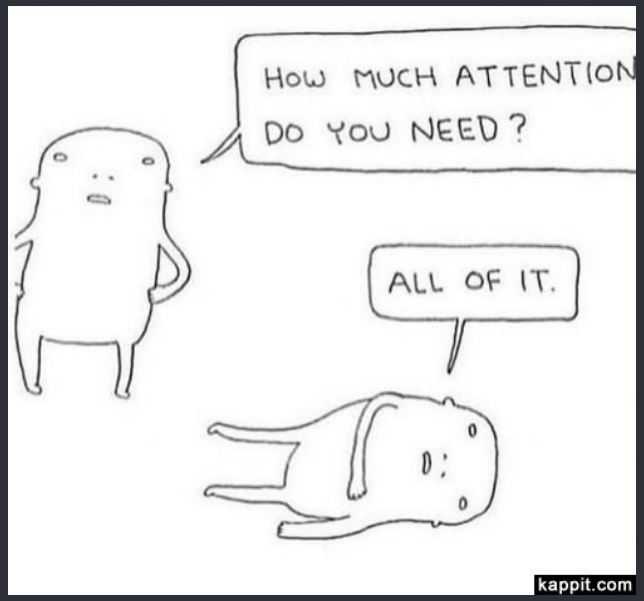
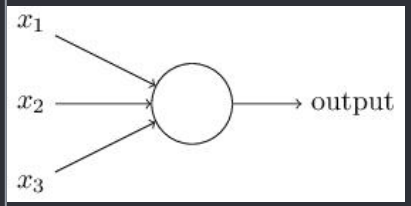
Attention Is All You Need

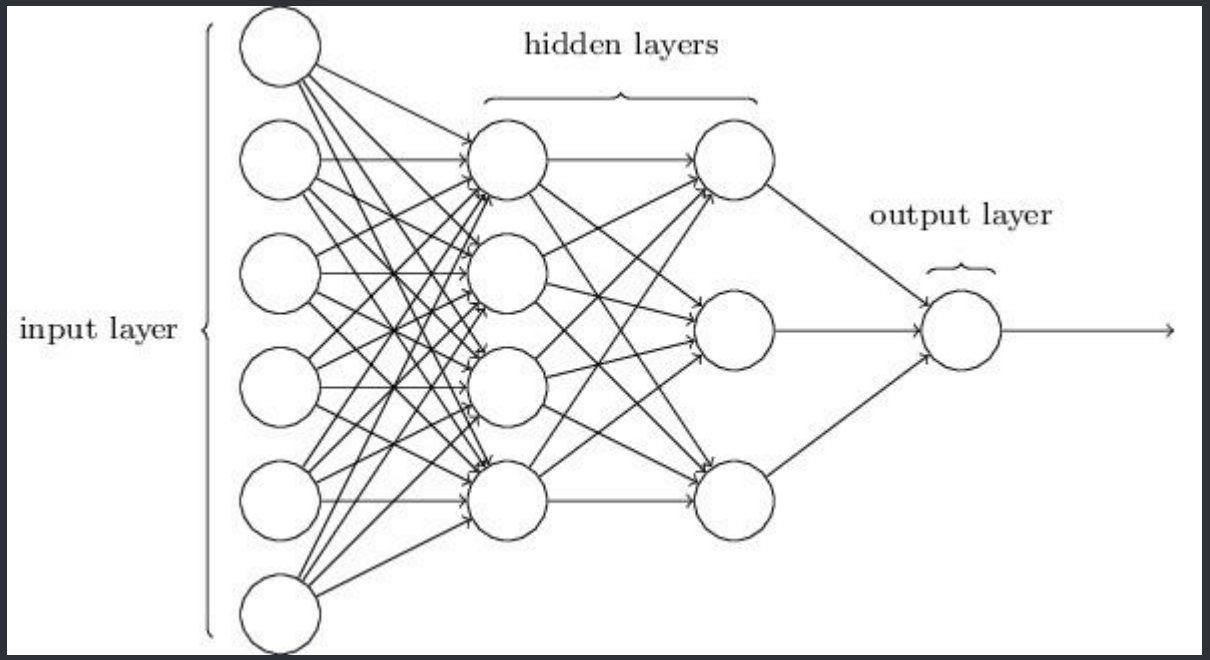


2017 - By:
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit,
Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin
From:
Google brain
Google research

