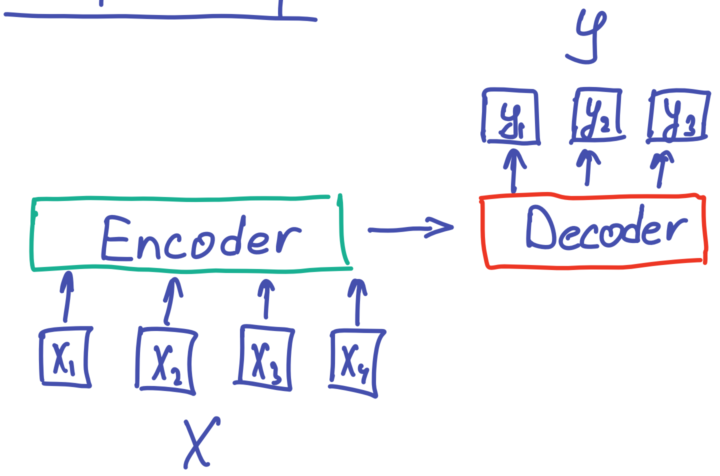
Seq 2 Seq



X - input sequence

$$P_\theta(\underset{\uparrow}{y}|X) = \prod_{i=1}^{n} P_\theta(y_i|X, y_{<i}) \to \max_\theta$$

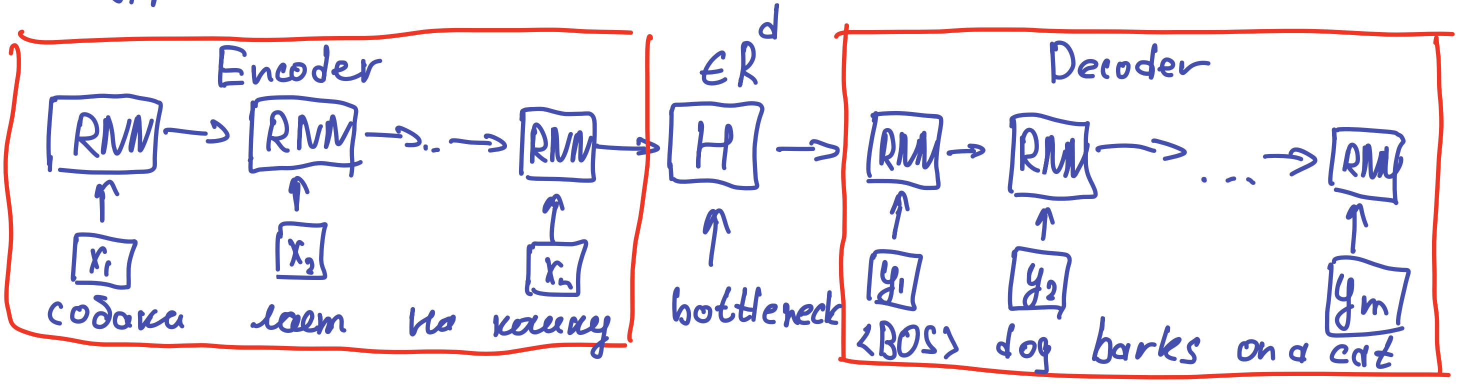
output sequence

$$\hat{y}_i = \arg\max_{y_i} P_\theta(y_i|X, \hat{y}_{<i})$$ — Жадная генерация

$$P_\theta(\hat{y}|X) = \prod_{i=1}^{n} \max_{y_i} P_\theta(y_i|X, \hat{y}_{<i}) \neq \max_{y} \prod_{i=1}^{n} P(y_i|X, y_{<i})$$

