Attention Is All You Need

**Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,
Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin**

[31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.]

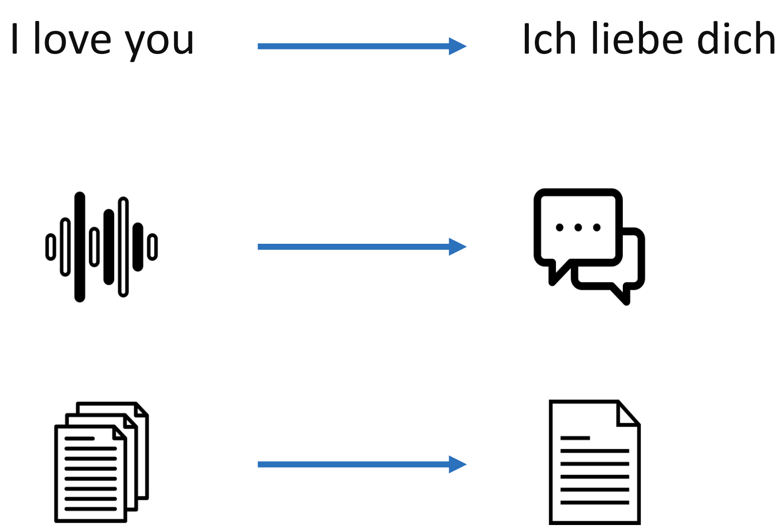
2024년 8월 13일

이규원

Data Mining & Intelligence Systems Lab
Department of Computer Science and Engineering
Chung-Ang University

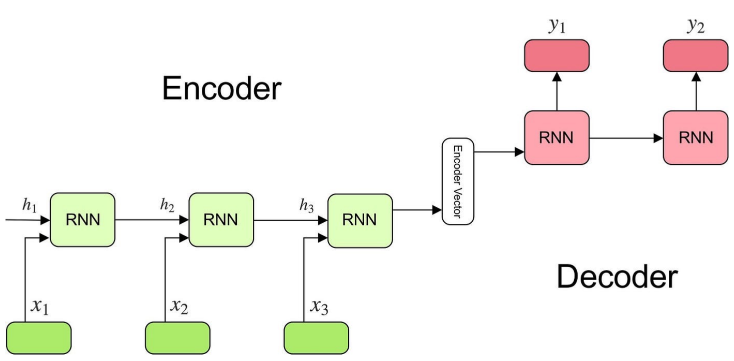
Background

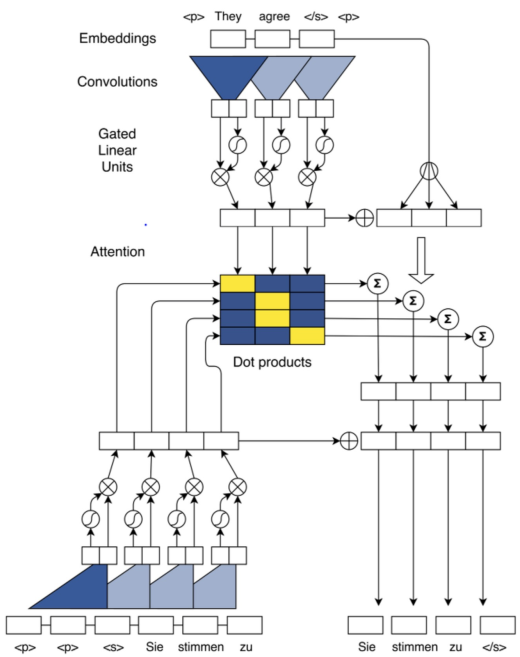
- **Sequence transduction**
    - Translation, Speech to Text, Summarize … etc.