Presented by: Aqeel Labash

Deep learning, algorithms inspired by brain functions & structure



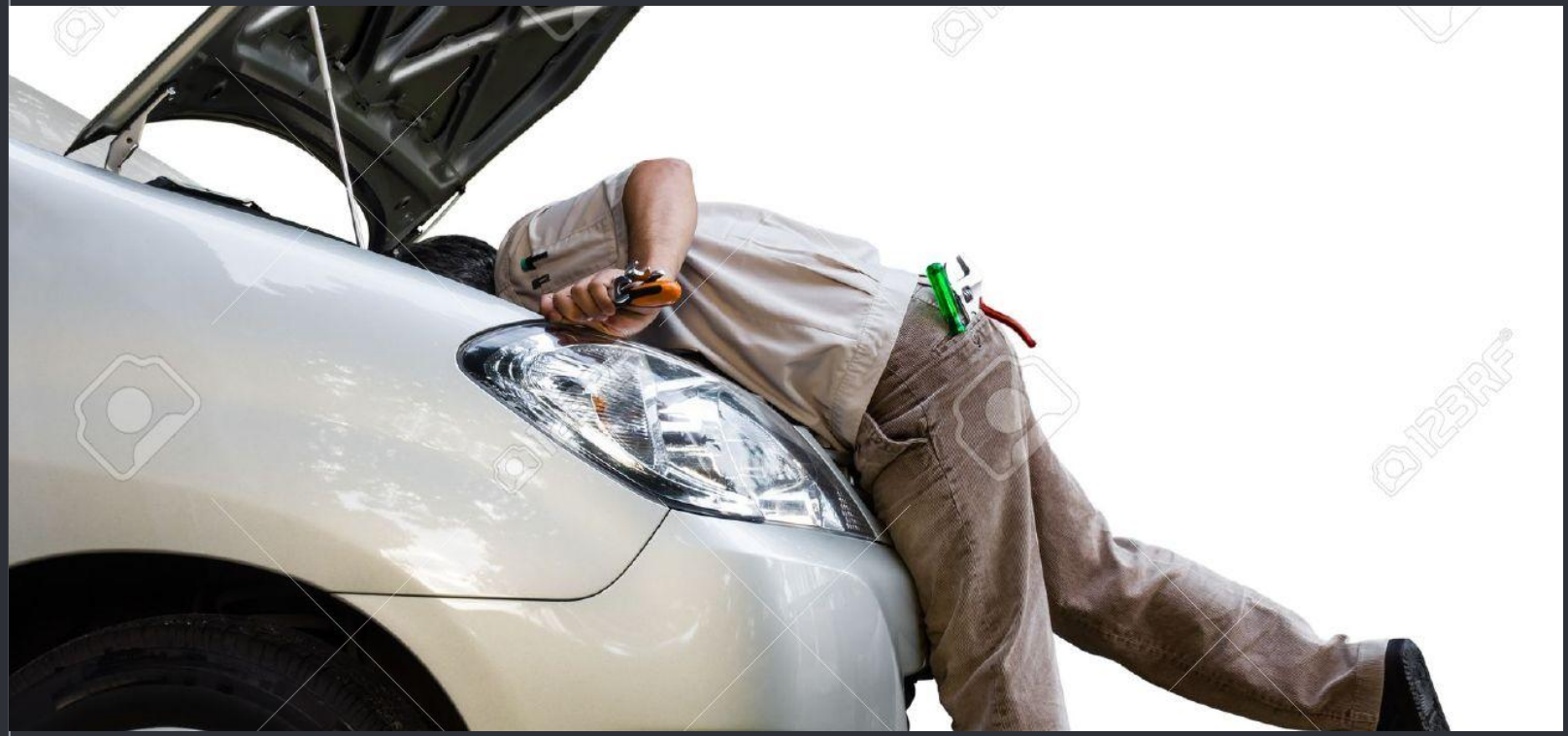


http://neuralnetworksanddeeplearning.com

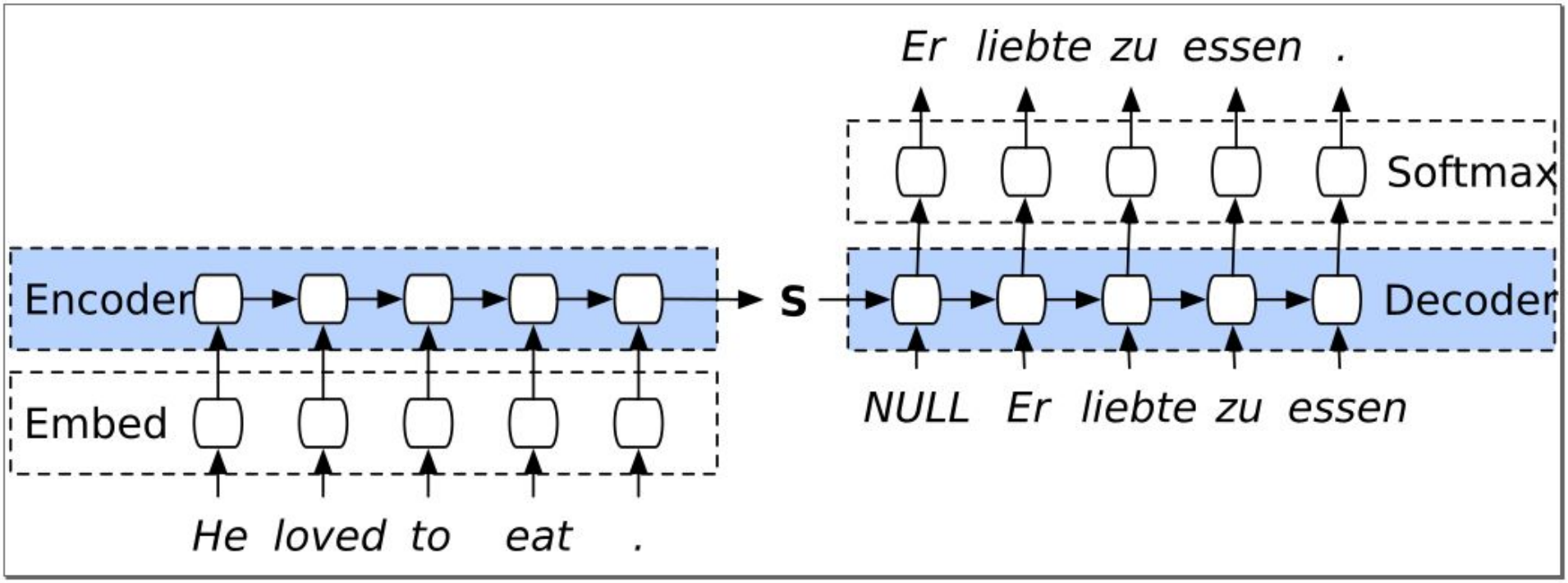
Machine translation is a blackbox

http://blog.webcertain.com/the-dangers-of-machine-translation/09/07/2014/

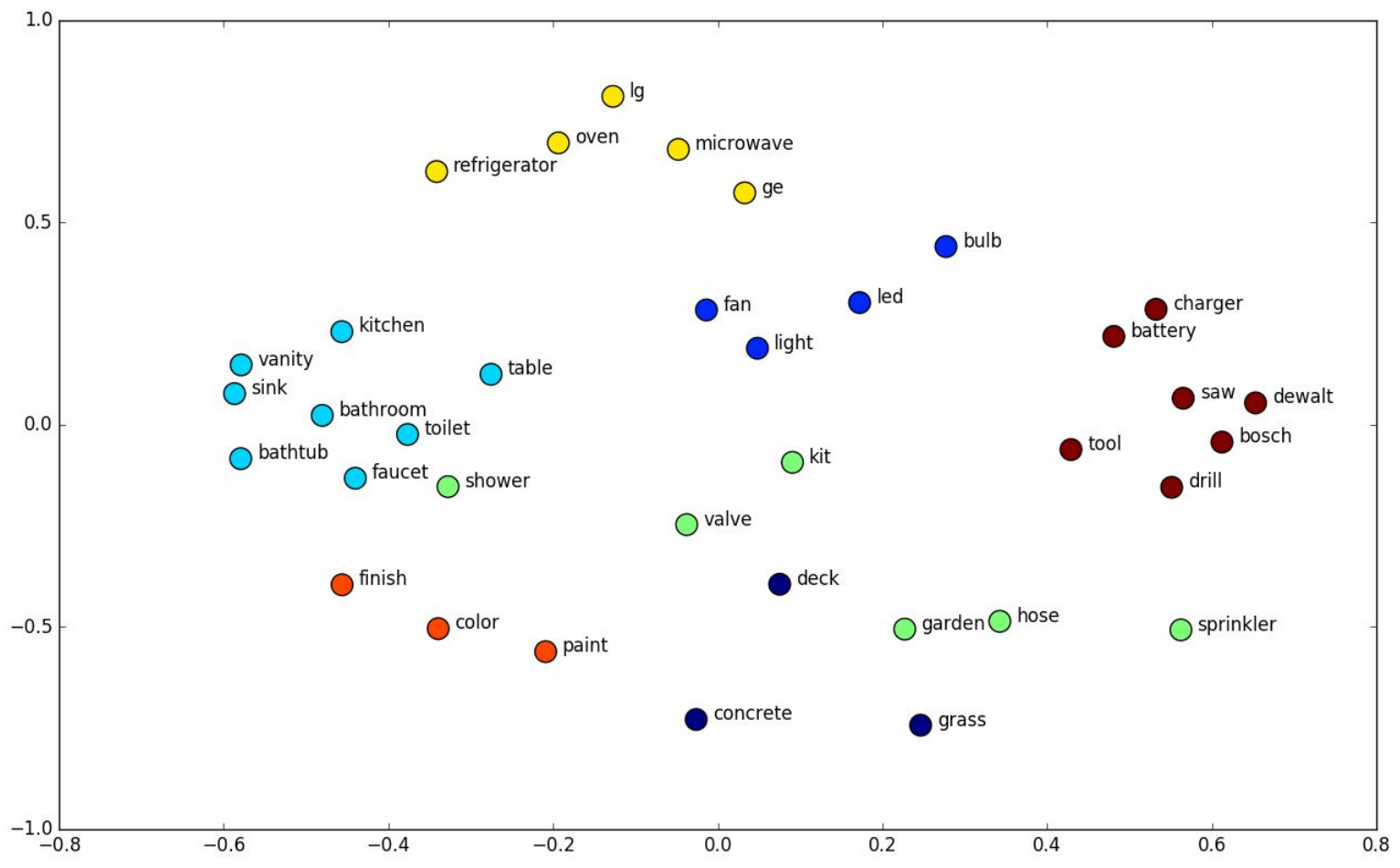
Taking a look under the hood



- But First How Machine Translation work ?

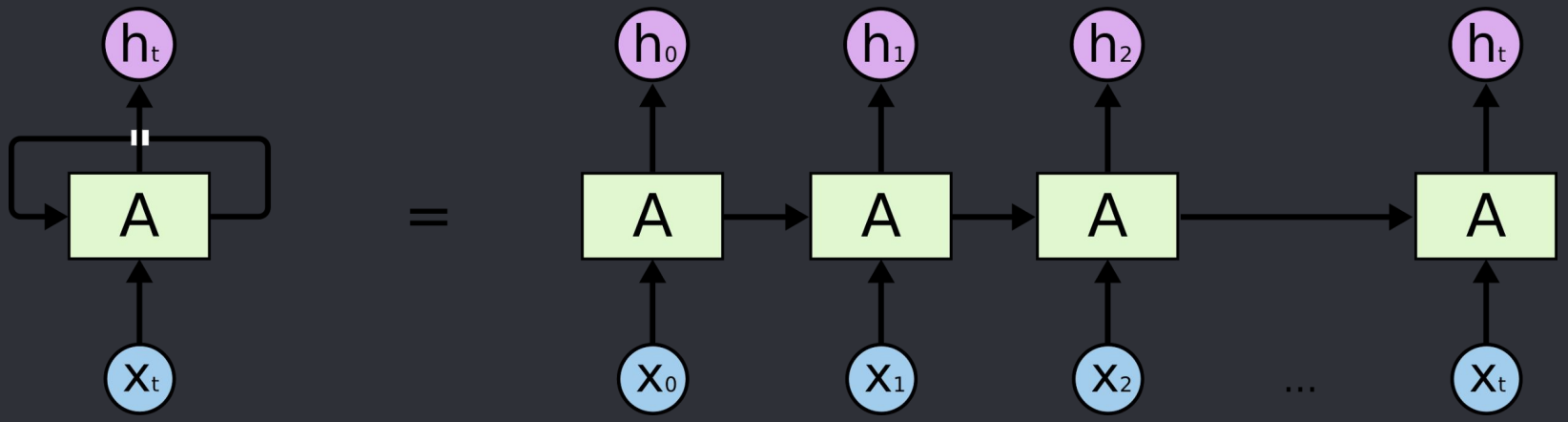


- Word embedding is about representing words in meaningful numerical vectors



https://www.doc.ic.ac.uk/~js4416/163/website/nlp/

Problem: RNN constrained by previous timestep computation



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

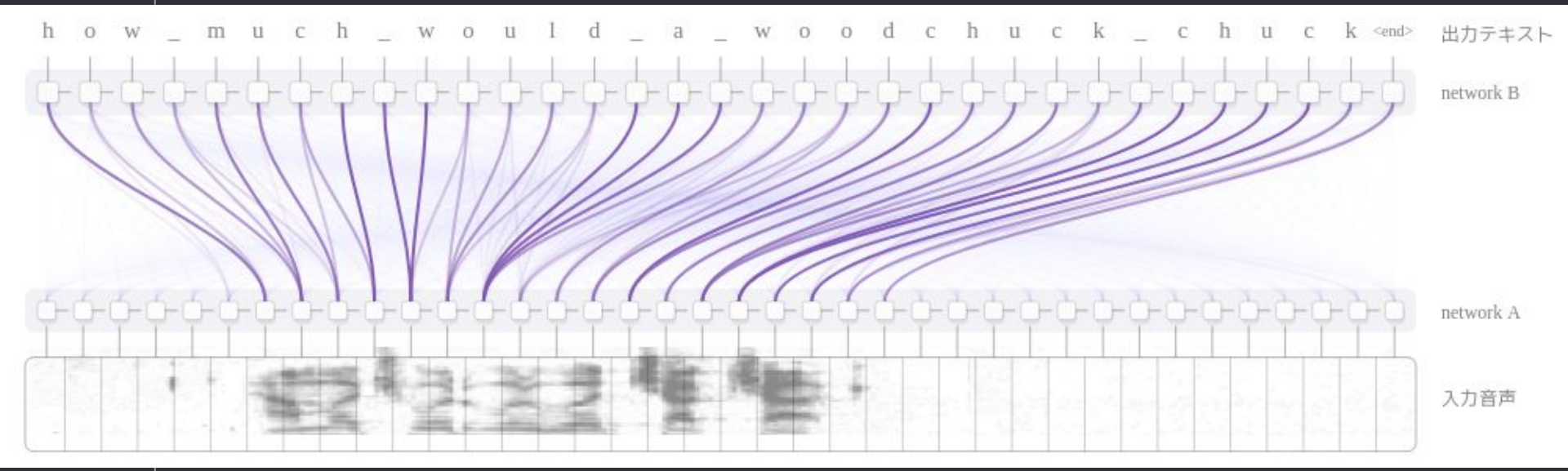
- Target is to Improve the perofmrance and get rid of sequential computation

