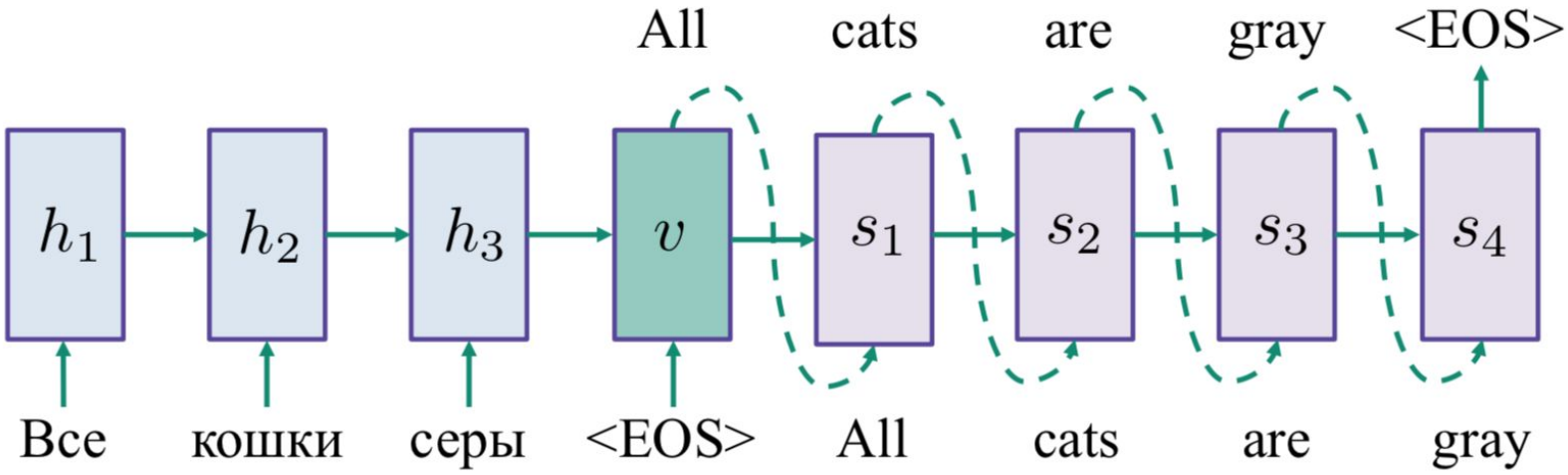
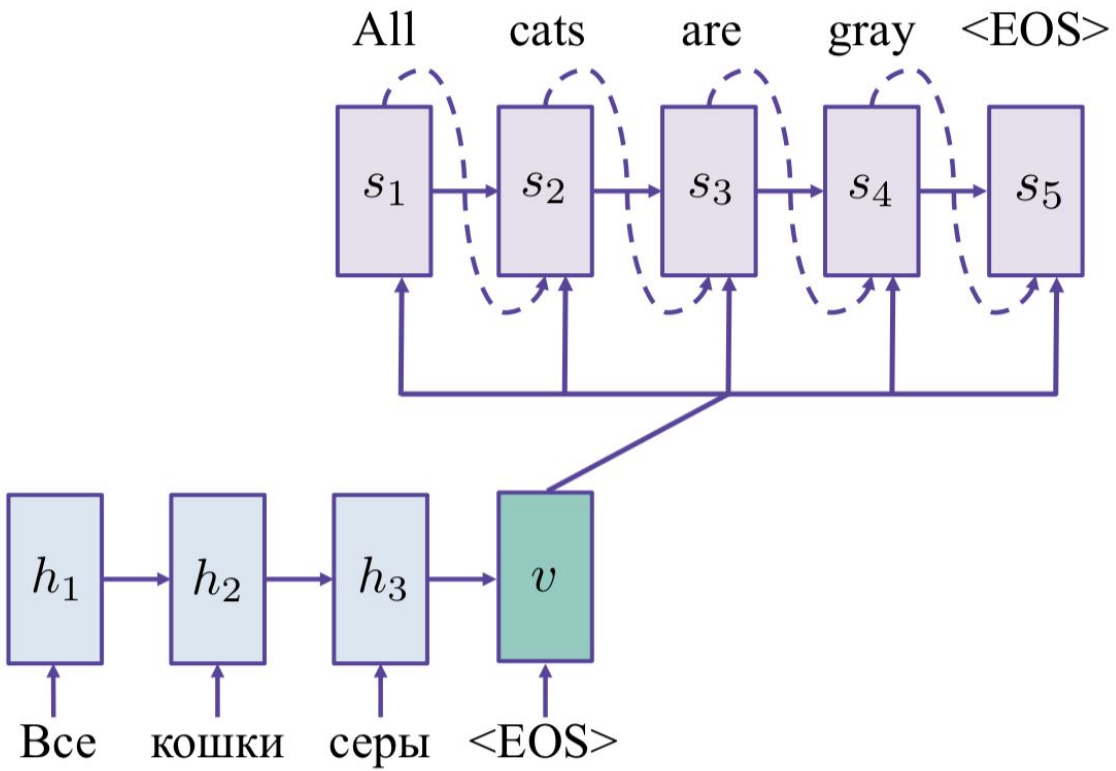
Attention is All You Need