Background

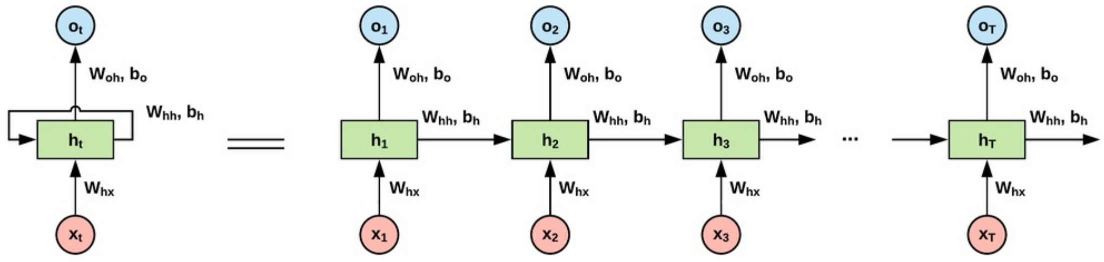
- **Sequence transduction models**
  - RNN(Recurrent Neural Network) base model
  - Convolution base model



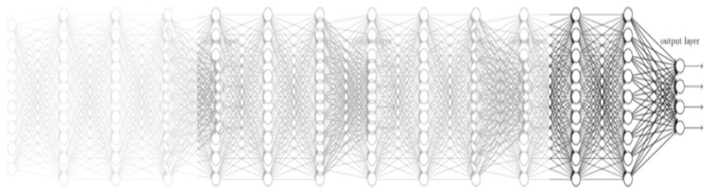


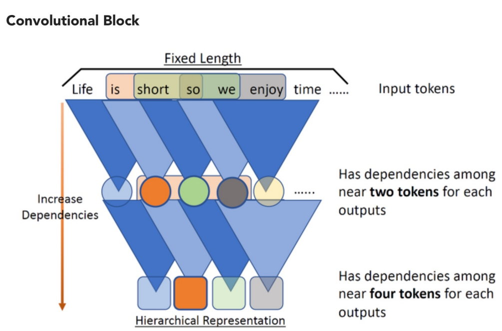
Motivation

- **Limitation of existing models**
  - Hard to parallelize



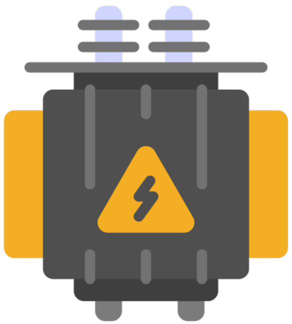
  - Difficult to learn dependencies between distant positions
    - Gradient vanishing problem
    - Increased computational complexity with positional distance





Motivation

- **Purpose**
  - Parallelize sequence transduction model
  - Learn dependencies between distant positions(faster)



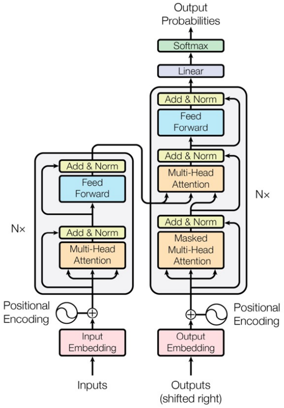
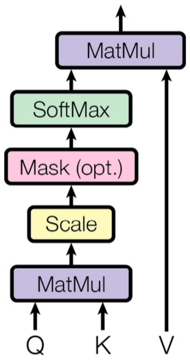


Figure 1: The Transformer - model architecture.