Transformer

1

https://www.universalstudioshollywood.com/things-to-do/rides-and-attractions/transformers-the-ride-3d/

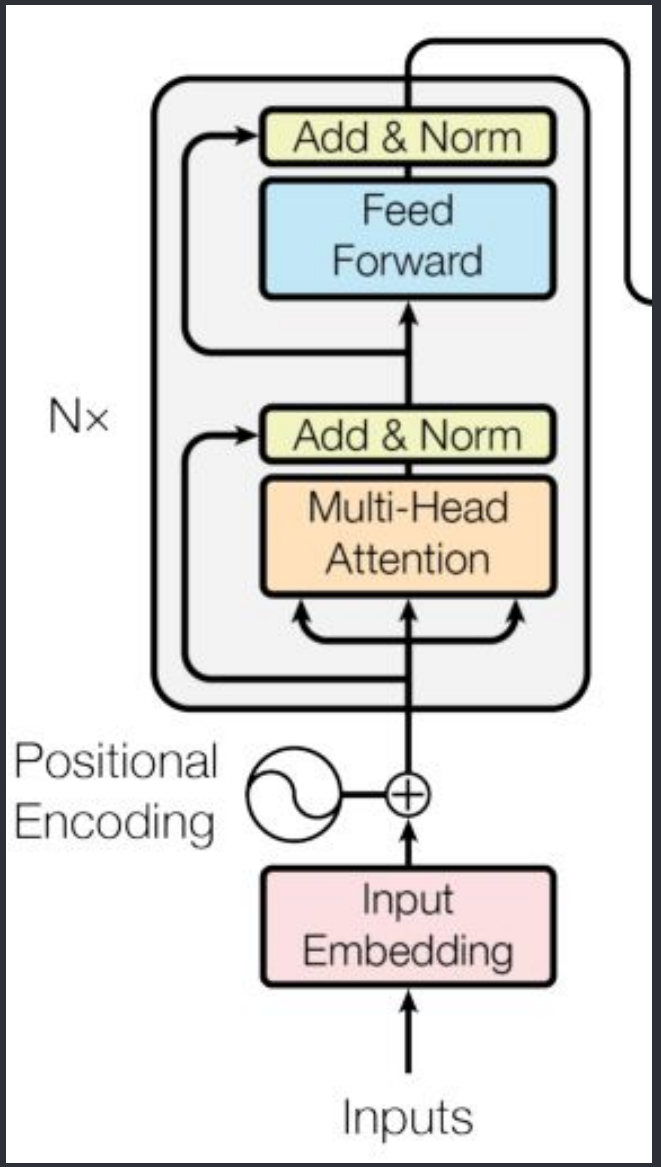
Self-Attention: focus on the important parts.



https://deepage.net/deep_learning/2017/03/03/attention-augmented-recurrent-neural-networks.html#attentionインターフェース

Model: Encoder

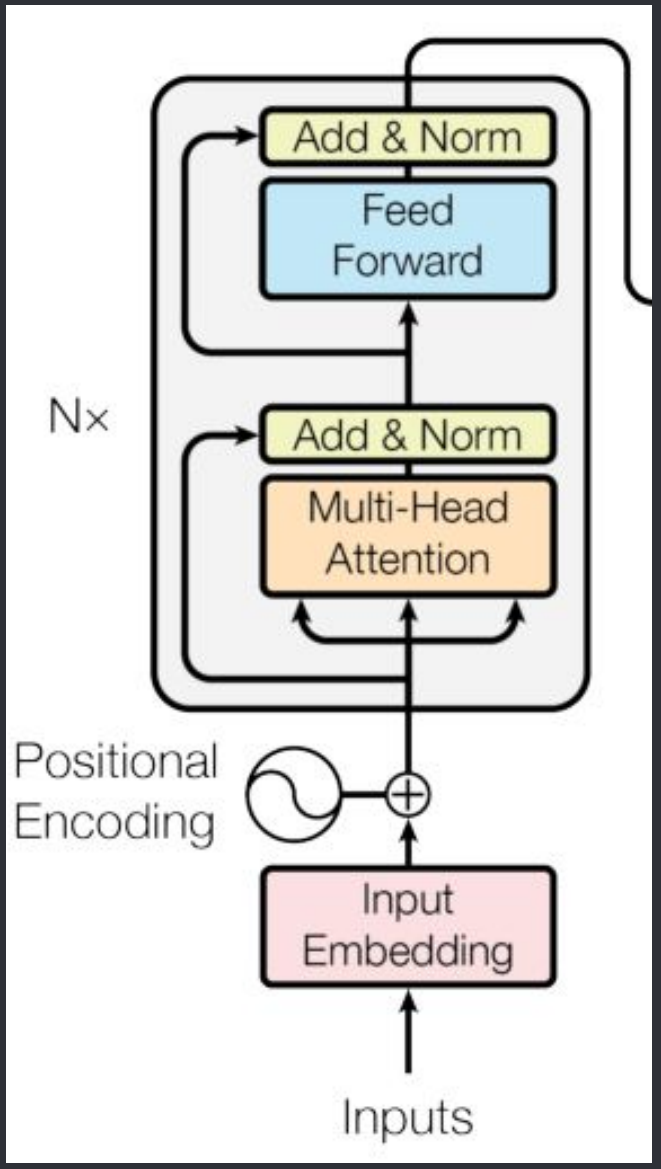
- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



https://arxiv.org/abs/1706.03762

Model: Encoder

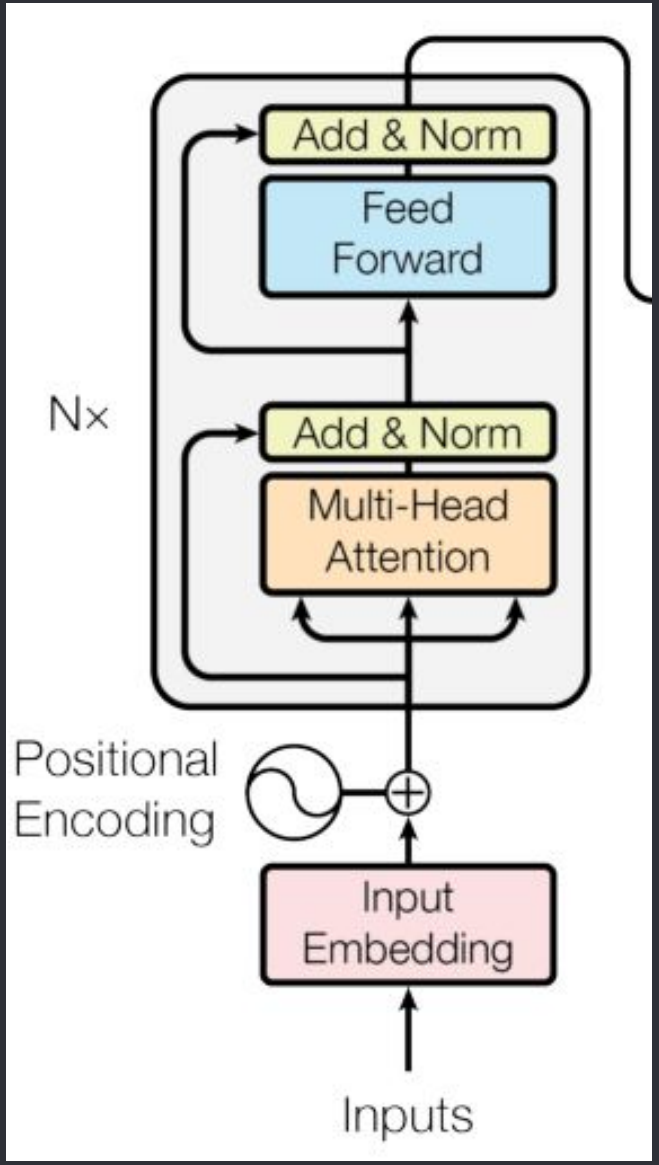
- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