Attention

Seq2seq


All
cats
are
gray
<EOS>
h_1
h_2
h_3
v
s_1
s_2
s_3
s_4
Все
кошки
серы
<EOS>
All
cats
are
gray

Seq2seq



Attention

- Additive

$$sim(h_i, s_j) = w^T \tanh(W_h h_i + W_s s_j)$$

- Multiplicative

$$sim(h_i, s_j) = h_i^T W s_j$$

- Dot-product

$$sim(h_i, s_j) = h_i^T s_j$$

RNN в моделировании последовательностей

+

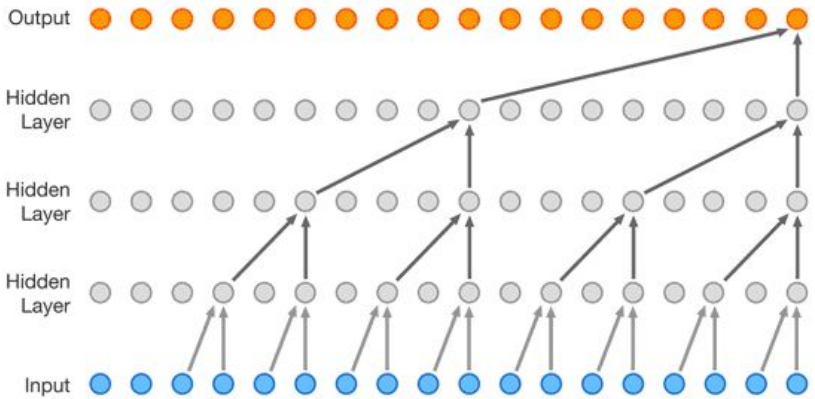
- успешны в задачах с последовательностями различающейся длины

–

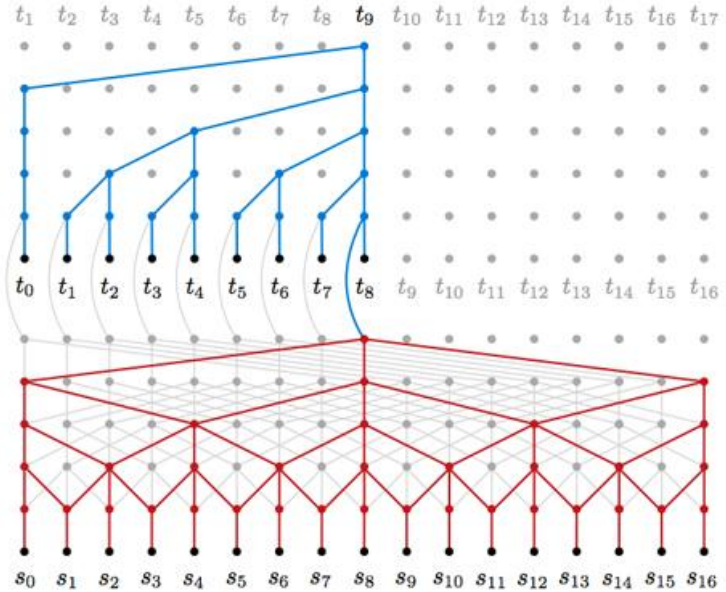
- Последовательная природа вычислений ограничивает распараллеливание
- Всё ещё сложно учитывать далёкий контекст