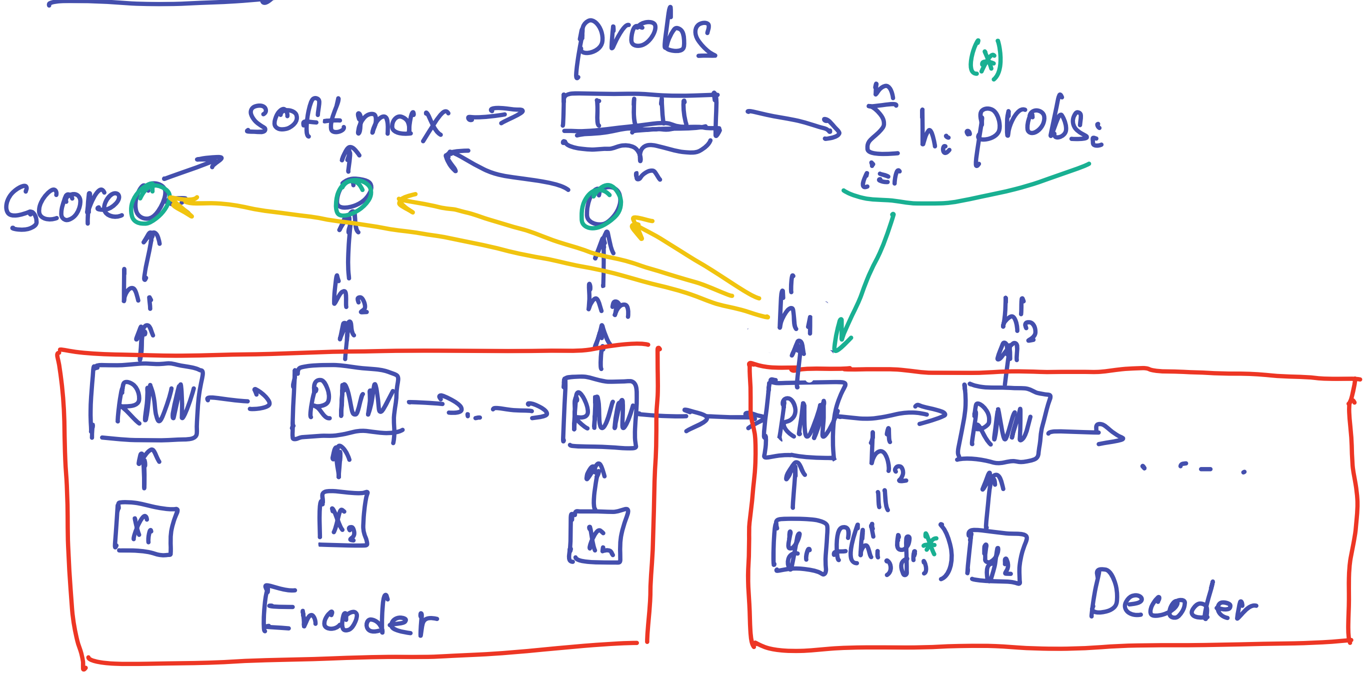
Для генерации лучше Beam Search

Рекуррентные сети не позволяют извлечь всю информацию



Обучение: Cross-Entropy

Механизм внимания. (2014)