https://arxiv.org/abs/1706.03762

Model: Encoder

- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



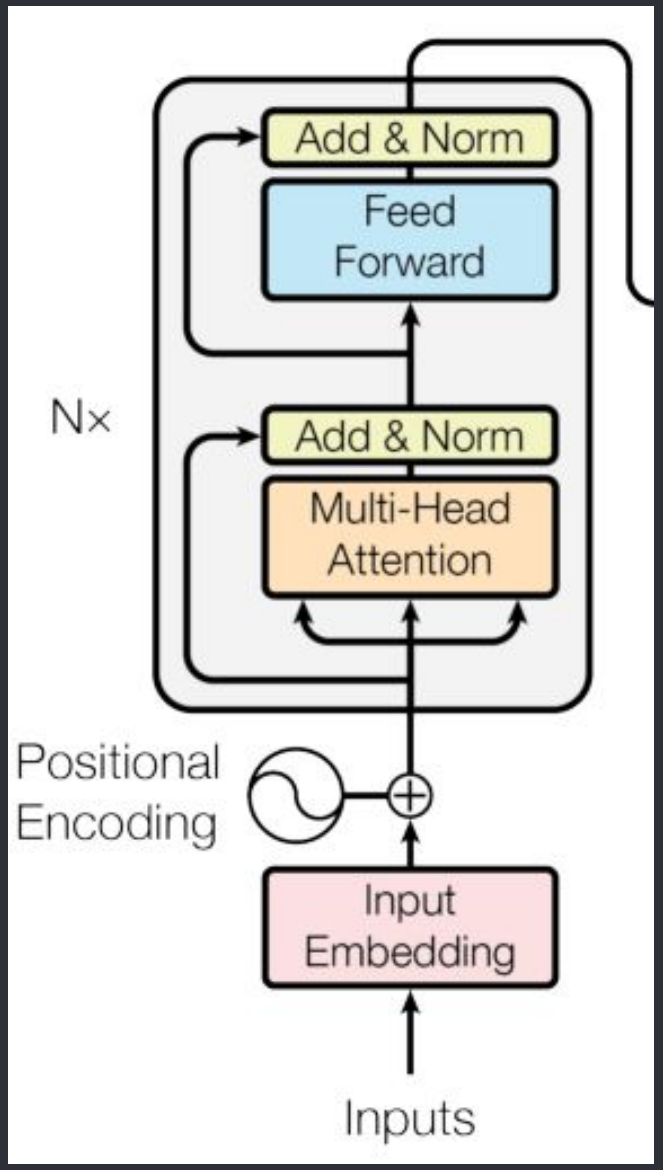
https://arxiv.org/abs/1706.03762

Model: Encoder

- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward

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

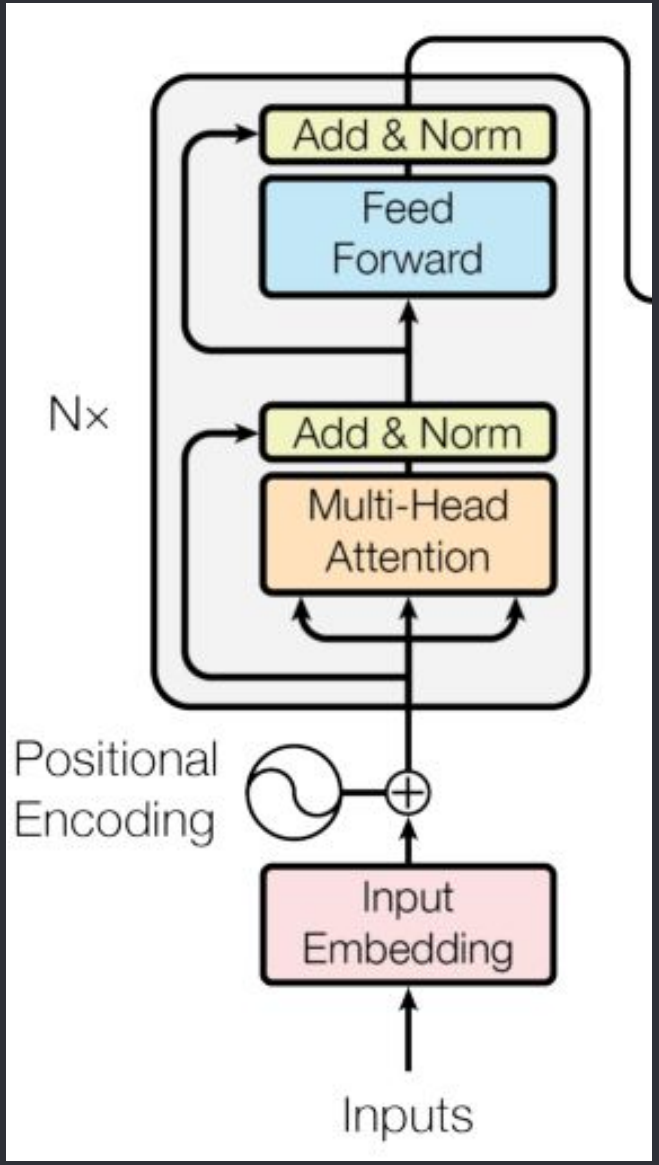
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$



https://arxiv.org/abs/1706.03762

Model: Encoder

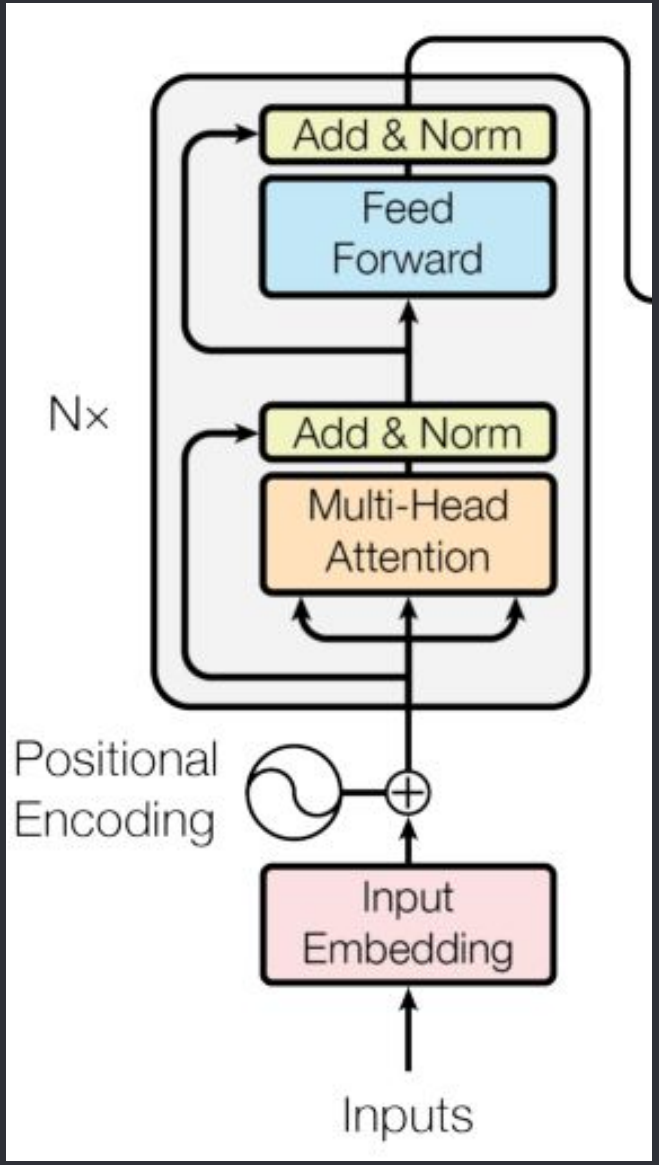
- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



https://arxiv.org/abs/1706.03762

Model: Encoder

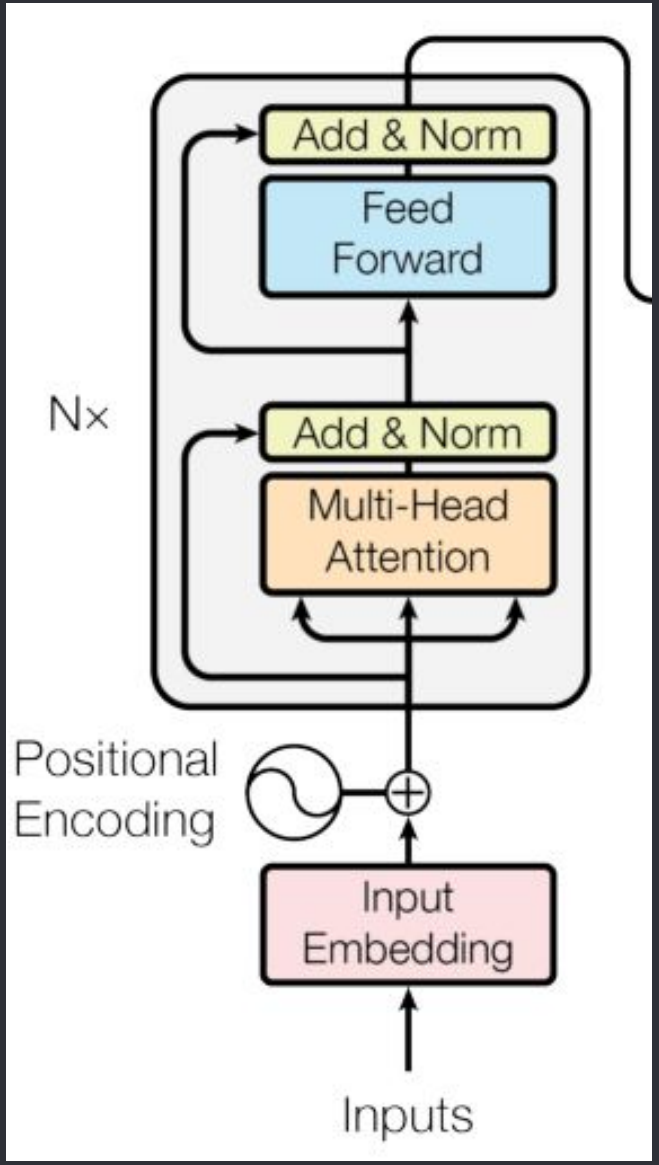
- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