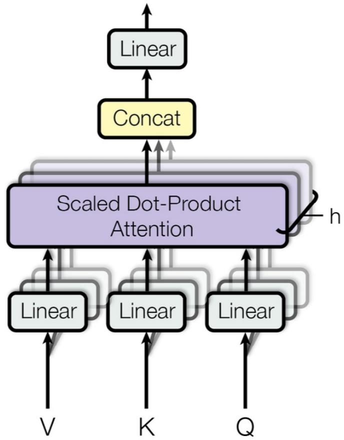
Proposed Model

- **Transformer**
    - Without RNN, Convolution
    - Self Attention
    - Multi-Head Attention
    - Positional Embedding

Scaled Dot-Product Attention



Multi-Head Attention



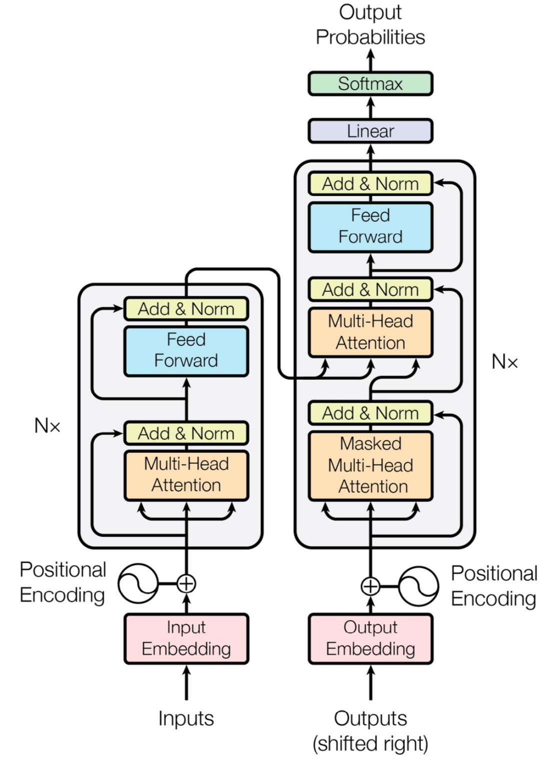
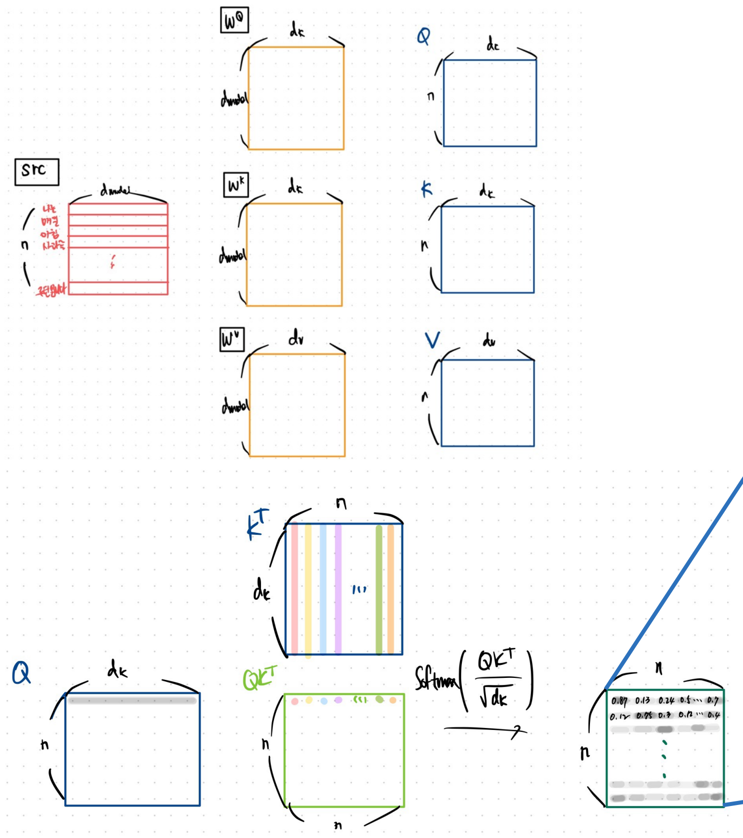
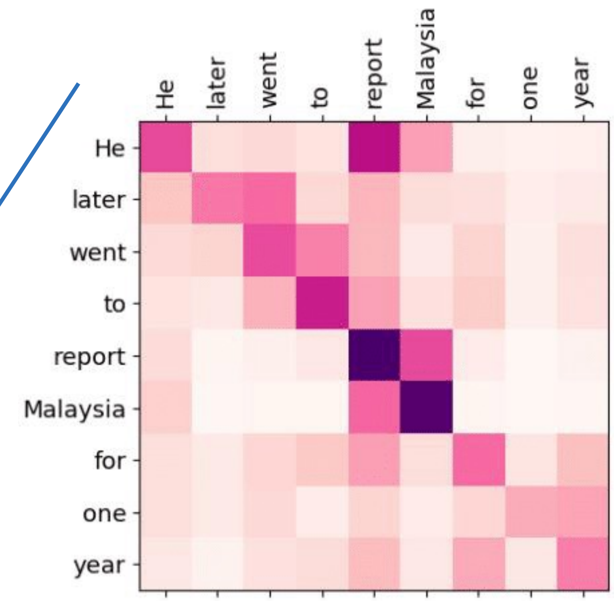