Auto-regressive CNNs

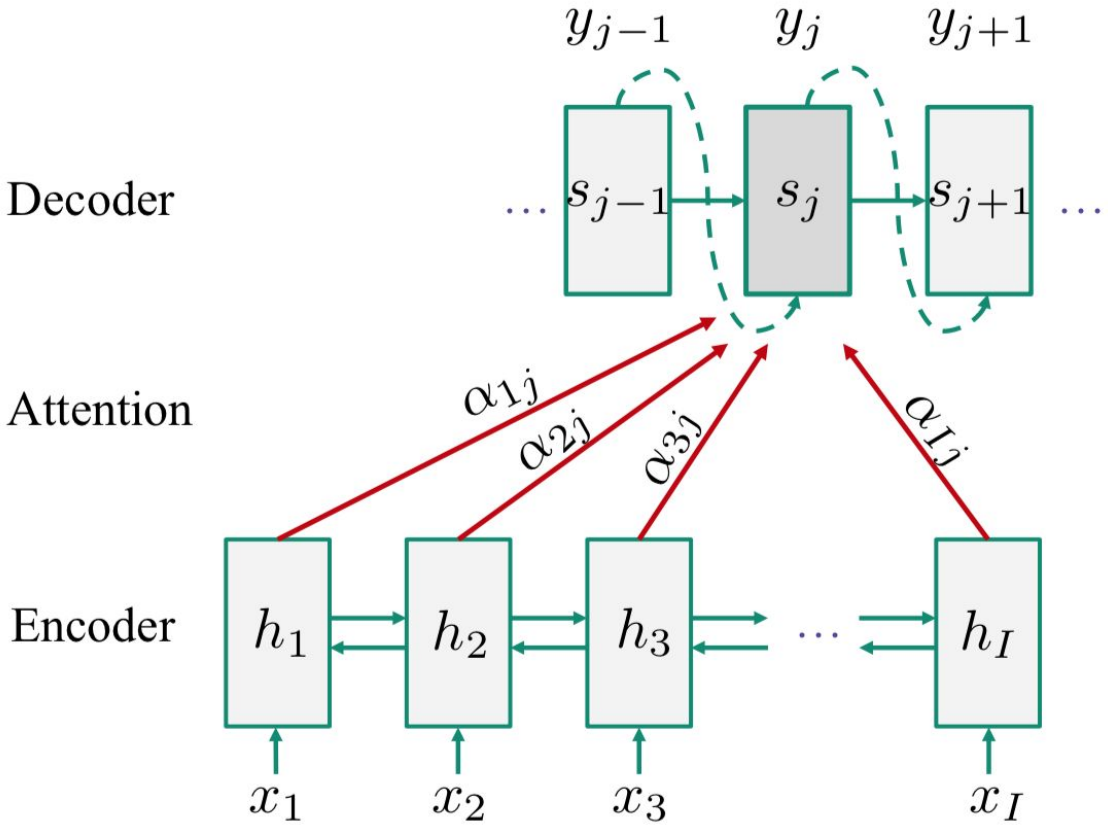
WaveNet



ByteNet



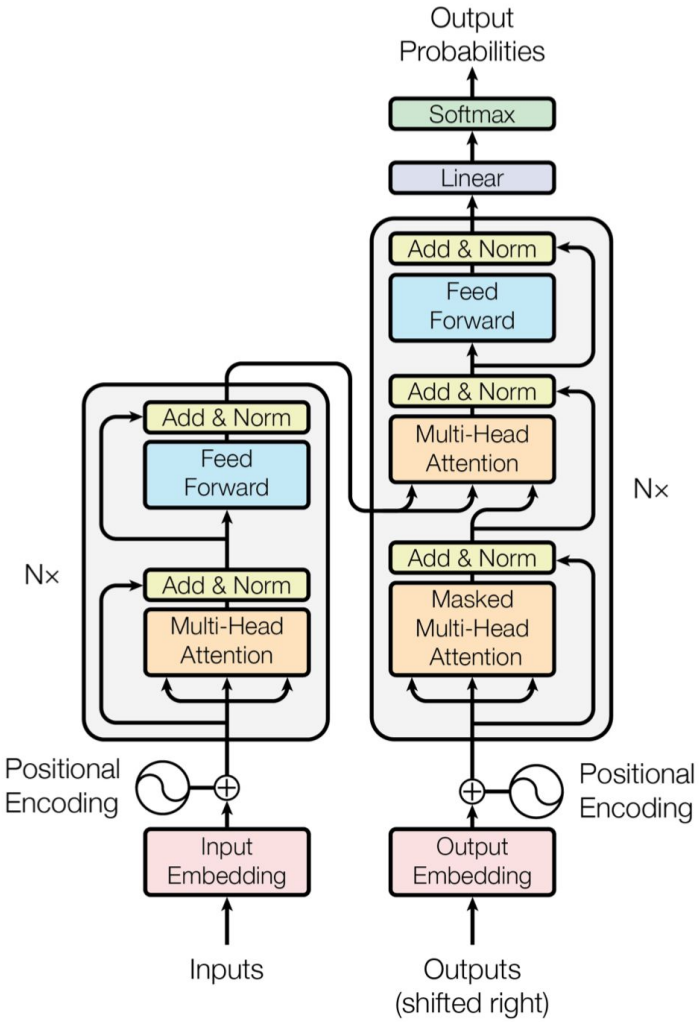
Attention


y_{j-1}
y_j
y_{j+1}
Decoder
s_{j-1}
s_j
s_{j+1}
Attention
α_{1j}
α_{2j}
α_{3j}
α_{Ij}
Encoder
h_1
h_2
h_3
h_I
x_1
x_2
x_3
x_I