https://arxiv.org/abs/1706.03762

Model: Encoder

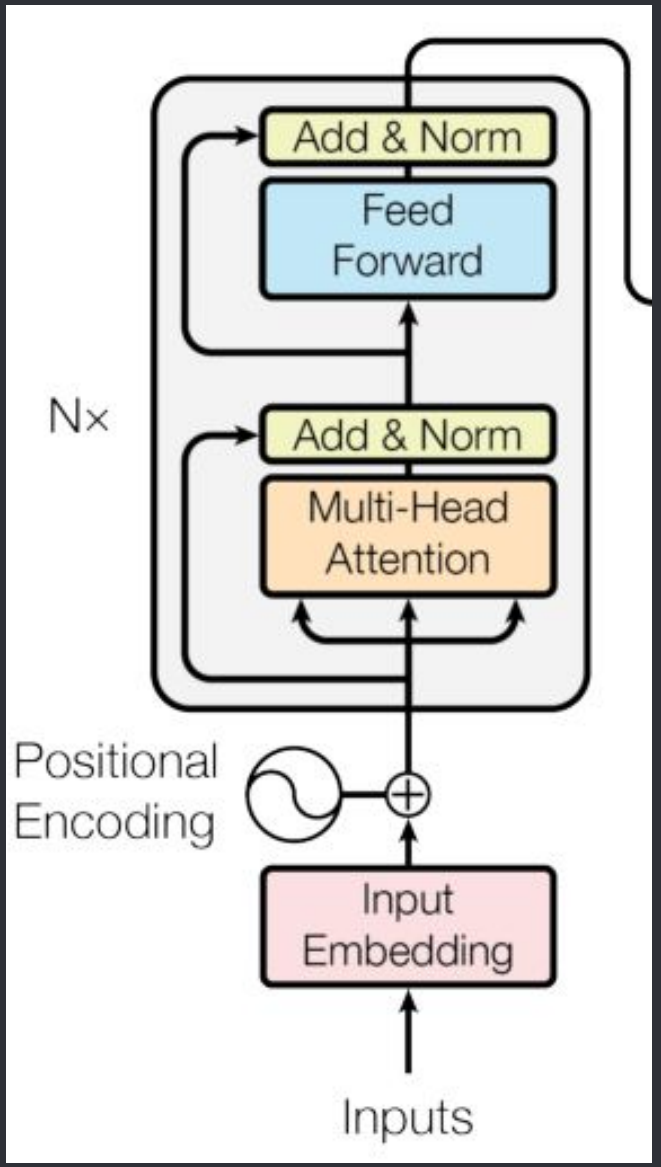
- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



https://arxiv.org/abs/1706.03762

Model: Encoder

- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward



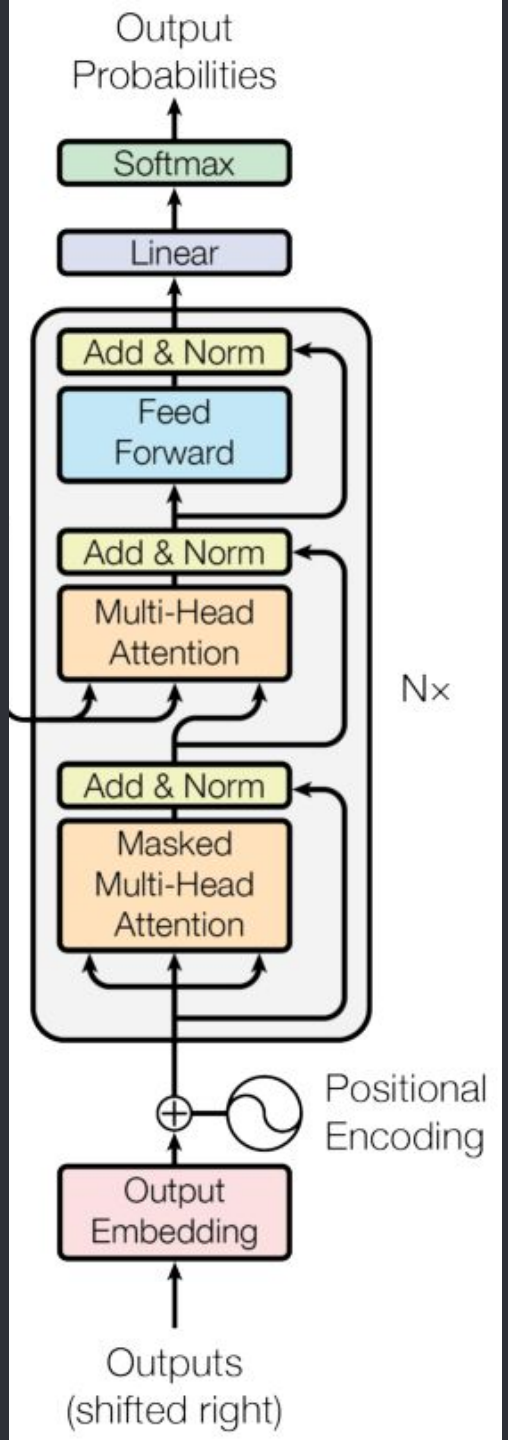
https://arxiv.org/abs/1706.03762

Model: Decoder

- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward
- Softmax

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

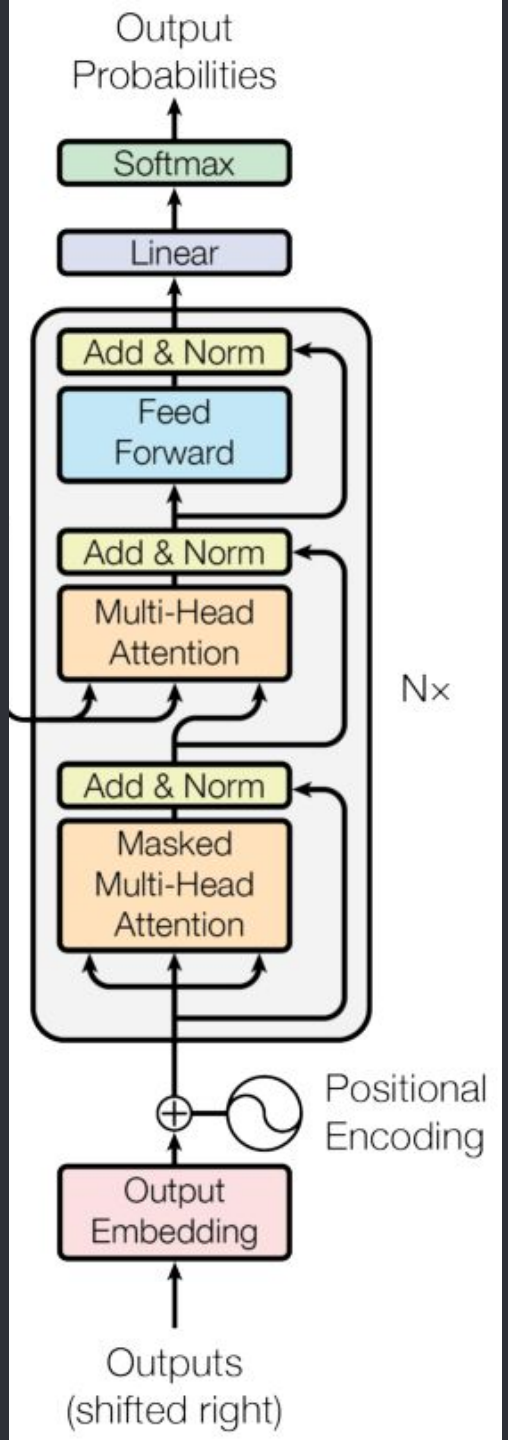
https://arxiv.org/abs/1706.03762



Model: Decoder

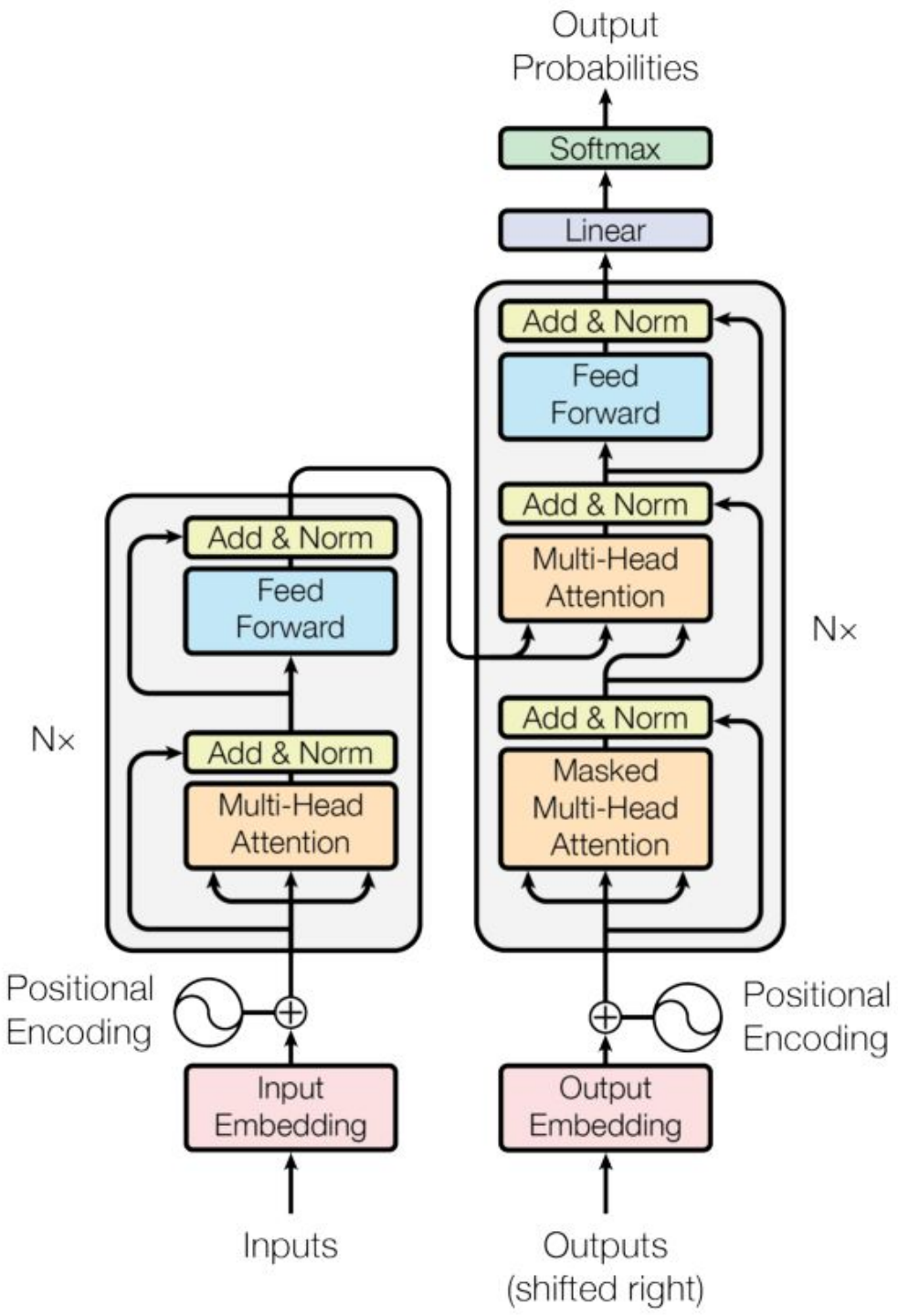
- N=6
- All Layers output size 512
- Embedding
- Positional Encoding
- Notice the Residual connection
- Multi-head Attention
- LayerNorm(x + Sublayer(x))
- Position wise feed forward
- Softmax

