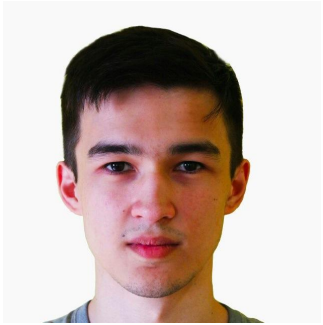
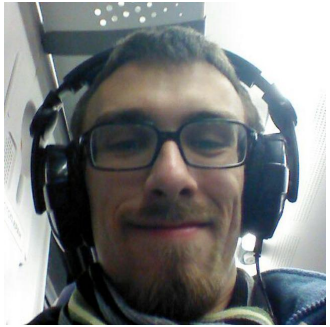
Pros&Cons of Encoder-Decoder architectures



Galim Turumtaev

Pauline Matavina

Valeriia Nemychnikova



Alexey Kulikov

Vladislav Shakhray

Elena Kirilenko

Google's NMT

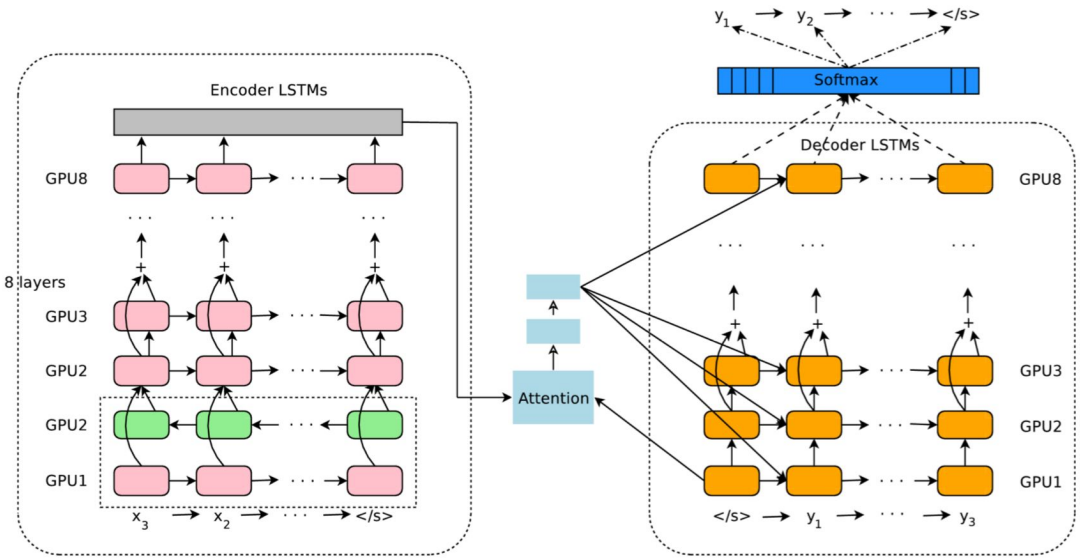


Table 4: Single model results on WMT En→Fr (newstest2014)

Model	BLEU	CPU decoding time per sentence (s)
Word	37.90	0.2226
Character	38.01	1.0530
WPM-8K	38.27	0.1919
WPM-16K	37.60	0.1874
WPM-32K	38.95	0.2118
Mixed Word/Character	38.39	0.2774
PBMT [15]	[illegible]	
LSTM (6 layers) [31]	31.5	
LSTM (6 layers + PosUnk) [31]	33.1	
Deep-Att [45]	37.7	
Deep-Att + PosUnk [45]	39.2	

Table 5: Single model results on WMT En→De (newstest2014)

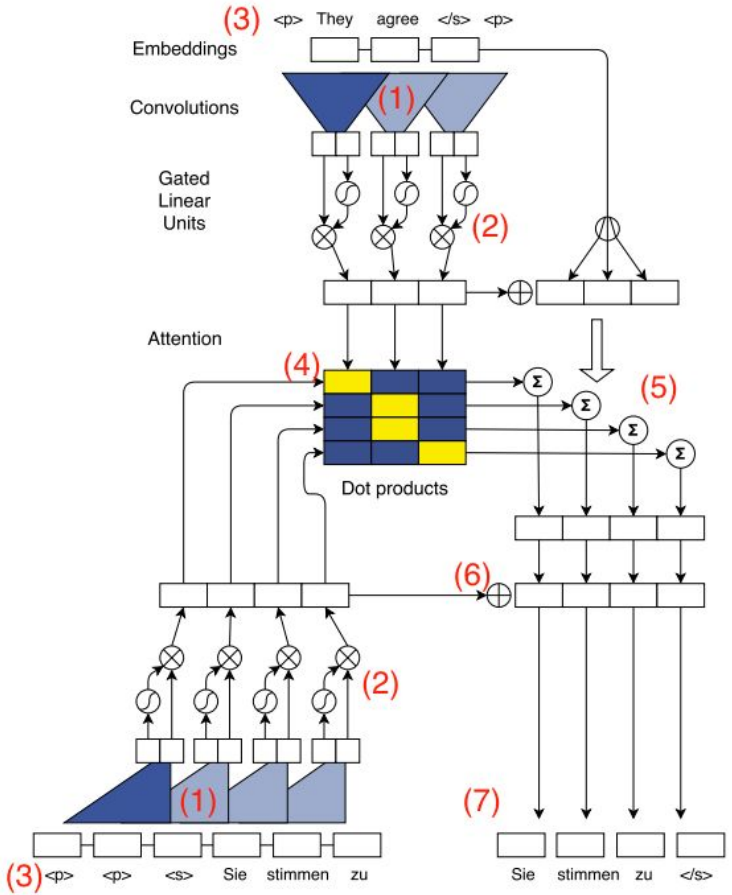
Model	BLEU	CPU decoding time per sentence (s)
Word	23.12	0.2972
Character (512 nodes)	22.62	0.8011
WPM-8K	23.50	0.2079
WPM-16K	24.36	0.1931
WPM-32K	24.61	0.1882
Mixed Word/Character	24.17	0.3268
PBMT [6]	20.7	
RNNSearch [37]	16.5	
RNNSearch-LV [37]	16.9	
RNNSearch-LV [37]	16.9	
Deep-Att [45]	20.6	