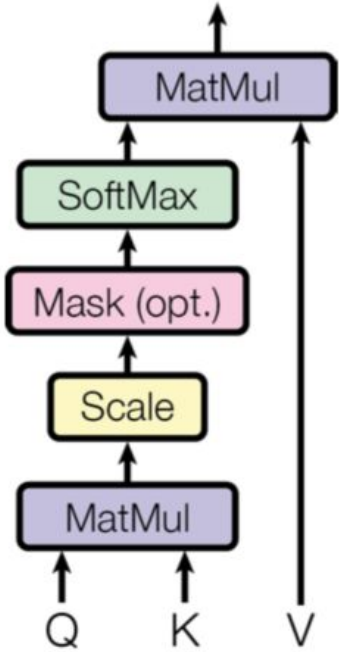
The Transformer

The Transformer

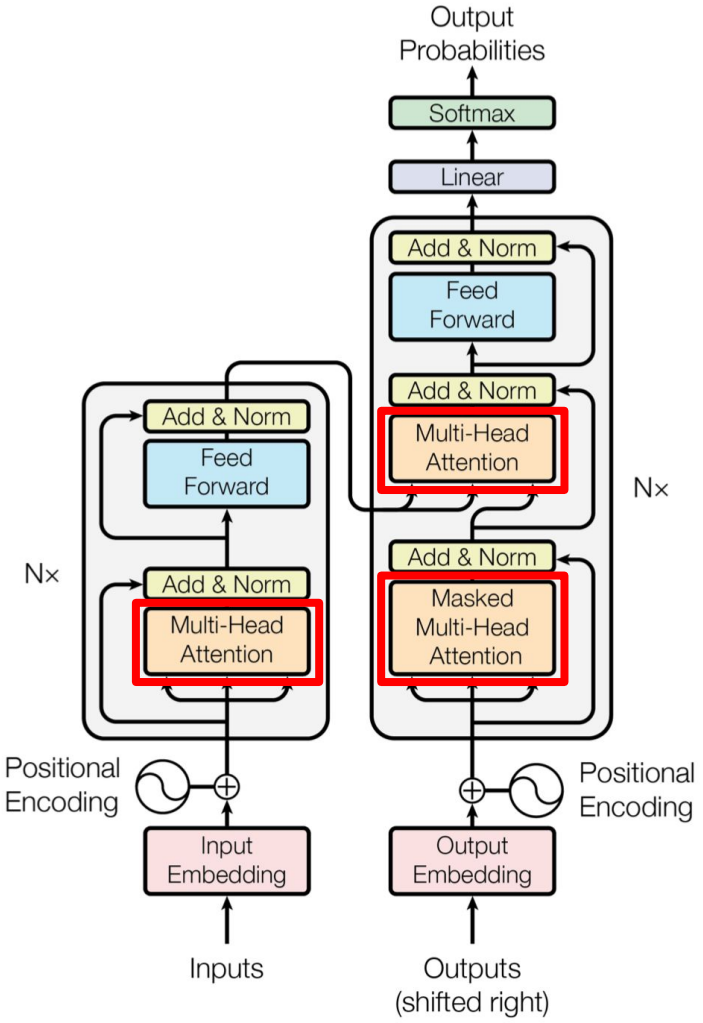
Архитектура



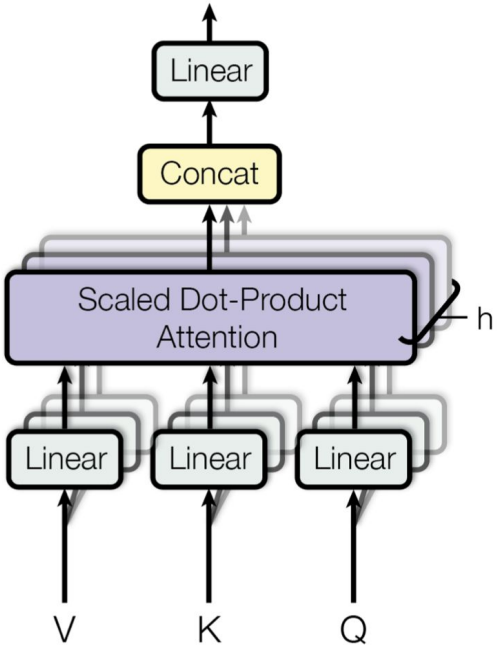
Attention



