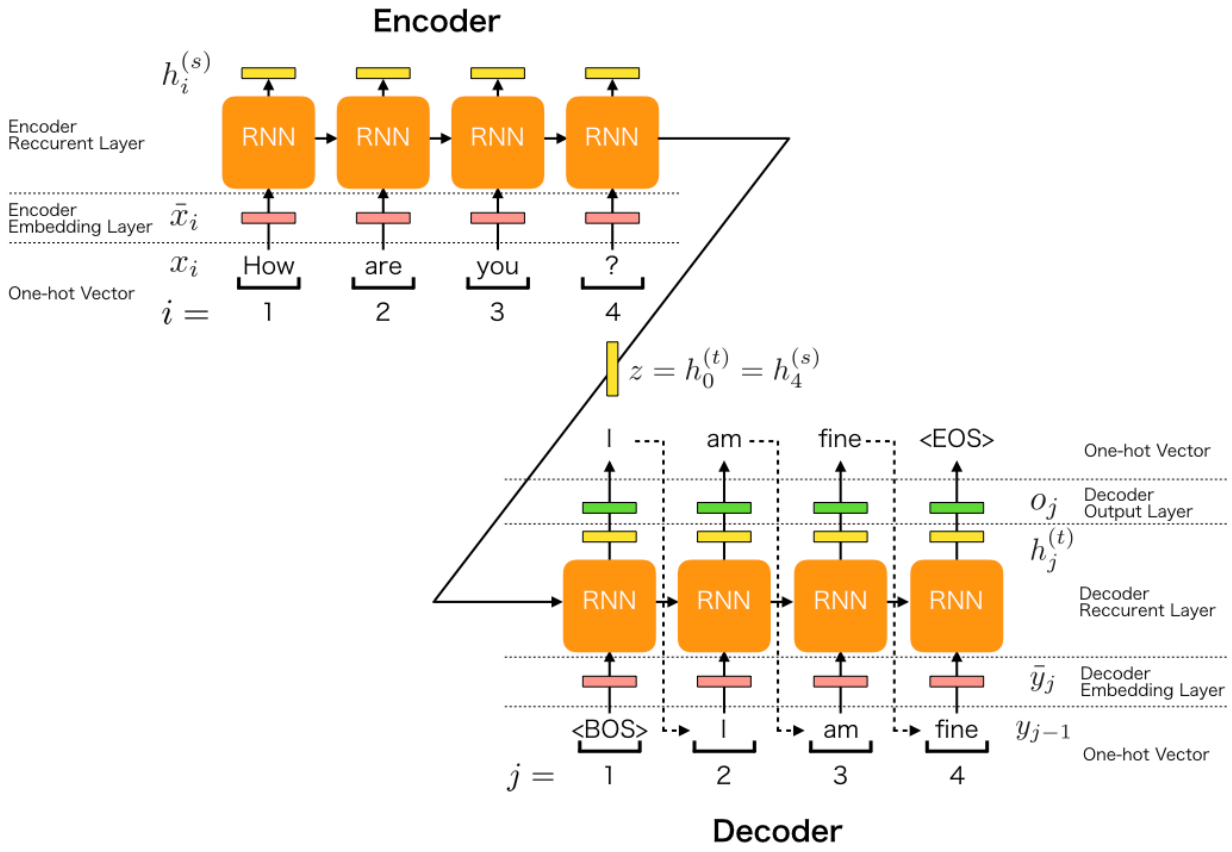
Attention Is All You Need

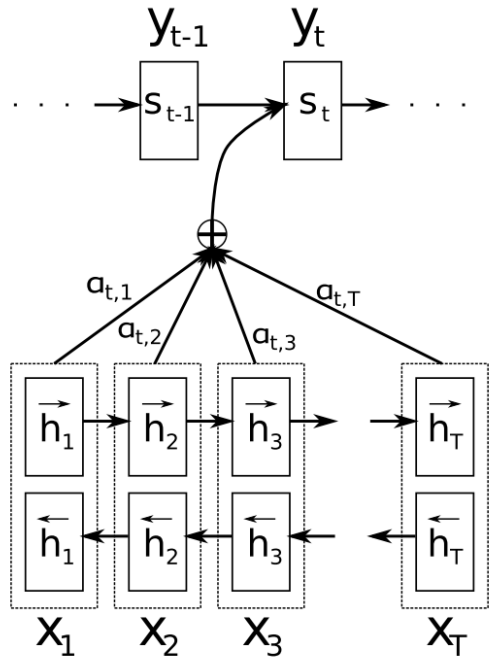
Прокопьева Дарья
НИУ ВШЭ
2019

Содержание

- Seq2seq, attention
- Transformer