Figure 1: The Transformer - model architecture.

Transformer Model Architecture

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





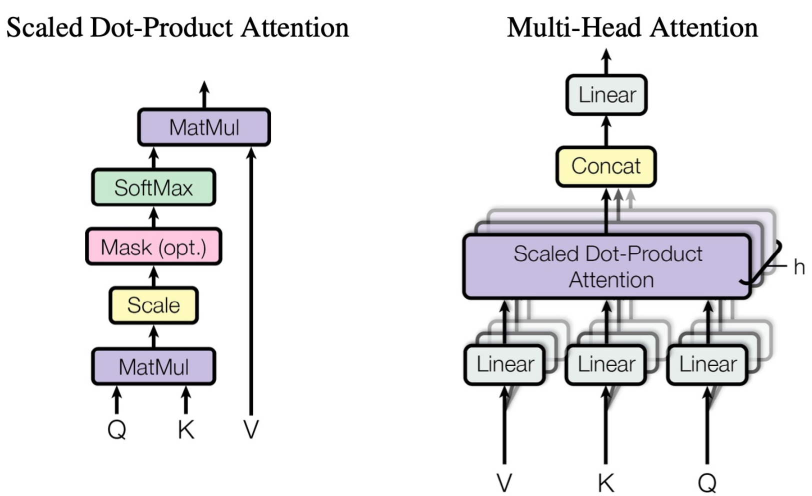
(self) attention score visualization

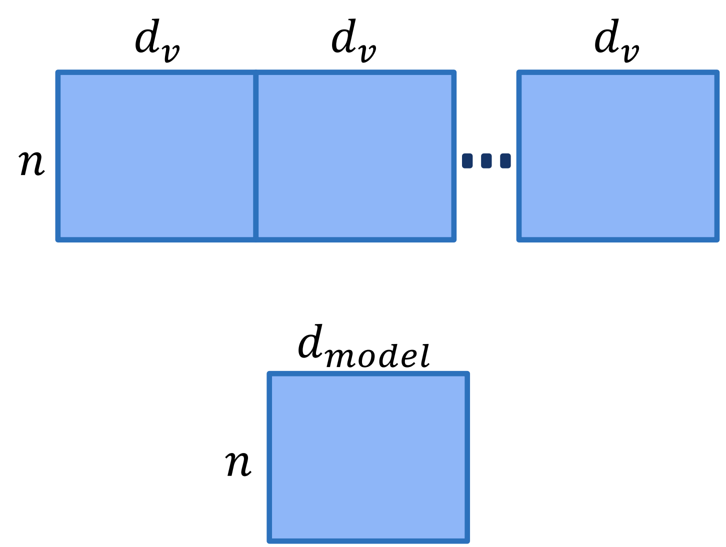
Du, S.; Wang, H. Addressing Syntax-Based Semantic Complementation: Incorporating Entity and Soft Dependency Constraints into Metonymy Resolution. *Future Internet* **2022**, *14*, 85. https://doi.org/10.3390/ fi14030085