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

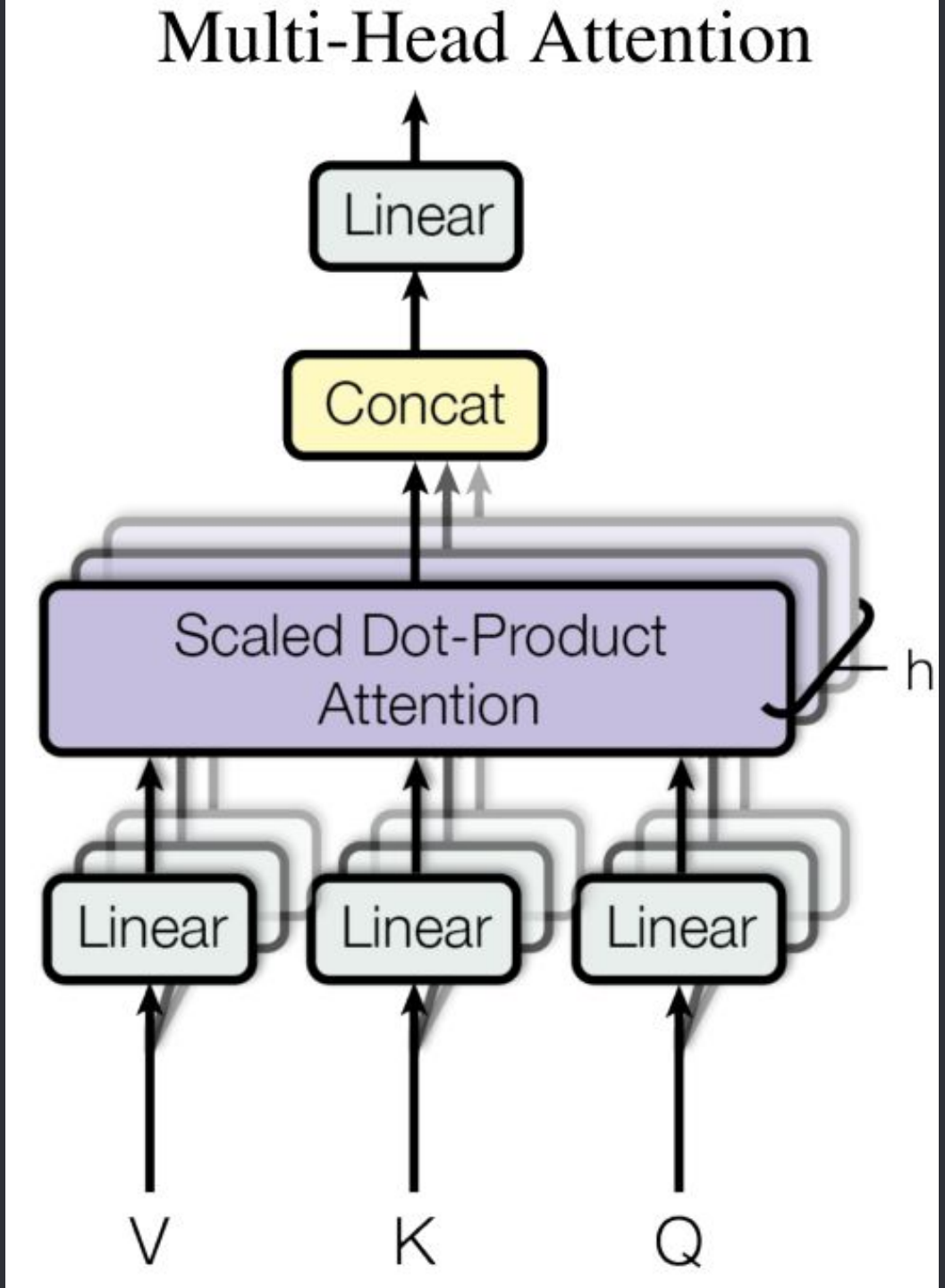


https://arxiv.org/abs/1706.03762

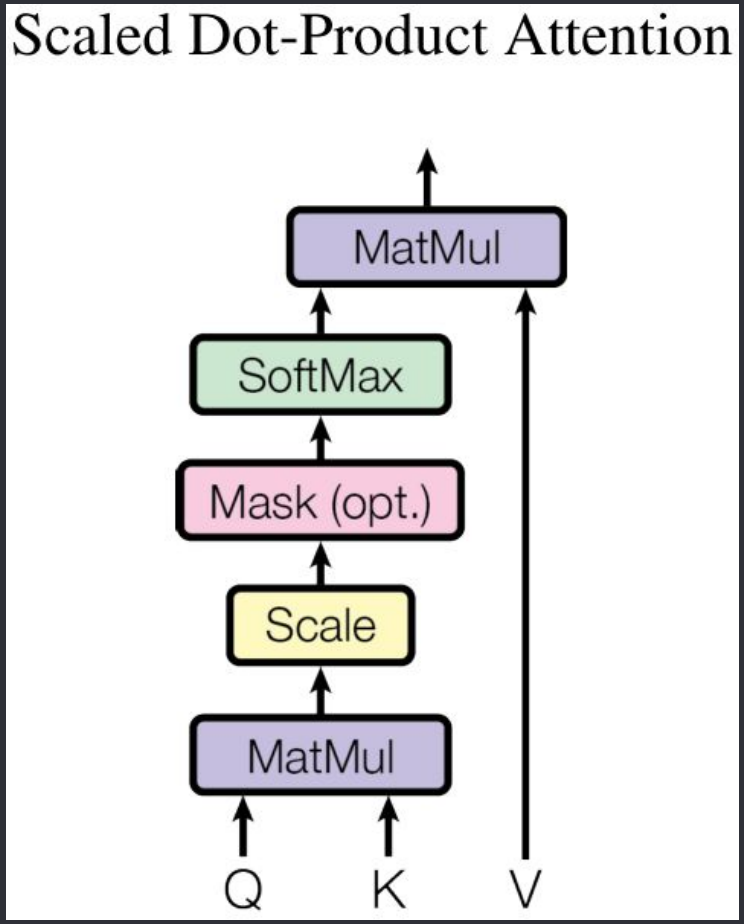
Model: Complete



https://arxiv.org/abs/1706.03762

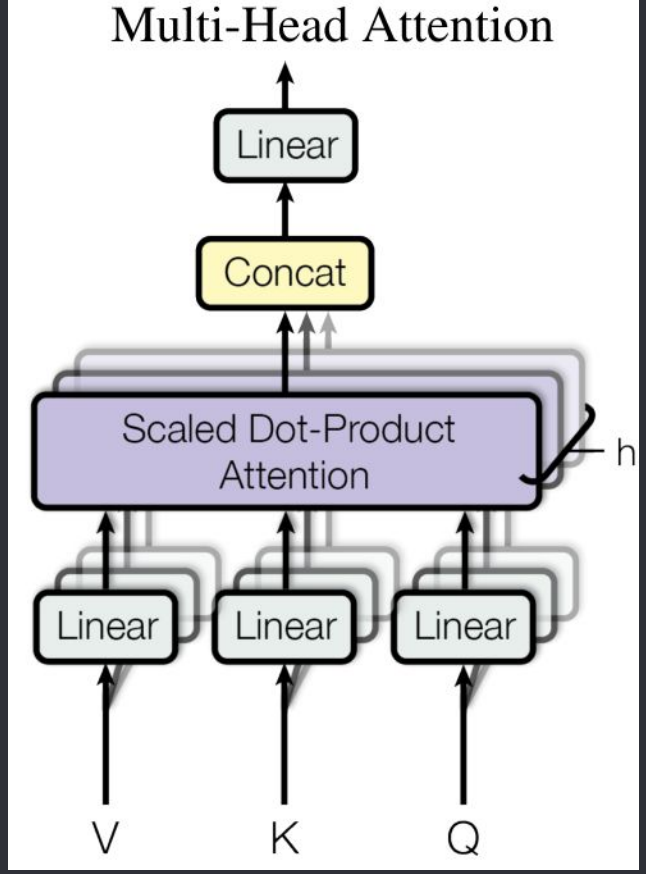


https://arxiv.org/abs/1706.03762



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

https://arxiv.org/abs/1706.03762



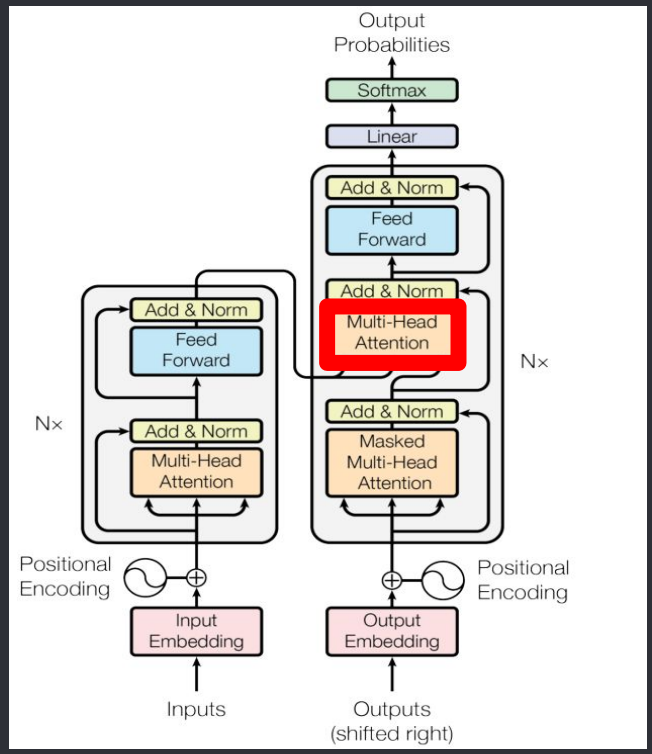
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$

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

https://arxiv.org/abs/1706.03762

Q,K,V

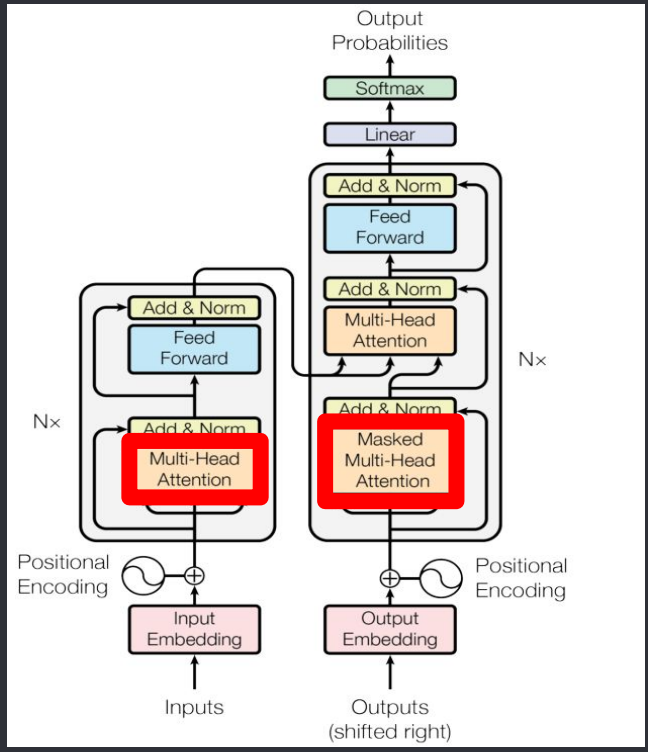
- "encoder-decoder attention" layers, the queries (Q) come from the previous decoder layer, and the memory keys (K) and values (V) come from the output of the encoder."
- Otherwise: all three come from previous layer (Hidden state)



https://www.reddit.com/r/MachineLearning/comments/6kc7py/d_where_does_the_query_keys_and_values_come_from/