- Результаты

Seq2seq, attention





Transformer – model architecture

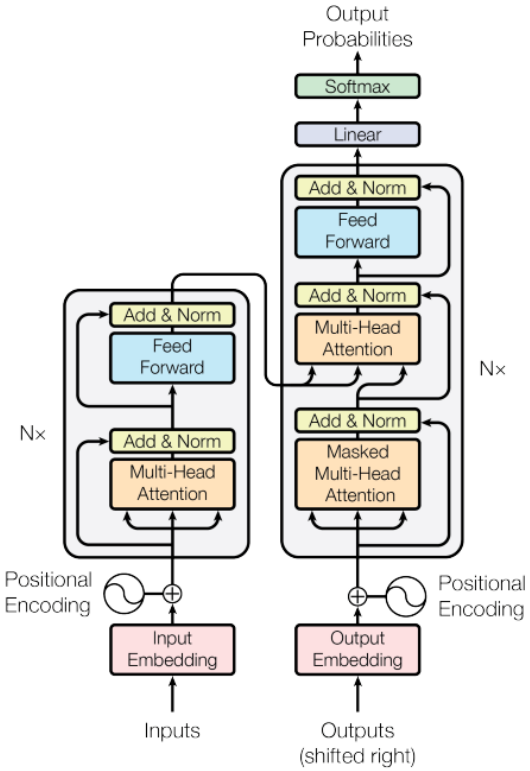
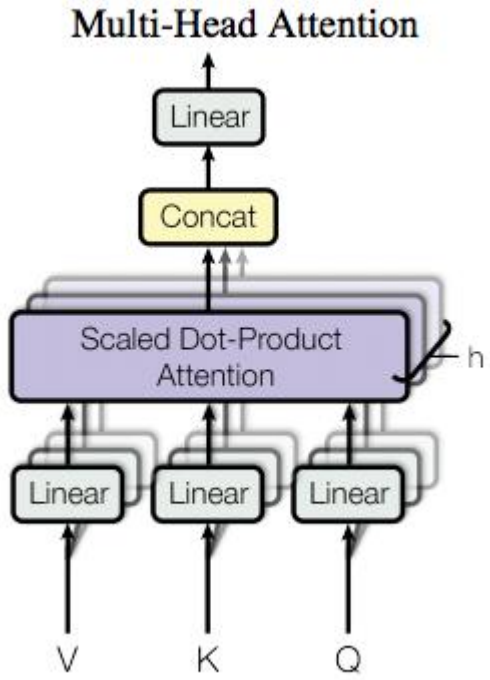
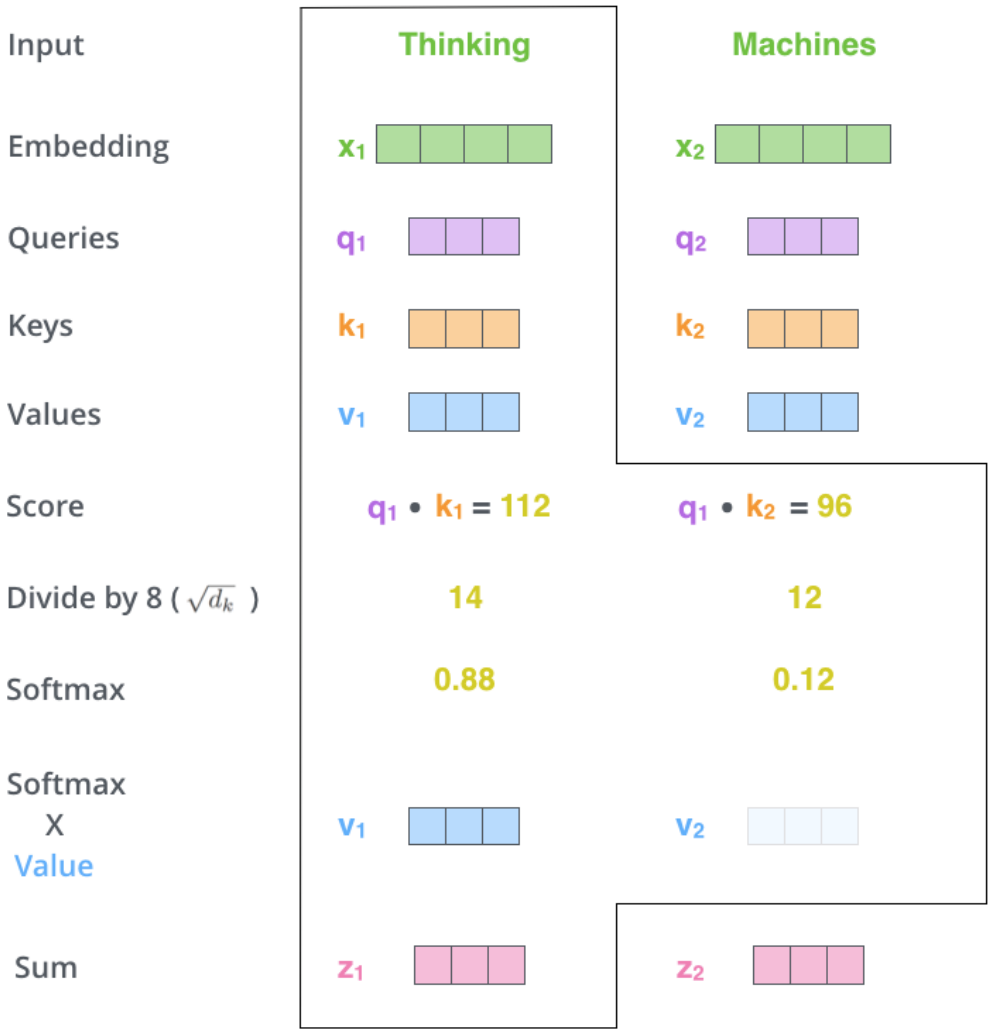