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

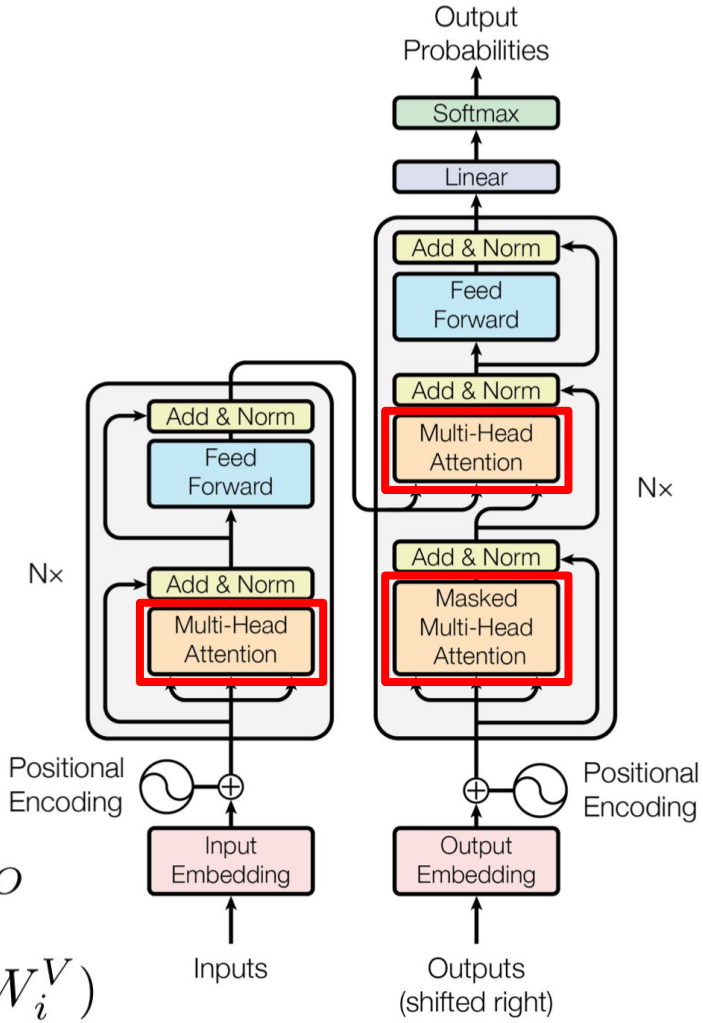


Multi-Head Attention



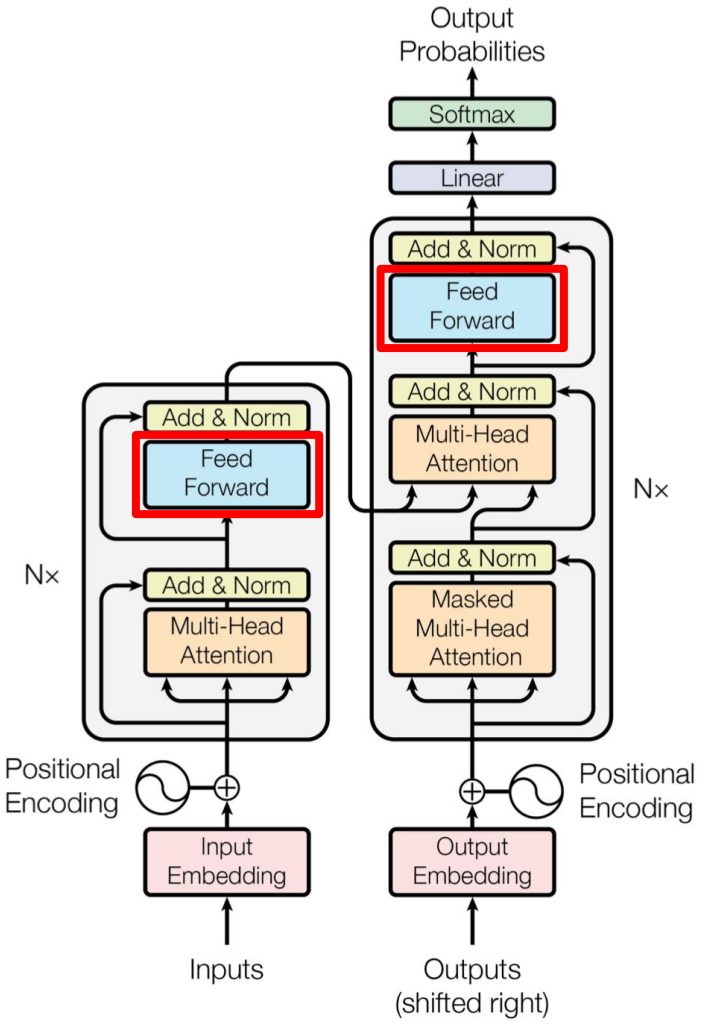
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$