Transformer Model Architecture

- **Multi-Head Attention**
  - $MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O$

$$where\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$





Transformer Model Architecture

- **Encoder Layer**
  - Positional Encoding
  - Self Attention
  - Feed Forward
  - Residual Connection

$$LayerNorm(x + Sublayer(x))$$

Self Attention

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

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

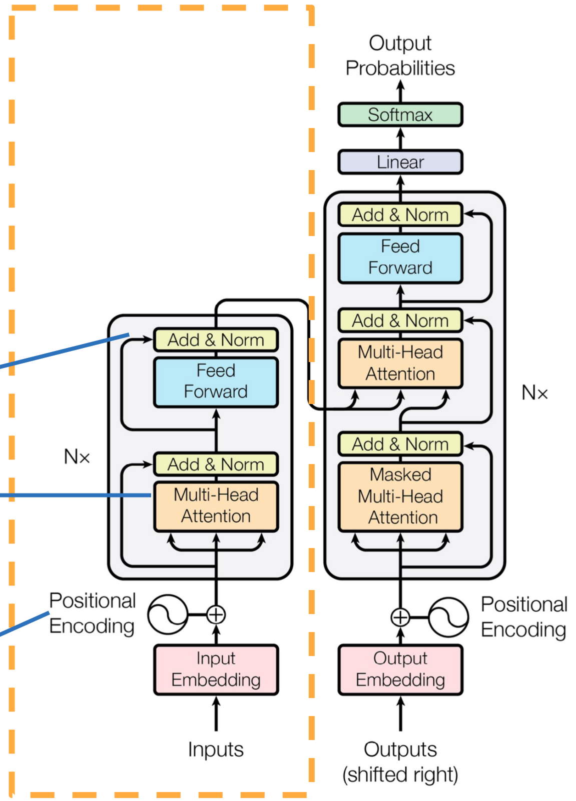


Figure 1: The Transformer - model architecture.

Transformer Model Architecture

- **Decoder Layer**
  - Positional Encoding
  - Masked Self Attention
  - Encoder Decoder Attention
  - Feed Forward
  - Residual Connection

Linear weights =
input/output embedding weights^T

Encoder Decoder
Attention

Masked Self Attention

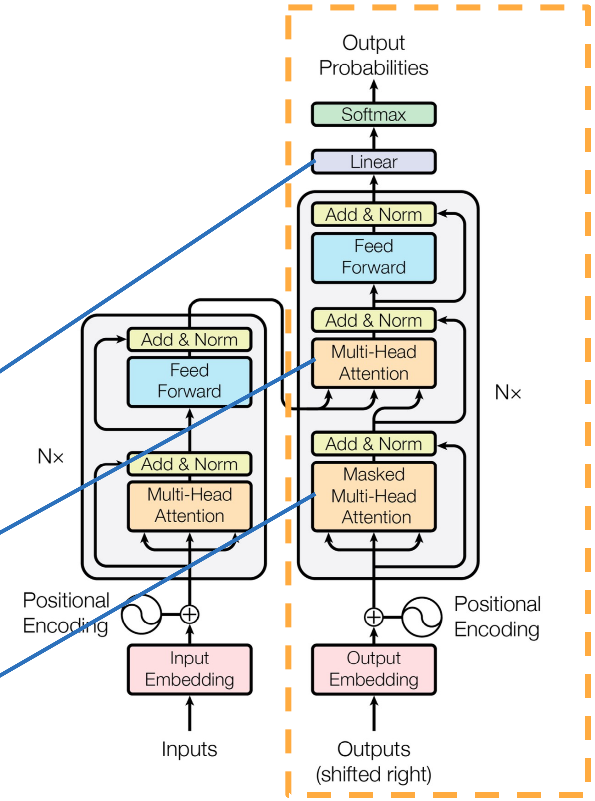
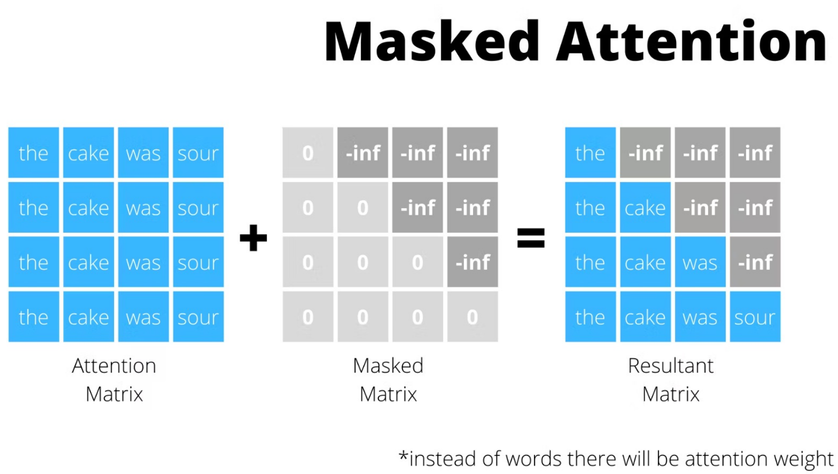