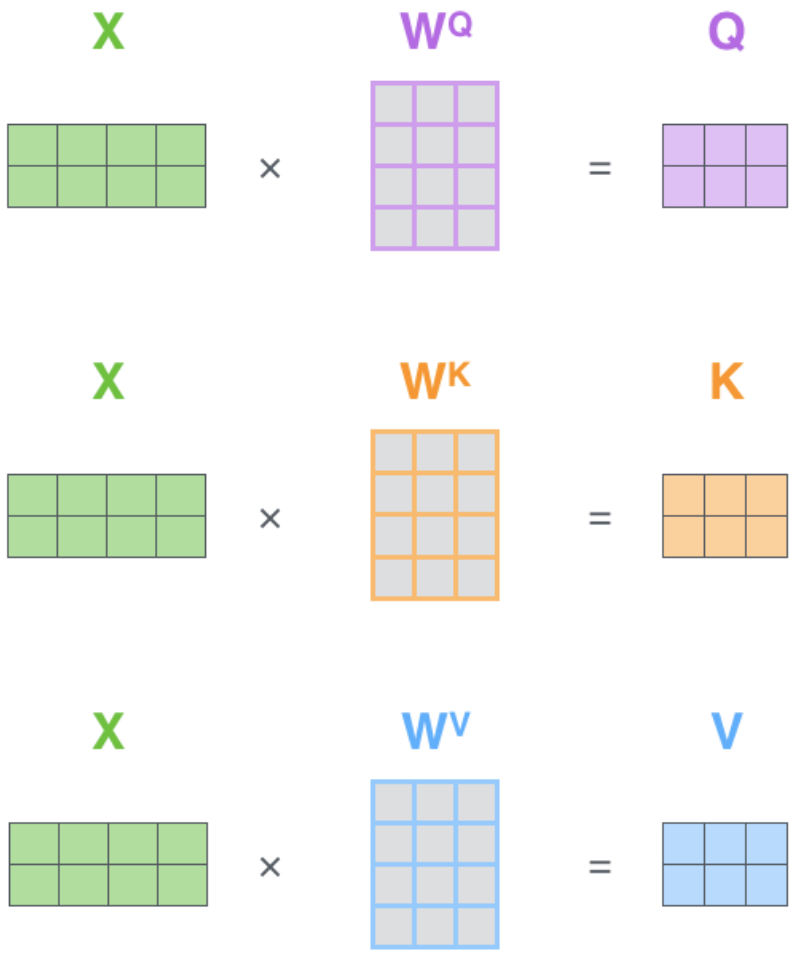
Figure 1: The Transformer - model architecture.