$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Feed Forward Network

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$



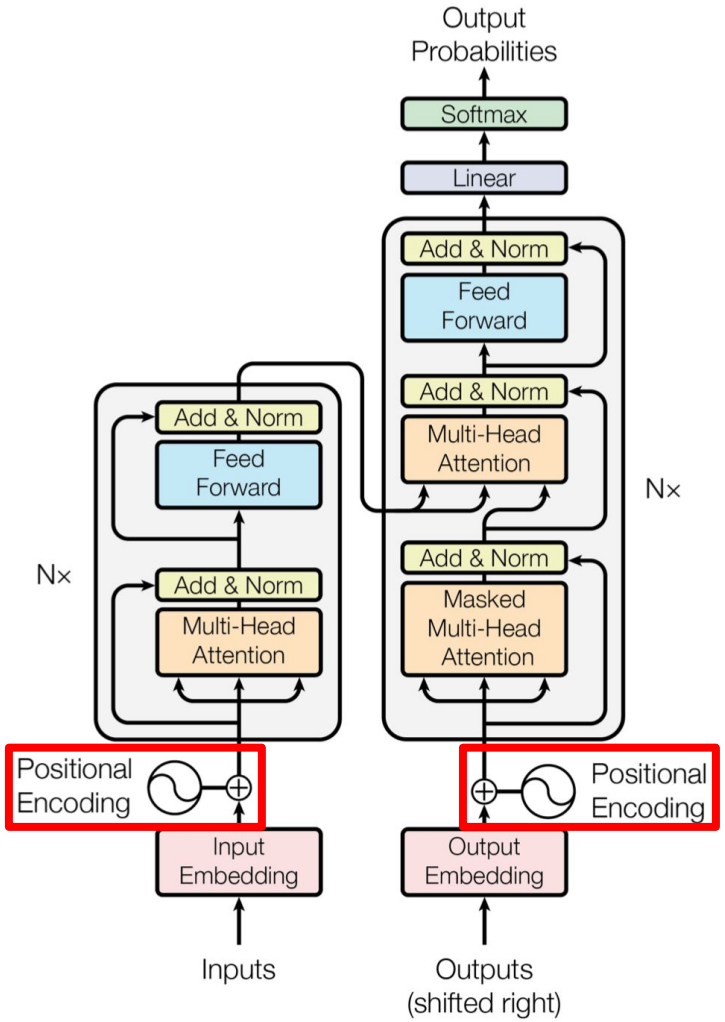
Positional encoding

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

pos — позиция

i — индекс в размерности



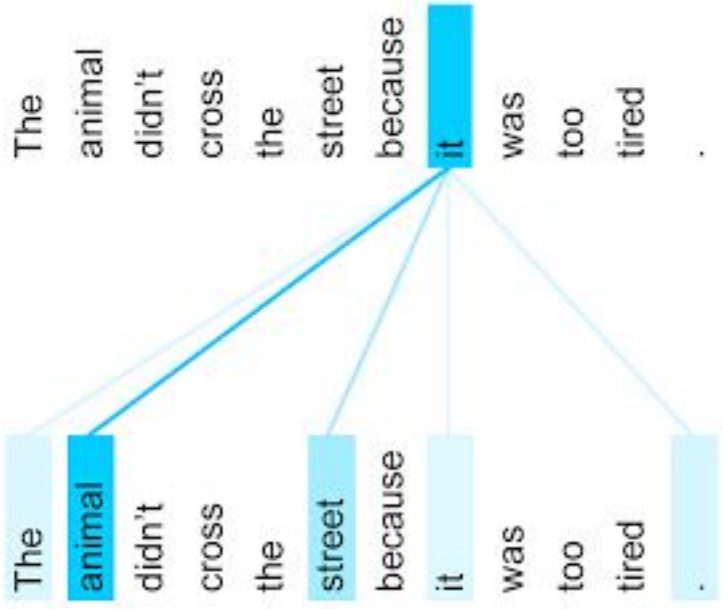
Обучение

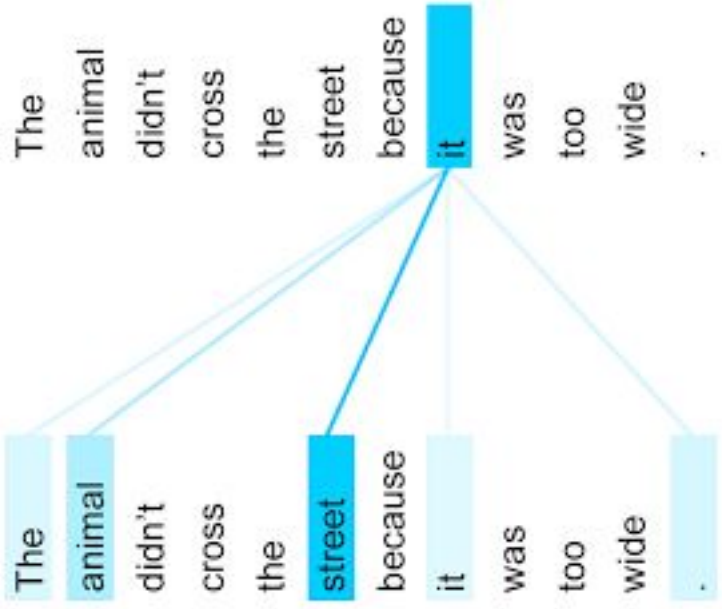
- Базовая модель тренировалась 12 часов на 8 GPU NVIDIA P100 (0.4 сек/шаг).
  - Большие модели тренировались 3.5 дня (1 сек/шаг)
- Алгоритм оптимизации — Adam
- Регуляризация
  - Residual Dropout
  - Attention Dropout
  - Label Smoothing