Figure 1: The Transformer - model architecture.

Transformer Model Architecture

- **Decoder Layer**
  - Positional Encoding
  - Masked Self Attention
  - Encoder Decoder Attention
  - Feed Forward
  - Residual Connection



Masked Self Attention

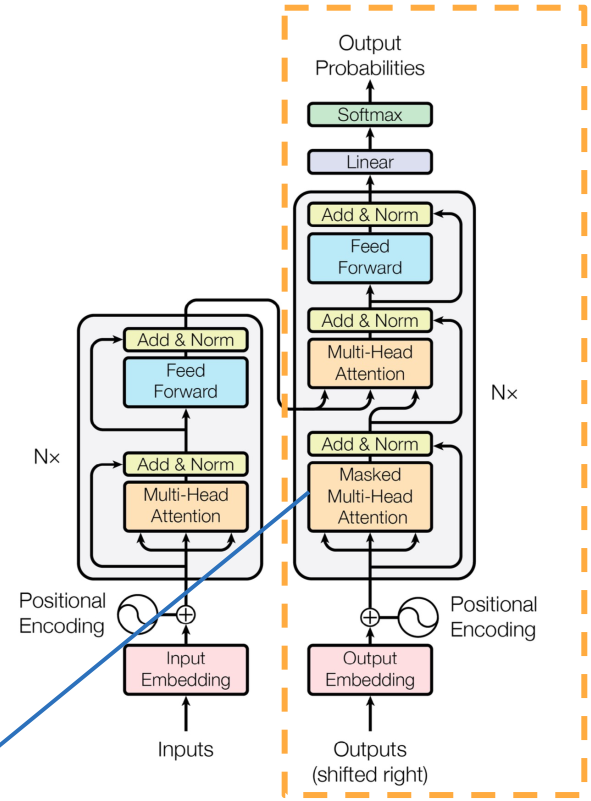
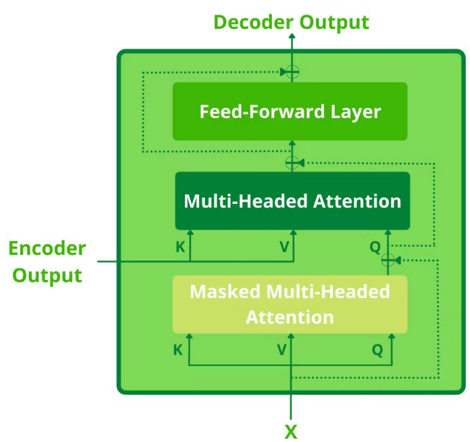


Figure 1: The Transformer - model architecture.