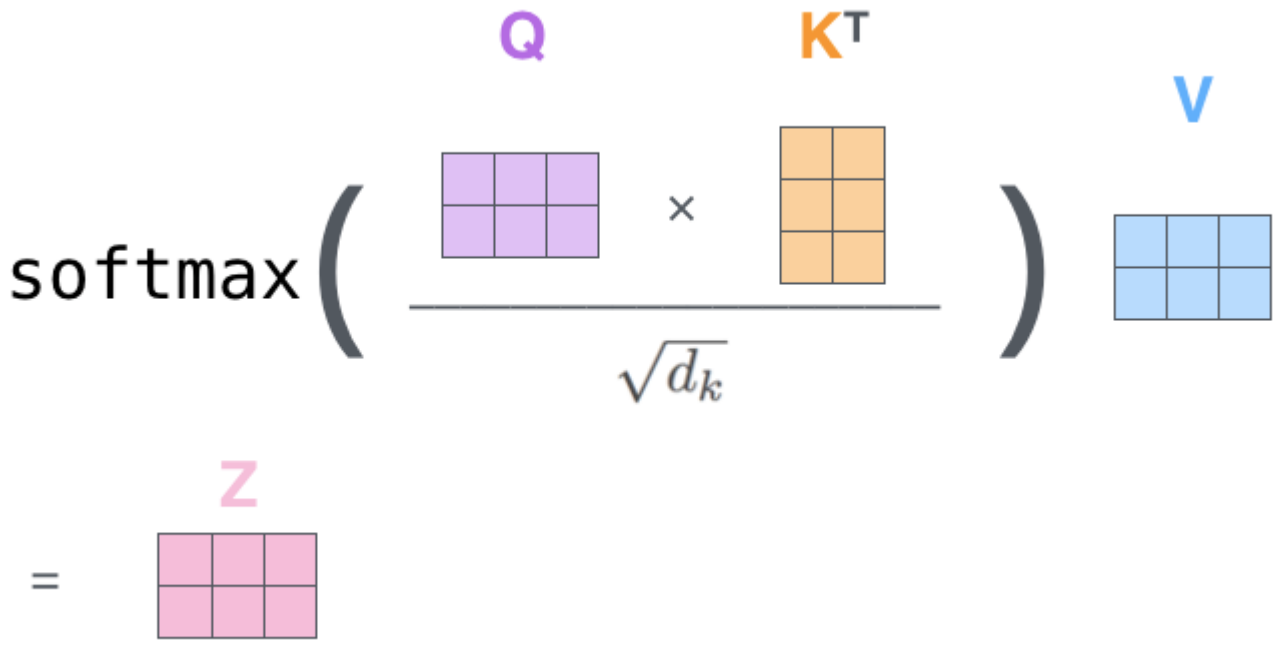
Self-attention



Self-attention



Self-attention



Multi-headed self-attention

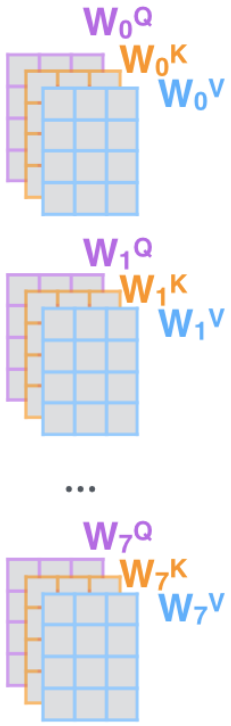
1) This is our input sentence*

Thinking
Machines

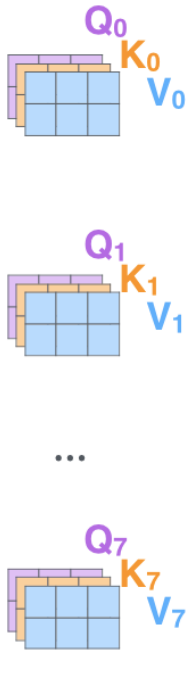
2) We embed each word*

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

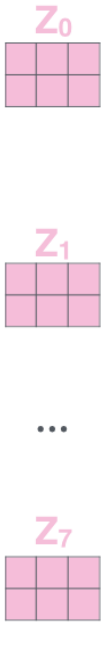
3) Split into 8 heads. We multiply X or R with weight matrices

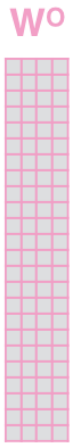


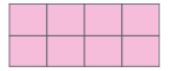
4) Calculate attention using the resulting Q/K/V matrices