Результаты

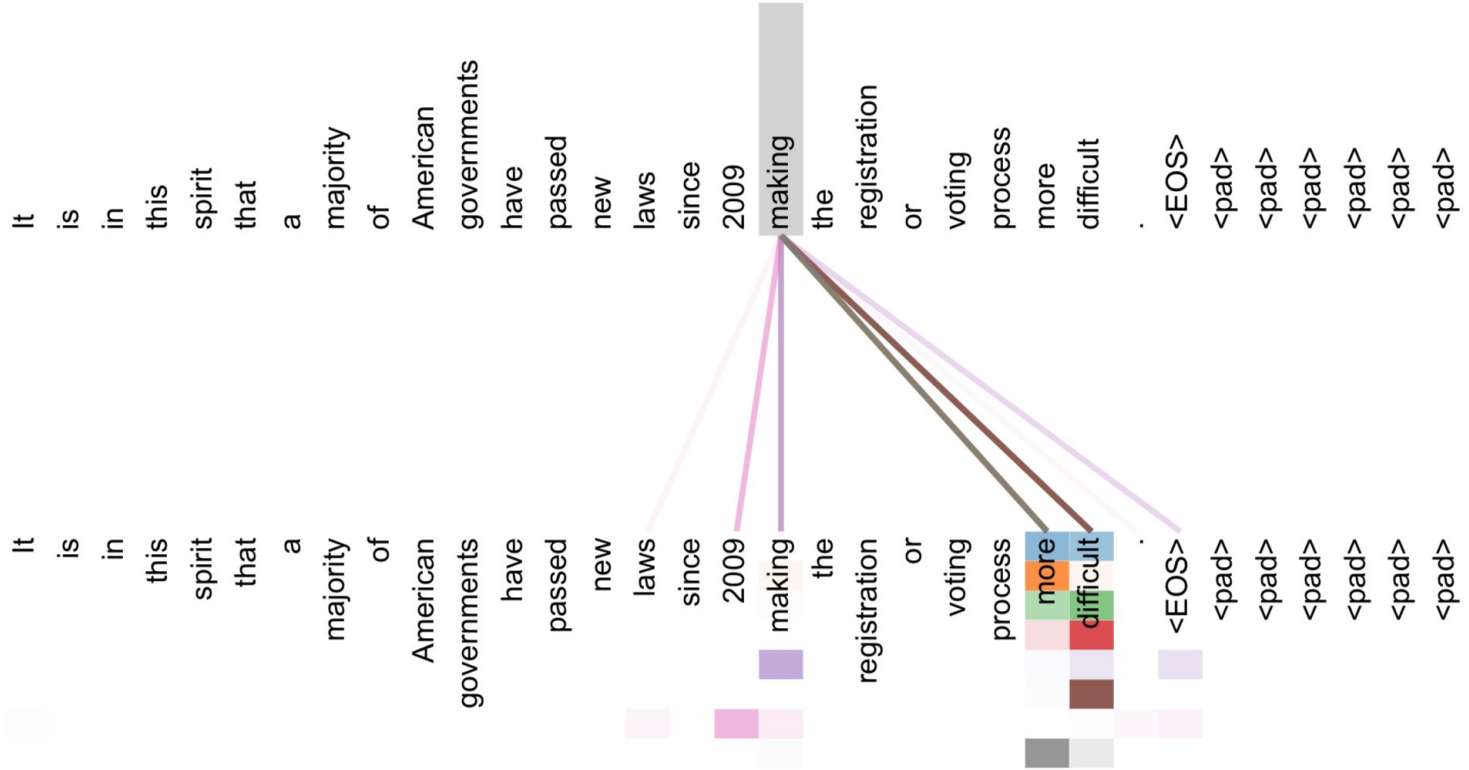
Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [17]	23.75			
Deep-Att + PosUnk [37]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [36]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [31]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [37]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [36]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Визуализация attention