- **word-piece** model (something like BPE)
- Quantized learning [on training]
- Attention, of course! (2-layer feed-forward net)
- **Deep** (x8) LSTMs (Google says it's crucial) + bidirection
- Featuring **residual connections** (otherwise: gradient problems)
- Attention is calculated only for the bottom layer

Convolutional seq2seq Learning

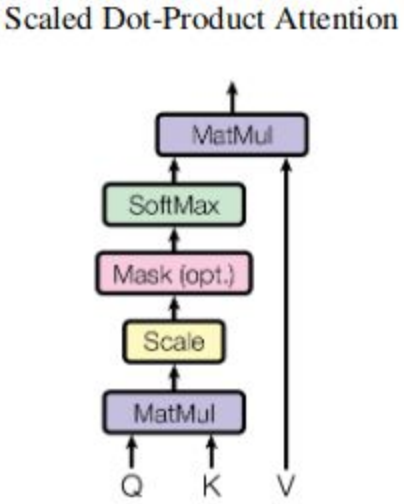
- Fully parallelized during training
- Multi-step Attention
- Generation Speed
- Copy activations that have not changed
- Position Embeddings
- Gated linear units (GLU) as non-linearity
- Residual connections
- For each decoder output we obtain "fraction of importance" for encoder

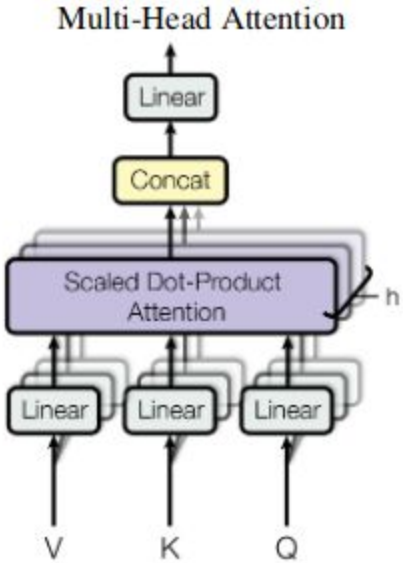
BLEU:
40.51 @ WMT14 Eng-Fr
25.16 @ WMT14 Eng-Ger



Attention is All You Need

BLEU 28.4 @ WMT14
Eng-German





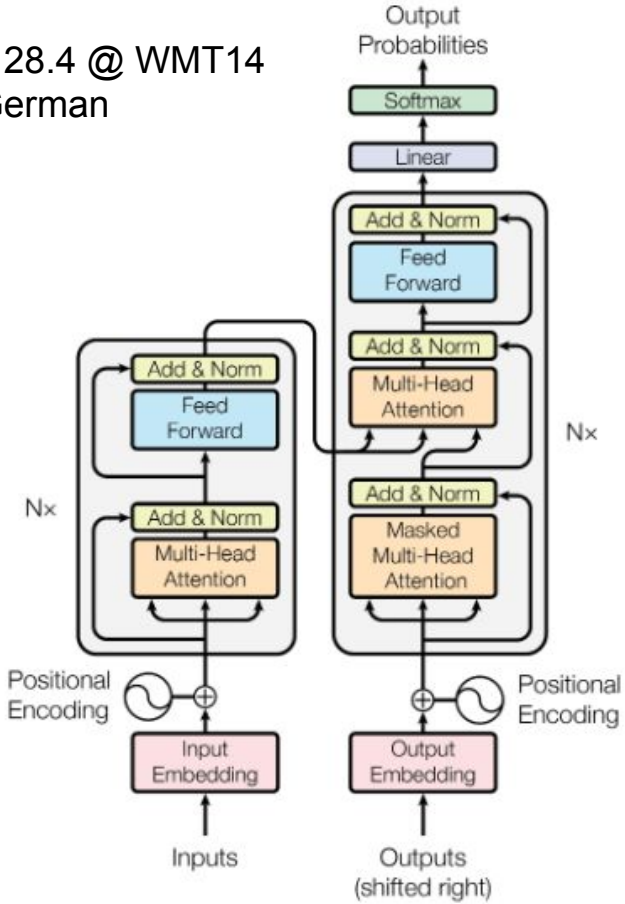


Figure 1: The Transformer - model architecture.

- ~~Recurrence Convolutions~~ **Attention**
- Multi-headed self-attention
- Encoding symbol position
- Much shorter training time! Parallelizable

Colorless Green Neural Networks Dream Hierarchically

- The **girl** you met yesterday through her friends **thinks**…
- RNN
- Nonce sentences
- Italian: compared to people

Accuracy				
LSTM	IT	EN	HE	RU
Original	92.18	1.0	94.7	96.1
Nonce	85.5	74.1	80.8	88.8