Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jacob Uszkoreit

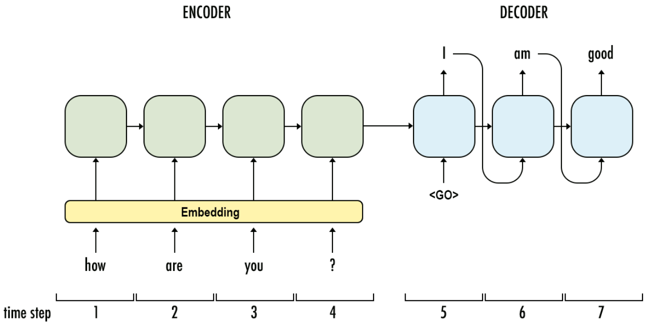
Google Brain, Google Research

April 11, 2019

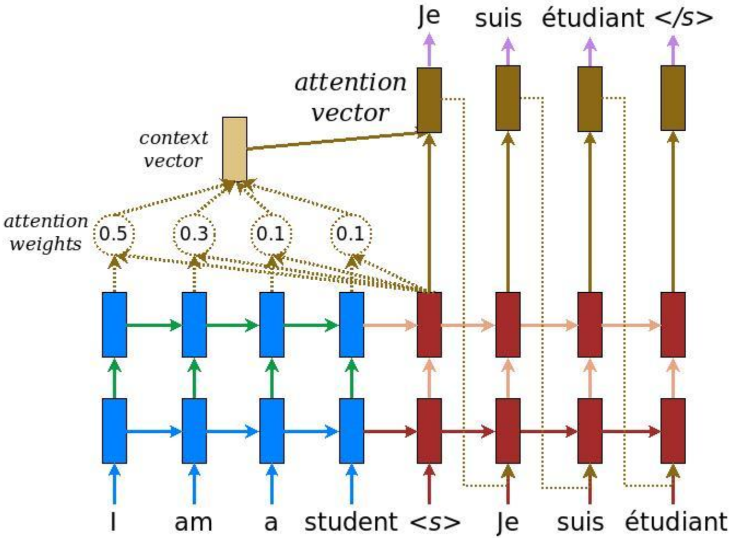
Overview

1. Background
    - Seq2Seq
    - Seq2Seq Attention
2. Model Architecture
    - The Transformer
    - Attention
    - Mask
    - Attentions in transformer
    - Positional Encoding
3. Results

