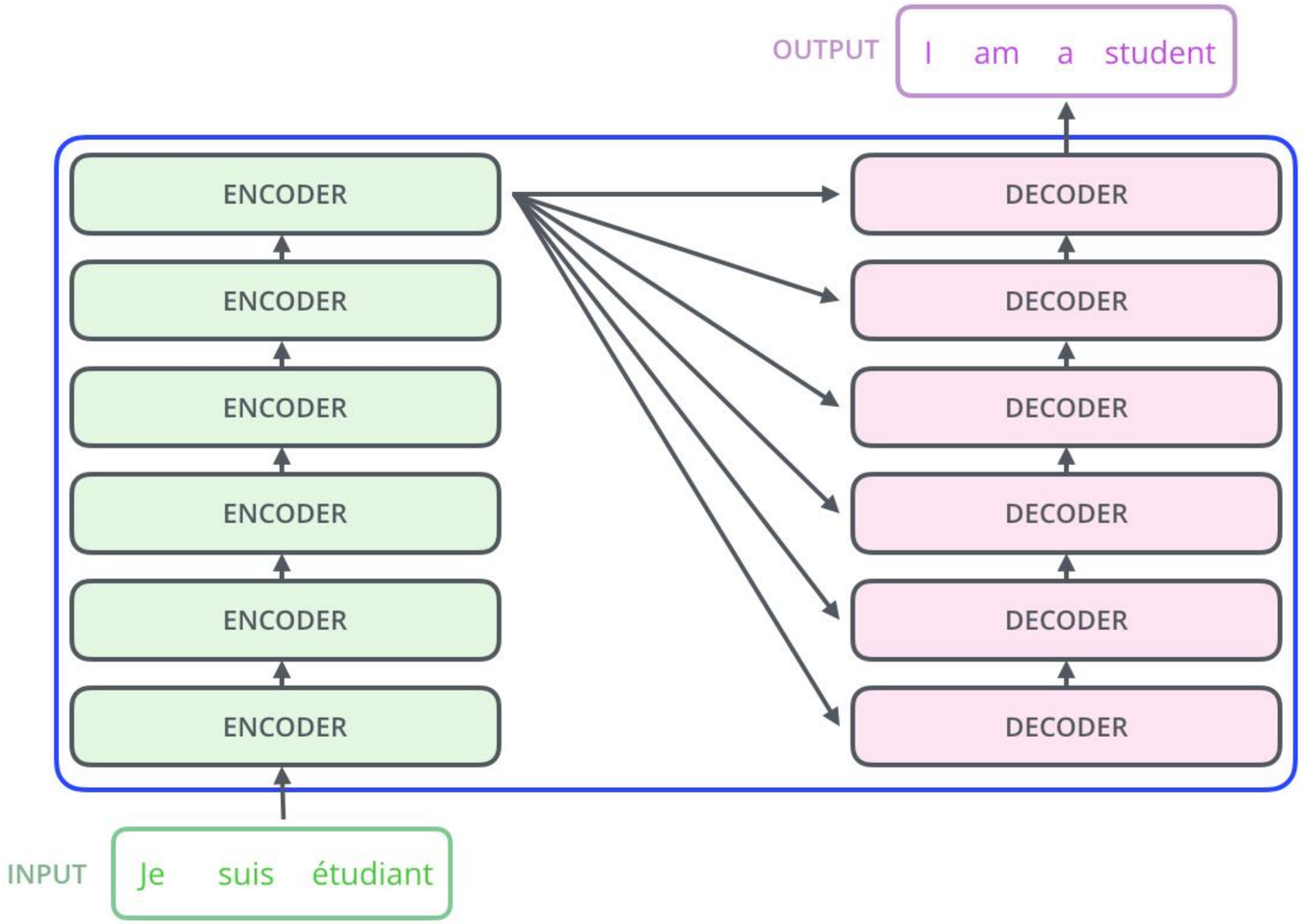






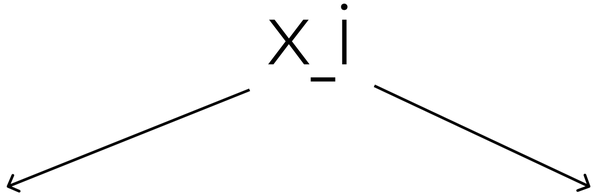
X5 Tech

overview