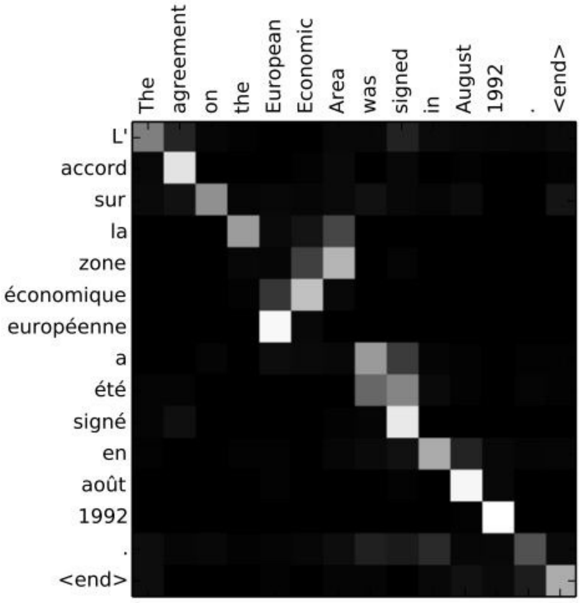


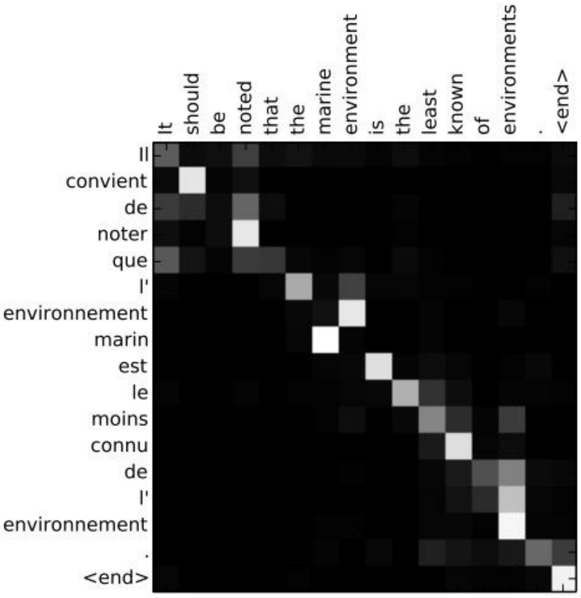
Score functions:

1) Dot-product: $\underset{\in \mathbb{R}^d}{s(h}, \underset{\in \mathbb{R}^d}{h')} = h^T h'$

2) Bilinear: $s(\underset{\mathbb{R}^d}{h}, \underset{\mathbb{R}^{d'}}{h'}) = \underline{h^T W_s h'}$

3) MLP: $s(h, h') = \vec{w}_{s'}^{\,T} \tanh\left(W_s \begin{bmatrix} h^T \\ h^T \end{bmatrix}\right)$



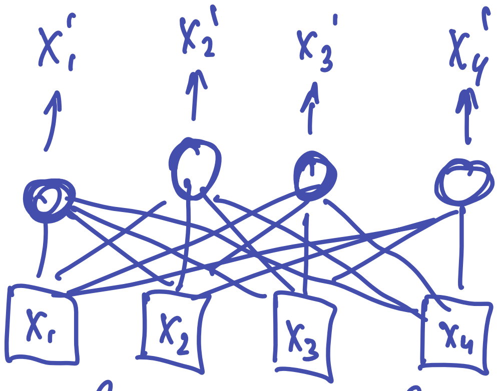


Learned Attention Matrices

Self-attention

$$\begin{cases} q_i = W_Q x_i \\ k_i = W_k x_i \\ v_i = W_v x_i \end{cases}$$

Encoder



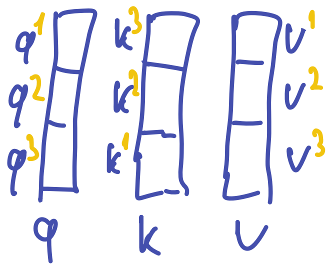
На двери замка висит замок

pos embed. (0 + 1 + 2 + 3 + 4)

$$\boxed{\begin{aligned} score_{ij} &= \frac{q_i^T k_j}{\sqrt{d}} \\ attention_i &= softmax(score_i) \\ x_i' &= \sum_{j=1}^{n} \underset{\substack{\uparrow \\ scalar}}{attention_{ij}} v_j \end{aligned}}$$

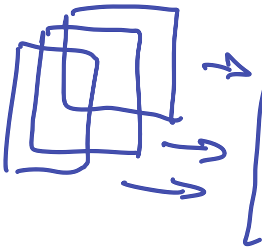
Multi-head Attention

Идея: разобьем q, k, v на n_{heads} векторов. Каждая голова будет извлекать свою информацию. Потом склеим все выходы вместе.