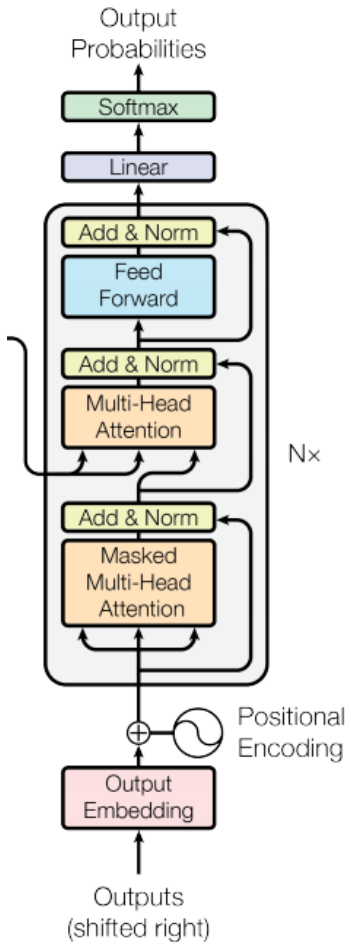
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer



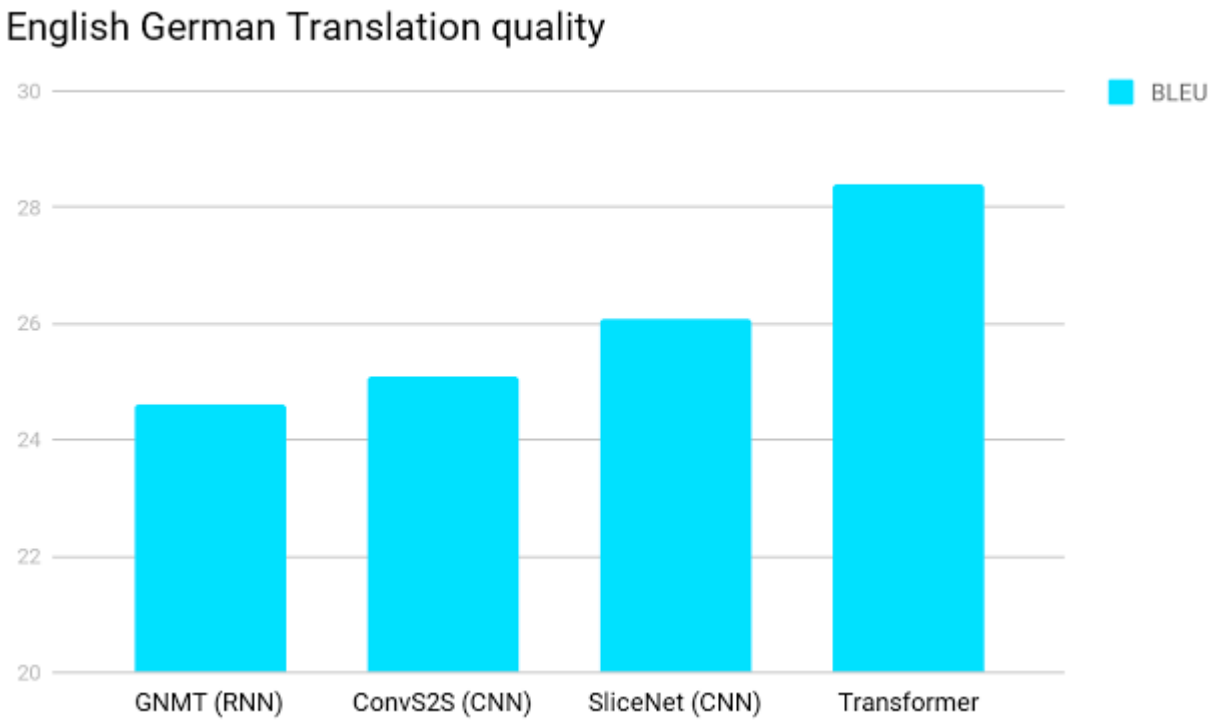




Decoder


Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Positional
Encoding
Output
Embedding
Outputs
(shifted right)

Результаты


English German Translation quality
BLEU
30
28
26
24
22
20
GNMT (RNN)
ConvS2S (CNN)
SliceNet (CNN)
Transformer

Результаты

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	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 [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	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**	**41.8**	$2.3 \cdot 10^{19}$	

Источники

- https://arxiv.org/pdf/1706.03762.pdf
- http://jalammar.github.io/illustrated-transformer/
- https://habr.com/ru/post/341240/