Transformer Model Architecture

- **Decoder Layer**
  - Positional Encoding
  - Masked Self Attention
  - Encoder Decoder Attention
  - Feed Forward
  - Residual Connection

Encoder Decoder Attention



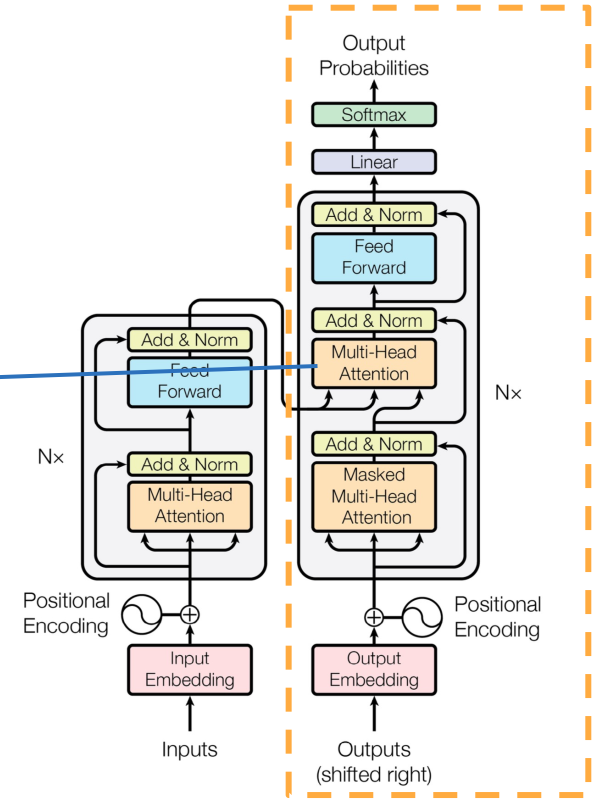


Figure 1: The Transformer - model architecture.

https://krypticmouse.hashnode.dev/attention-is-all-you-need

Experimental Results

- **Translation quality and training costs**

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 [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	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

BLEU: Bilingual Language Evaluation Understudy (The higher the better)

FLOPs: Floating point operations (The lower the better)

Experimental Results

- **Variations on the Transformer architecture**

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

PPL: Perplexity(the lower the better) / 헷갈리는 정도

ϵ_{ls}: Label smoothing epsilon / 정답에 1 대신 $1-\epsilon$

Experimental Results

- **English constituency parsing(영어 구성 구문 분석)**

Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ)

Parser	Training	WSJ 23 F1
Vinyals & Kaiser el al. (2014) [37]	WSJ only, discriminative	88.3
Petrov et al. (2006) [29]	WSJ only, discriminative	90.4
Zhu et al. (2013) [40]	WSJ only, discriminative	90.4
Dyer et al. (2016) [8]	WSJ only, discriminative	91.7
Transformer (4 layers)	WSJ only, discriminative	91.3
Zhu et al. (2013) [40]	semi-supervised	91.3
Huang & Harper (2009) [14]	semi-supervised	91.3
McClosky et al. (2006) [26]	semi-supervised	92.1
Vinyals & Kaiser el al. (2014) [37]	semi-supervised	92.1
Transformer (4 layers)	semi-supervised	92.7
Luong et al. (2015) [23]	multi-task	93.0
Dyer et al. (2016) [8]	generative	$\boxed{93.3}$

RNN base

Conclusion

- **Existing Problems**
  - Hard to Parallelize
  - Difficult to Learn Dependencies Between Distant Positions
- **Proposed Model**
  - Without Recurrent, Convolution Layers
  - Multi-Head Self-Attention
  - Positional Encoding
- **Experiments**
  - WMT 2014 Performance: BLEU 점수에서 기존 모델을 능가 / 적은 연산
  - English Constituency Parsing: 이전 모델 대부분을 능가