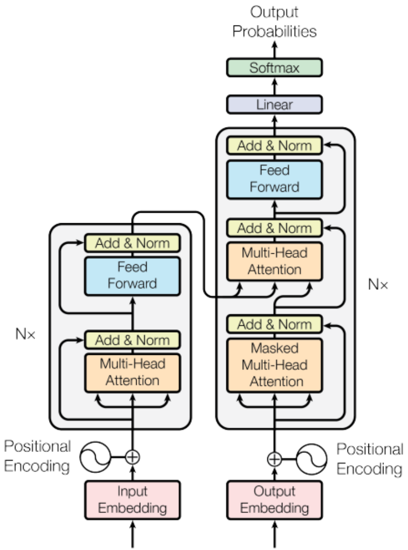
Sequence to sequence



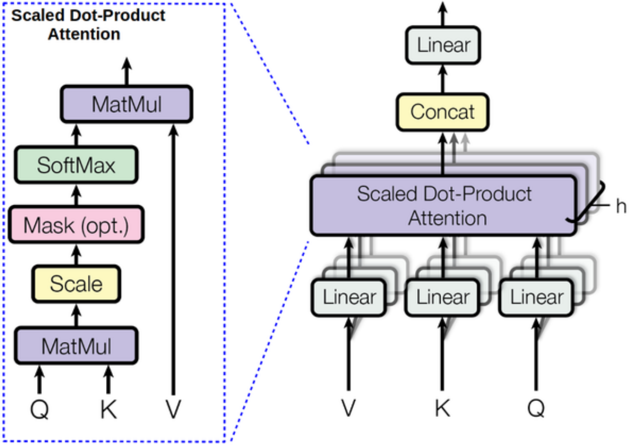
Seq2Seq Attention



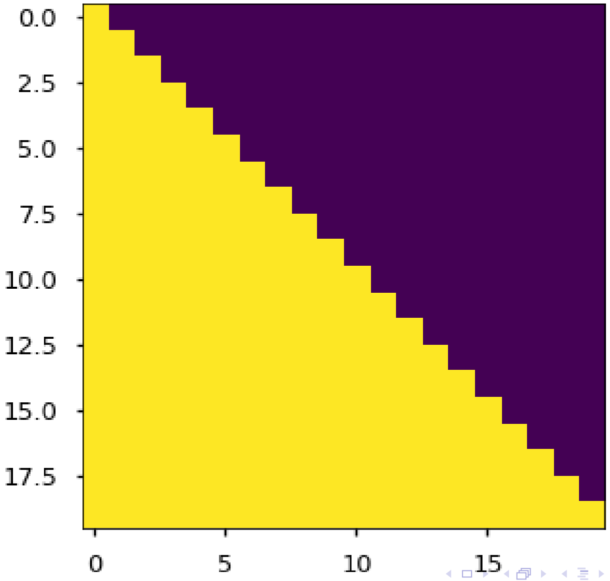
The transformer



Attention


Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

Mask



Attentions in transformer

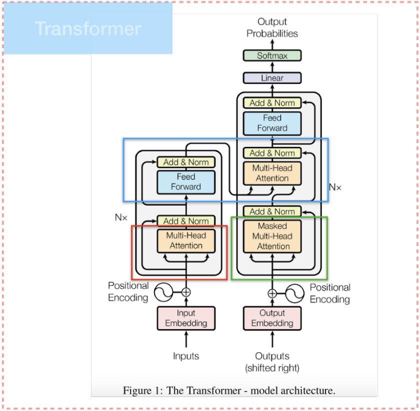


Figure 1: The Transformer - model architecture.

encoder self attention

1. **Multi-head Attention**
2. **Query=Key=Value**

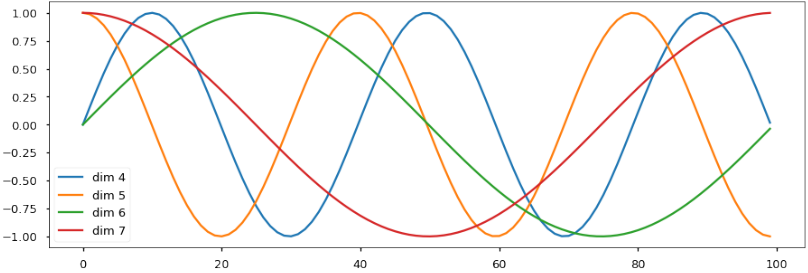
decoder self attention

1. **Masked Multi-head Attention**
2. **Query=Key=Value**

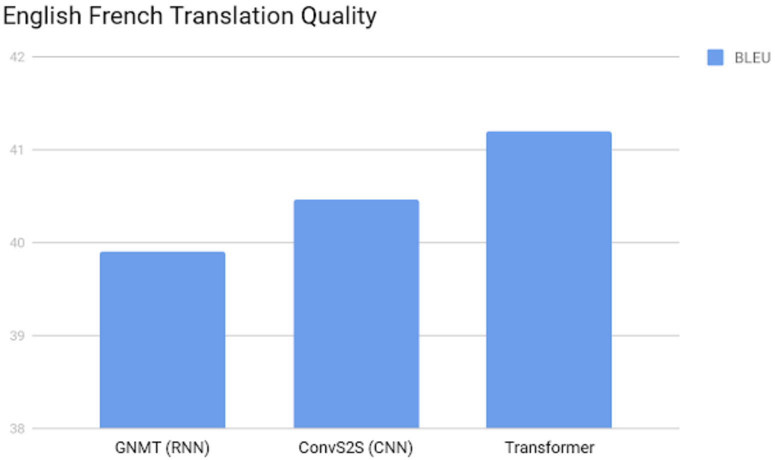
encoder-decoder attention

1. **Multi-head Attention**
2. **Encoder Self attention=Key=Value**
3. **Decoder Self attention=Query**

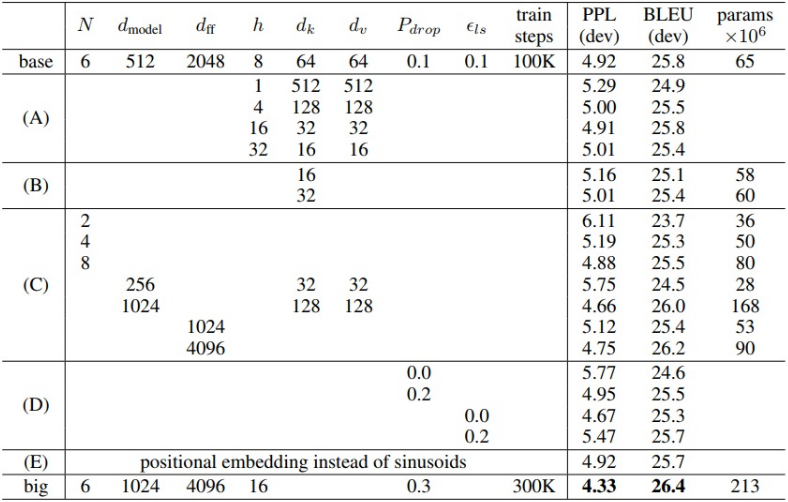
Positional encoding



Machine translation quality



Machine translation on different Transformer architectures



	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train	PPL	BLEU	params
									steps	(dev)	(dev)	$\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Conclusion

- Transformer performs better on Seq2Seq problems
- Transformer trains faster and requires smaller datasets
- Transformer provides interpretability
- It can be applied to any data - texts, images, audio and video

References

- Vaswani et al. *Attention is all you need*. arXiv, 2017
- Kalchbrenner et al. *Neural Machine Translation in linear time*. arXiv, 2002.
- Gu et al. *Non-autoregressive neural machine translation*. arXiv, 2017.
- Lukasz Kaiser and Samy Bengio. *Can active memory replace attention?* NIPS, 2016.