https://arxiv.org/abs/1706.03762

Q,K,V

- "encoder-decoder attention" layers, the queries (Q) come from the previous decoder layer, and the memory keys (K) and values (V) come from the output of the encoder."
- Otherwise: all three come from previous layer (Hidden state)



https://www.reddit.com/r/MachineLearning/comments/6kc7py/d_where_does_the_query_keys_and_values_come_from/
https://arxiv.org/abs/1706.03762

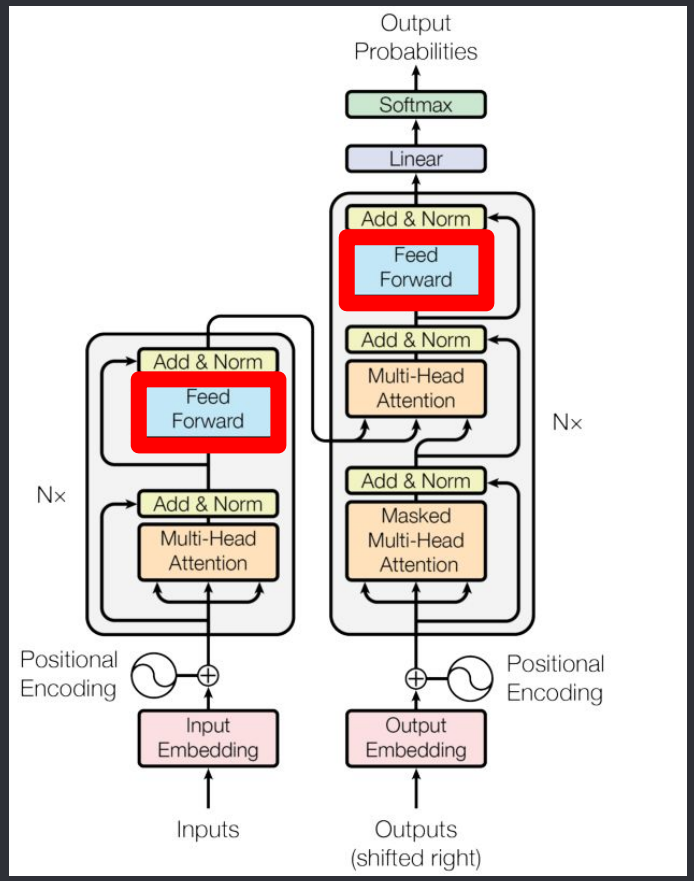
Complexity

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

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)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

https://arxiv.org/abs/1706.03762

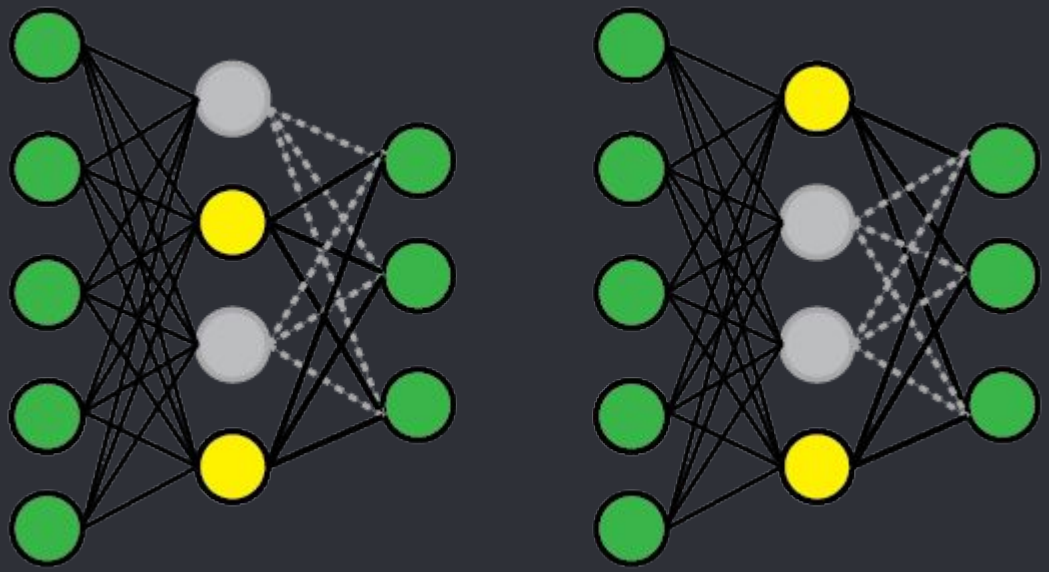
Position-wise Feed-Forward network



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

https://arxiv.org/abs/1706.03762

Dropout



- Used for Residual connections

https://stats.stackexchange.com/questions/201569/difference-between-dropout-and-dropconnect/201891