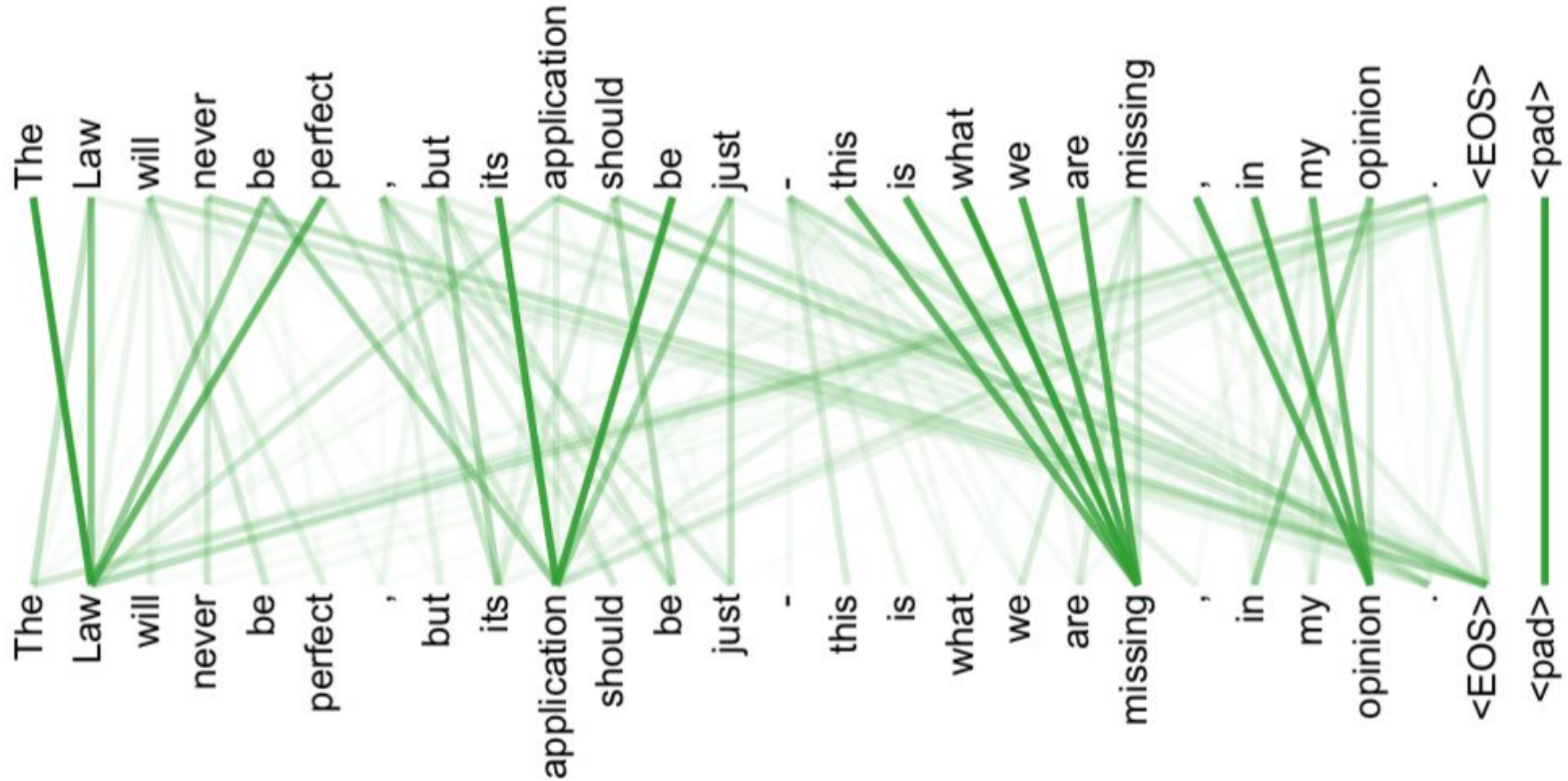




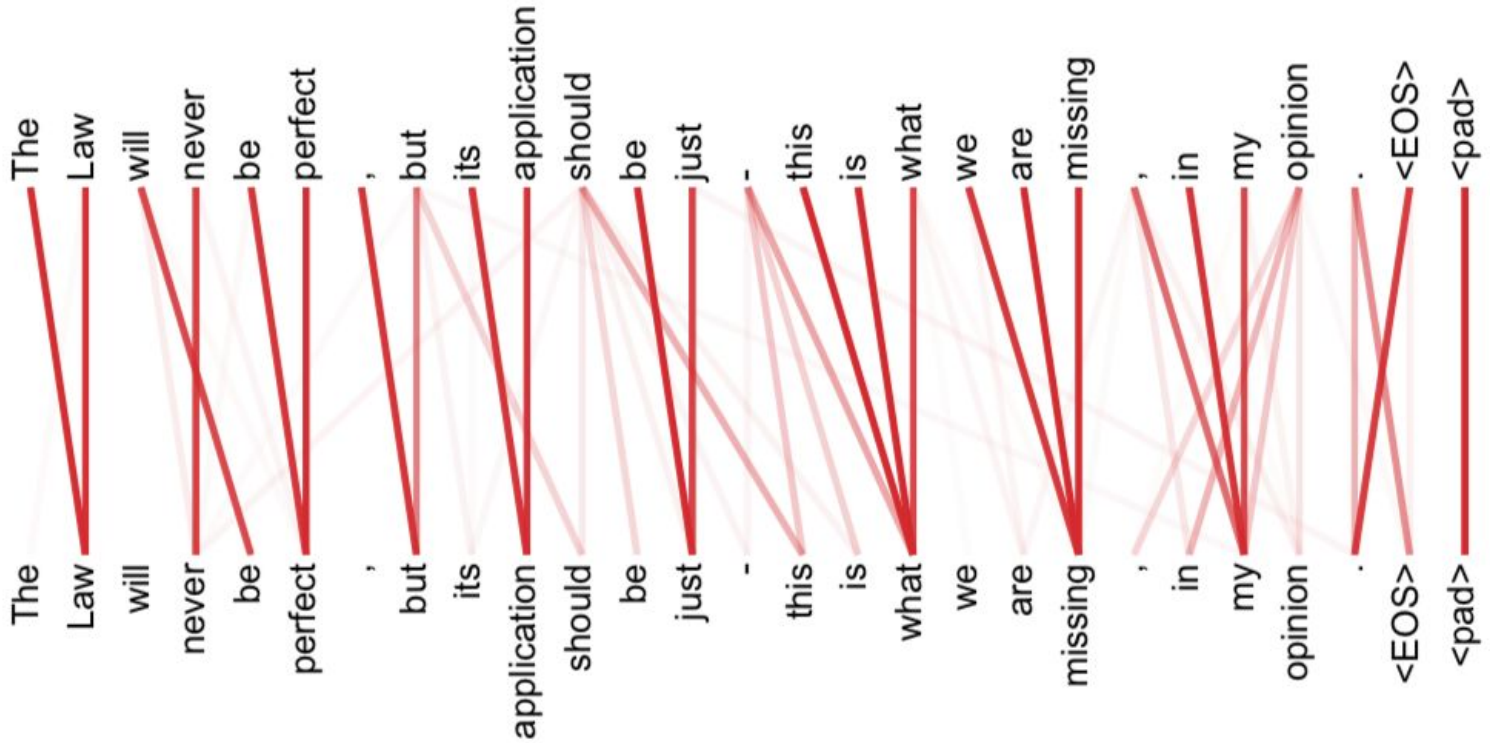
Визуализация attention



Визуализация attention



Визуализация attention



Источники

- Attention is all you need
- Выступление одного из авторов об этой модели и её расширении
- Заметка в блоге Google Research