X5 Tech

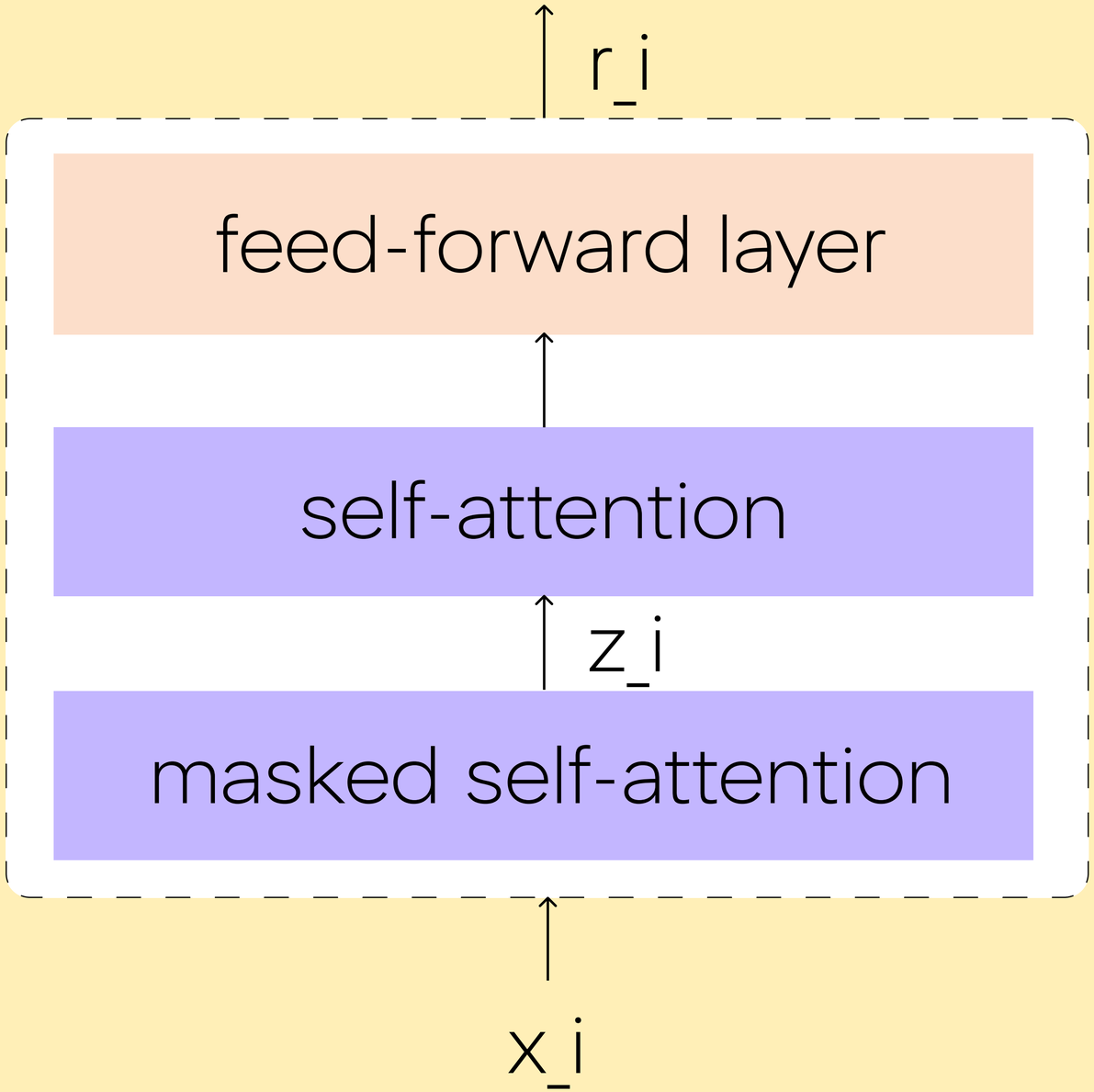
decoder



embedding

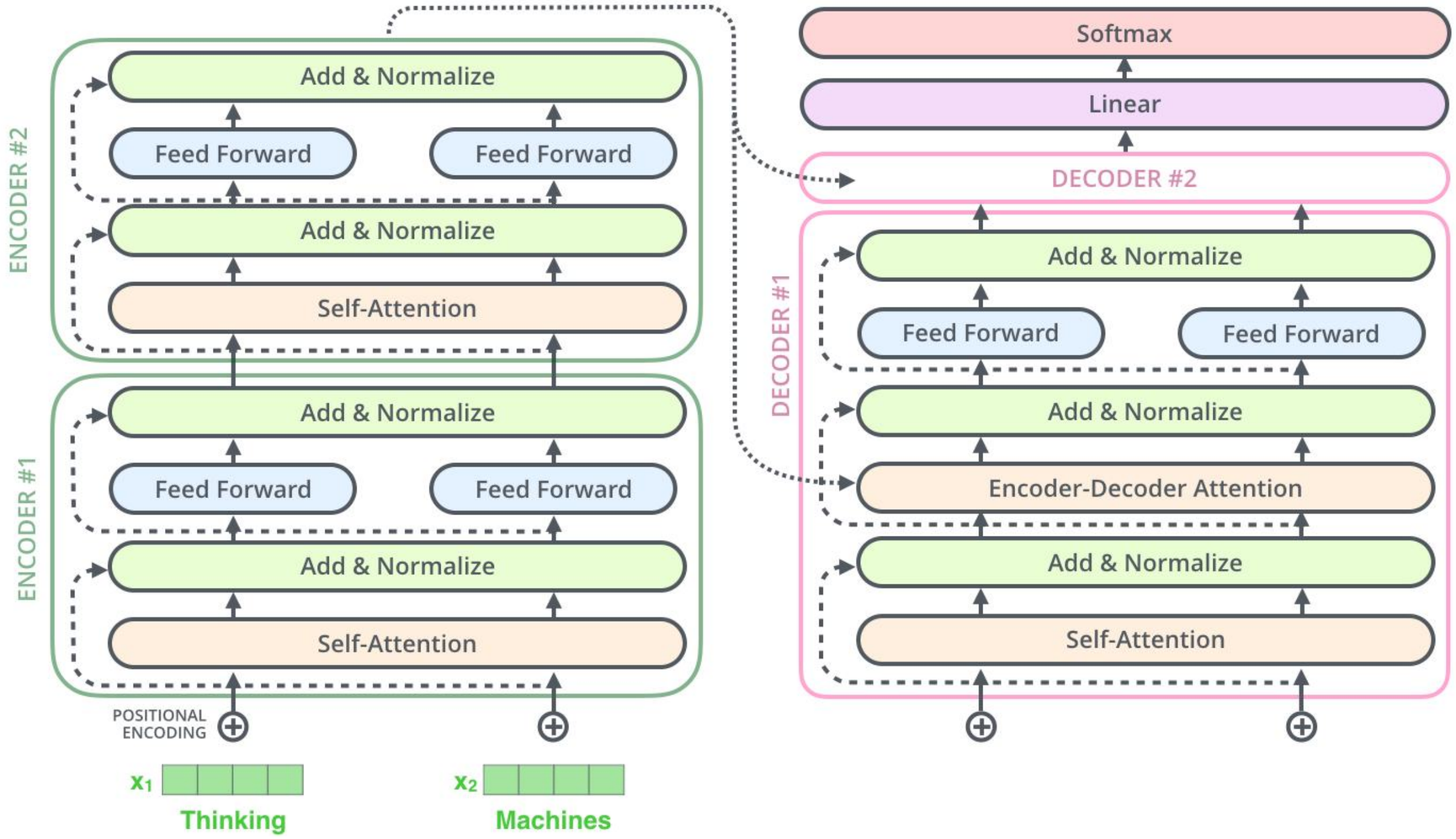
$q_i : W_Q * x\_i$
$k_i : W_K * x\_i$
$v_i : W_V * x\_i$

attention:



X5 Tech

overview



X5 Tech