Training

- Data sets:
  - WMT 2014 English-German: 4.5 million sentences pairs with 37K tokens.
  - WMT 2014 English-French: 36M sentences, 32K tokens.
- Hardware:
  - 8 Nvidia P100 GPus (Base model 12 hours, big model 3.5 days)

http://www.prioritiesusa.org/recommendations-for-sports-training/

Training

- Data sets:
  - WMT 2014 English-German: 4.5 million sentences pairs with 37K tokens.
  - WMT 2014 English-French: 36M sentences, 32K tokens.
- Hardware:
  - 8 Nvidia P100 GPus (Base model 12 hours, big model 3.5 days)

http://www.prioritiesusa.org/recommendations-for-sports-training/

Optimizing the parameters

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
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

https://arxiv.org/abs/1706.03762

Results

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

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**	$2.3 \cdot 10^{19}$	

https://arxiv.org/abs/1706.03762

Results

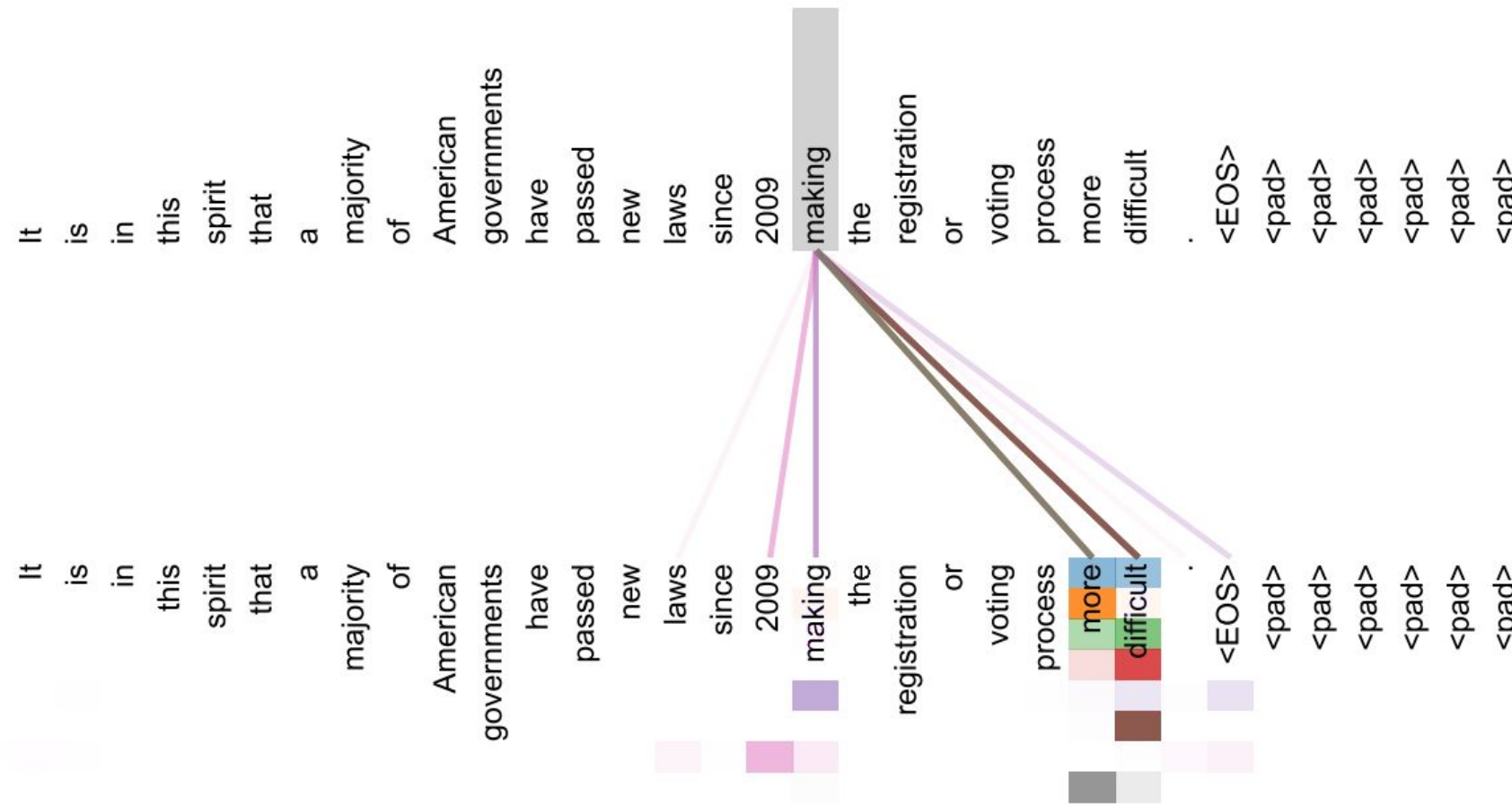


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.

https://arxiv.org/abs/1706.03762

Results

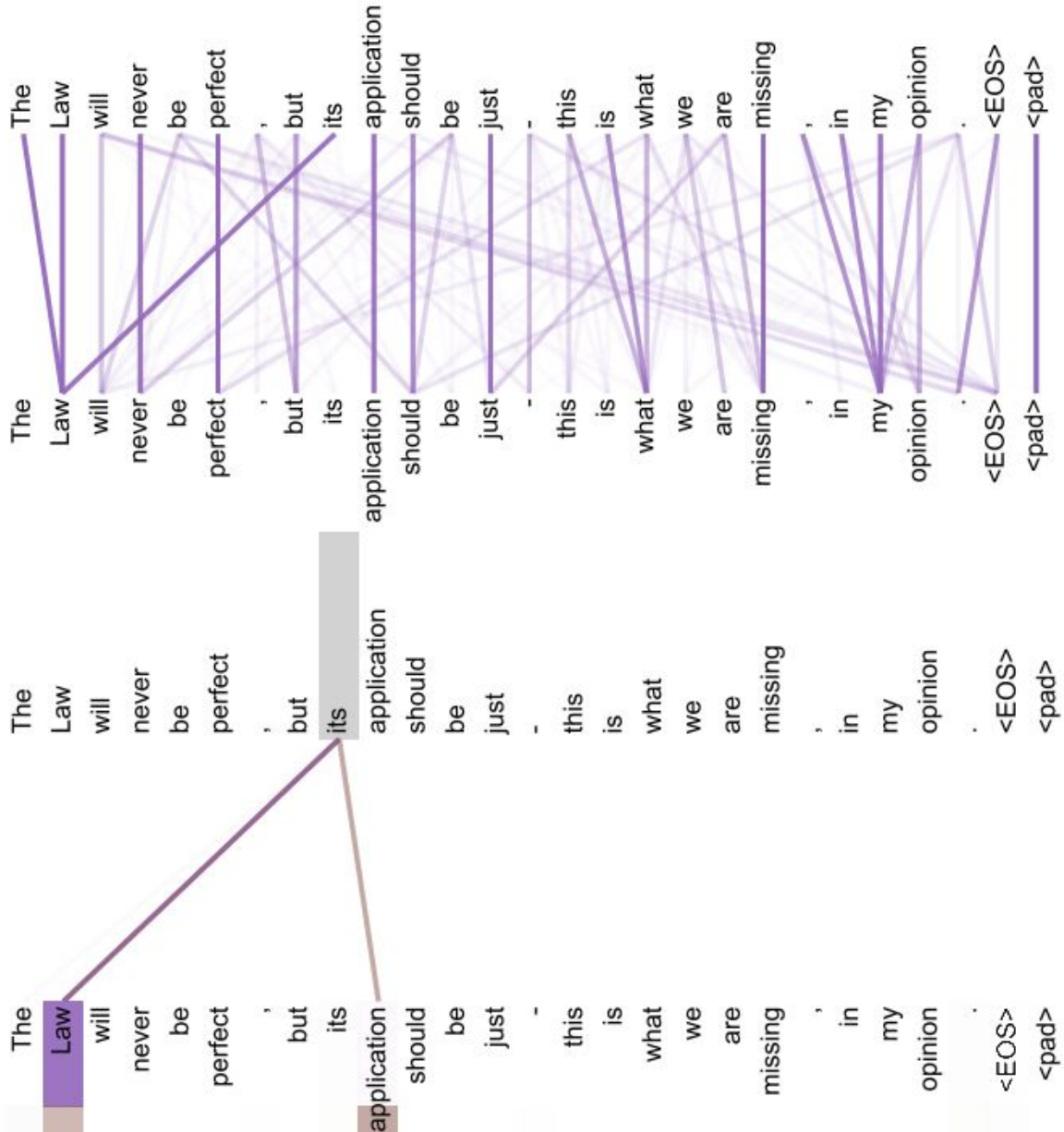
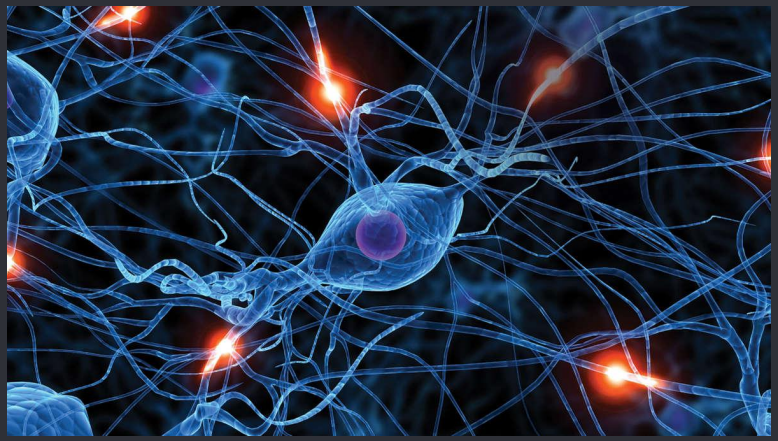


Figure 4: Two attention heads, also in layer 5 of 6, apparently involved in anaphora resolution. Top
Full attentions for head 5. Bottom: Isolated attentions from just the word 'its' for attention heads 5
and 6. Note that the attentions are very sharp for this word.

https://arxiv.org/abs/1706.03762

How is this paper linked to neuroscience ??

?



https://www.astrazeneca.com/our-focus-areas/neuroscience.html
http://blog.webcertain.com/machine-translation-technology-the-search-engine-takeover/18/02/2015/


THANKS FOR
LISTENING
ANY QUESTIONS?
NO?
GREAT!
KeepCalmAndPosters.com