$$attention_i^s = softmax\left(\frac{q_i^{s^T} k_j^s}{\sqrt{d}}\right)$$

$$x_i^s = \sum_{j=1}^{n} attention_i^s v_j^s$$



$$\begin{bmatrix} x_i^{1'} \\ x_i^{2'} \\ x_i^{3'} \end{bmatrix} = x_i'$$

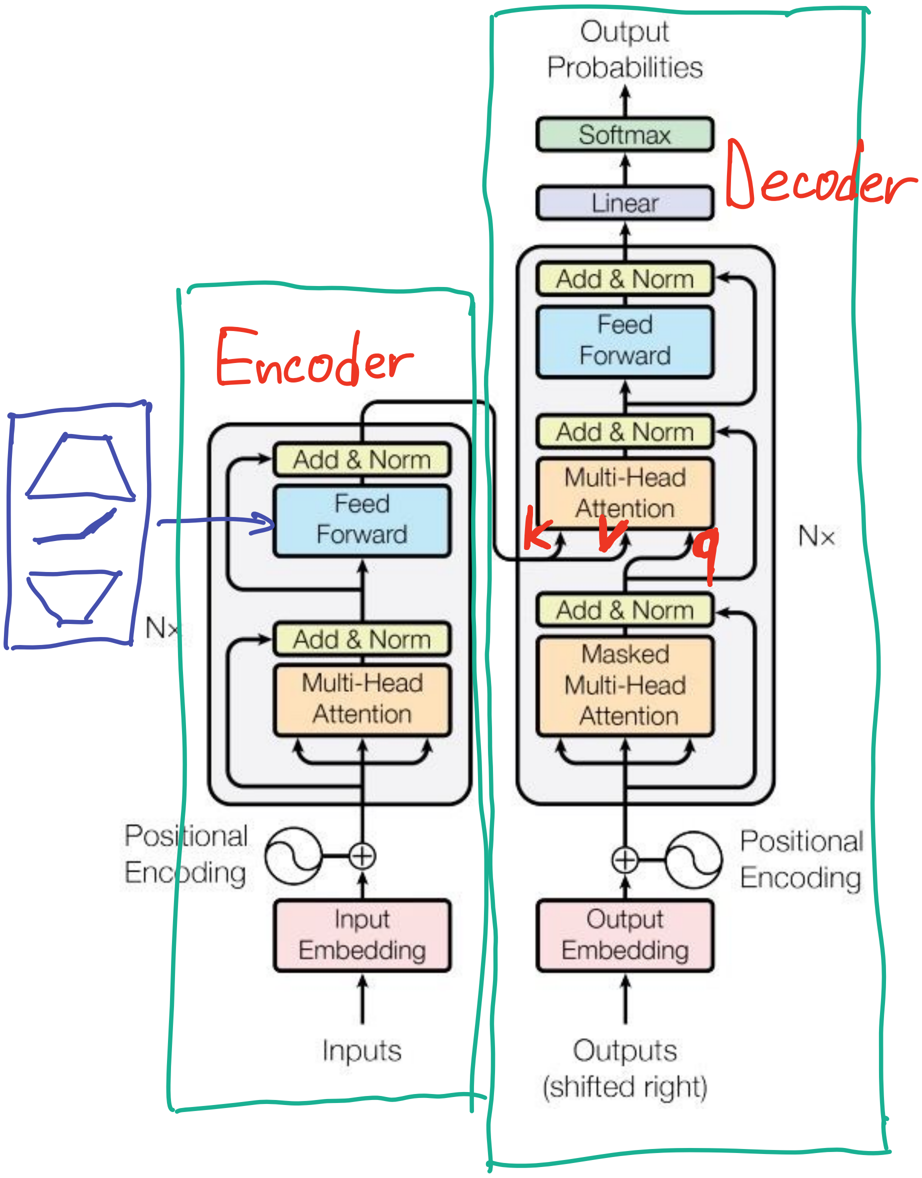
Masked Self-Attention (Decoder)

$$attention_{ij} = softmax\left(\frac{q_i^T k_j}{\sqrt{d}} + \underline{mask_{ij}}\right)$$

$$mask_{ij} = \begin{cases} 0, & i \le j \\ -\infty, & i > j \end{cases}$$

$$\Rightarrow \forall i > j : \; attention_{ij} = 0$$

Зануляем внимание так, чтобы декодер при обучении не мог смотреть вперед


Output Probabilities
Softmax
Linear
Decoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
k
v
q
Nx
Add & Norm
Masked Multi-Head Attention
Positional Encoding
Output Embedding
Outputs (shifted right)
Encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs