Attention is All You Need

DeepSync, South Korea

Transformer 구조

Deep Sync

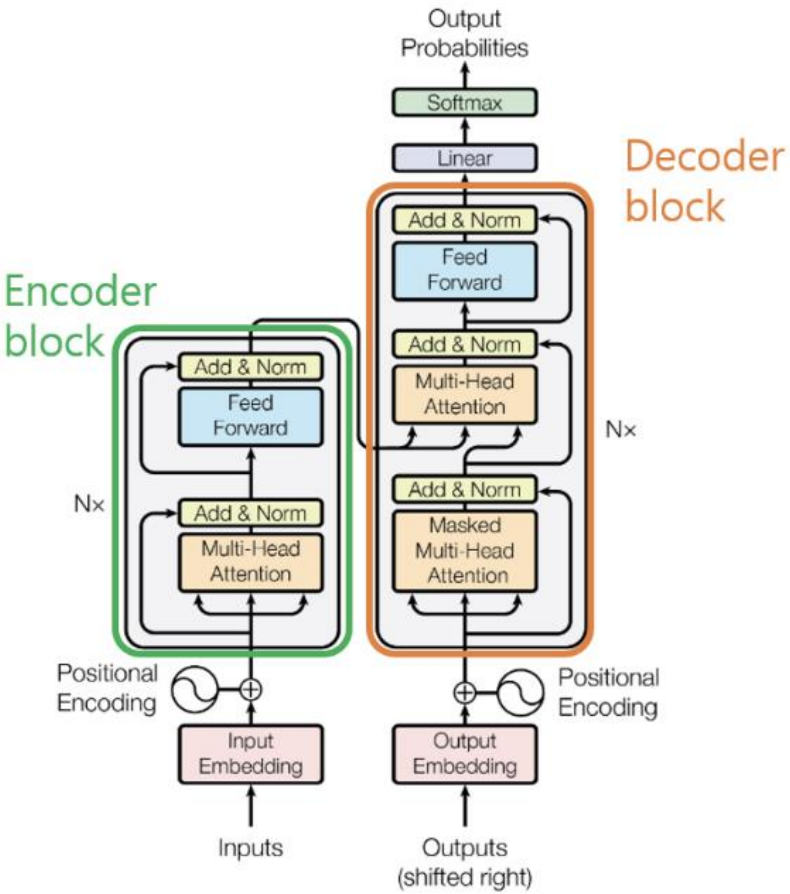
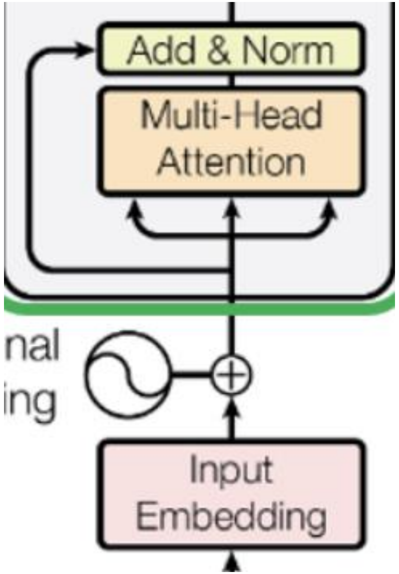
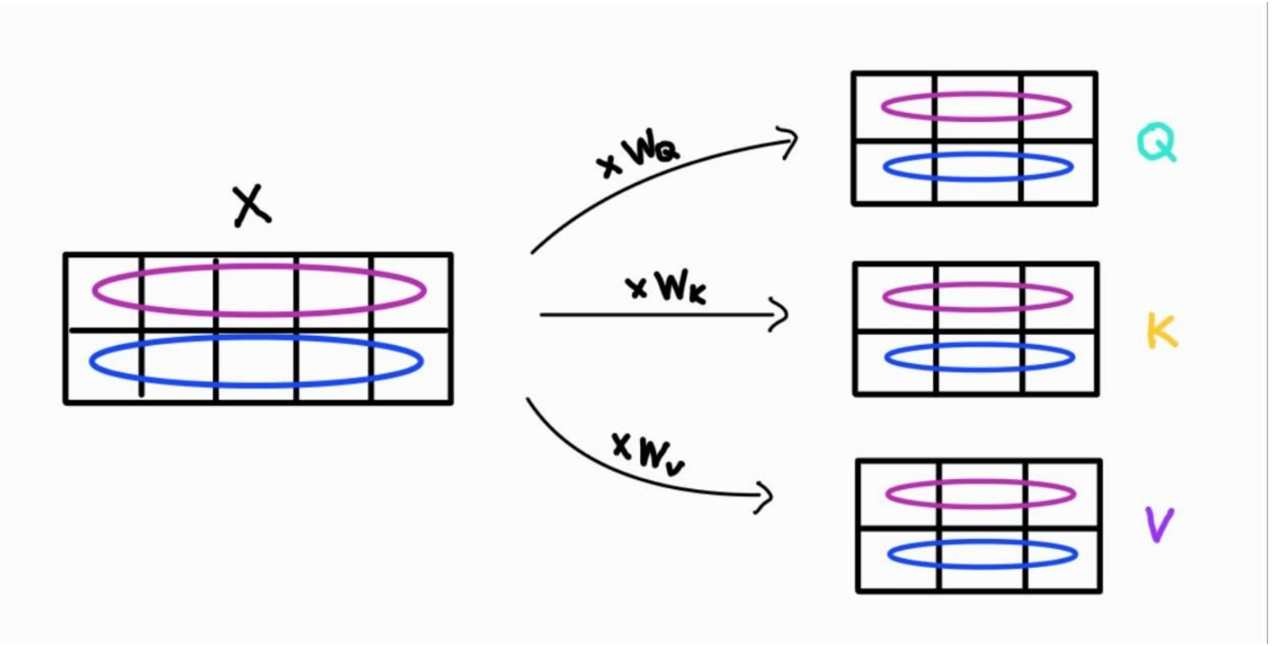


Figure 1: The Transformer - model architecture.

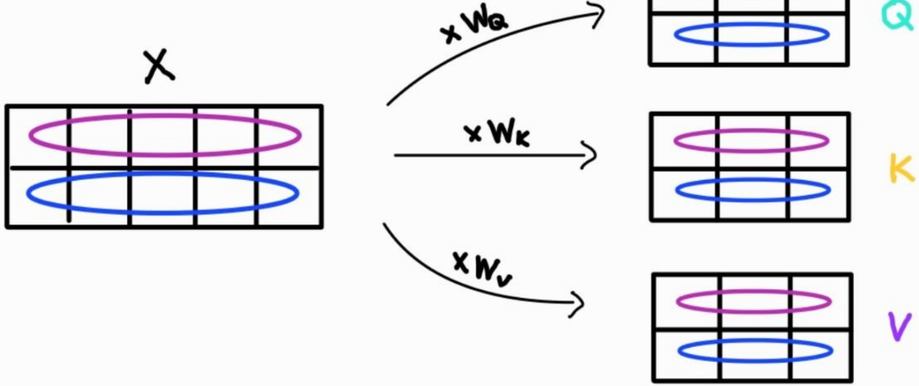
1. Multi-Head Attention





W_Q, W_K, W_V 학습

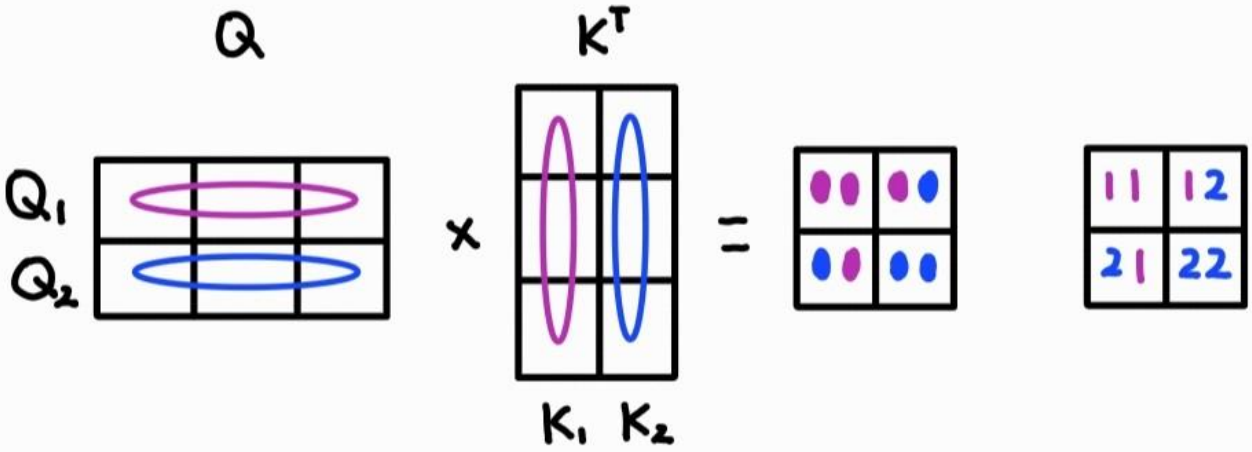
1. Multi-Head Attention



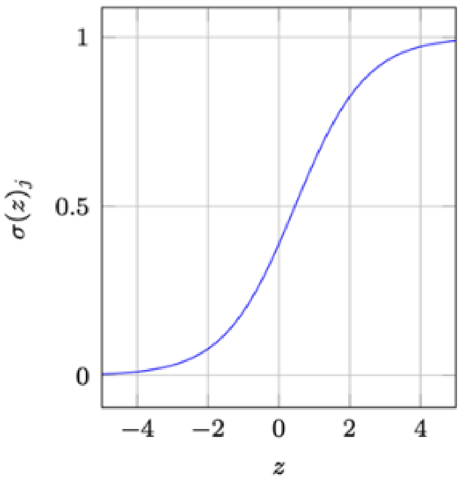
- Input vectors: X (Shape: $N_X \times D_X$)
- key matrix: W_K (Shape: $D_X \times D_K$)
- Value matrix: W_V (Shape: $D_X \times D_V$)
- Query matrix: W_Q (Shape: $D_X \times D_Q$)

- Query vectors: $Q = XW_Q$ (Shape: $N_X \times D_Q$)
- key vectors: $K = XW_K$ (Shape: $N_X \times D_K$)
- Value vectors: $V = XW_V$ (Shape: $N_X \times D_V$)
- Similarities: $E = QK^T$ (Shape: $N_X \times N_X$)

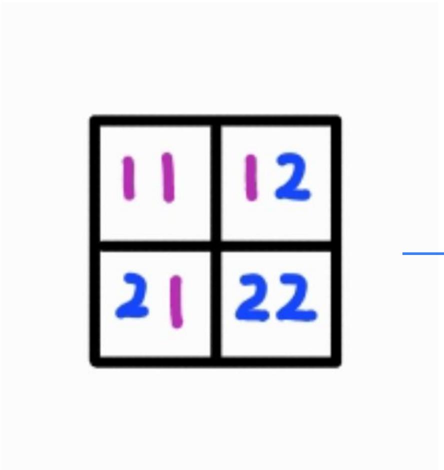
1. Multi-Head Attention(self attention)



1. Multi-Head Attention(self attention)

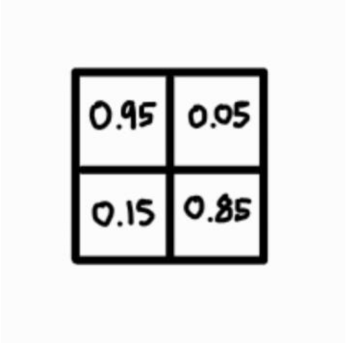


(b) Softmax activation function.

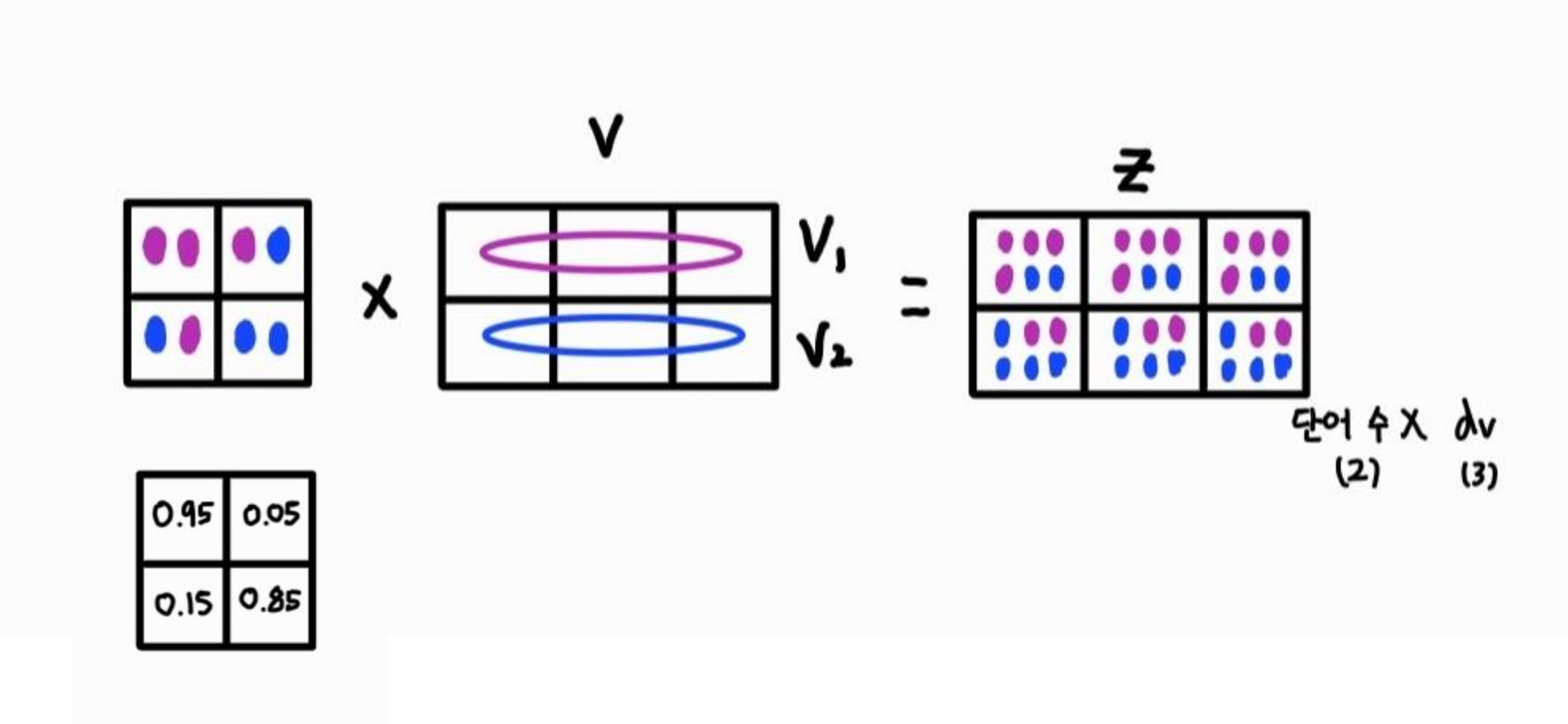


d_K 로 나눠 줌

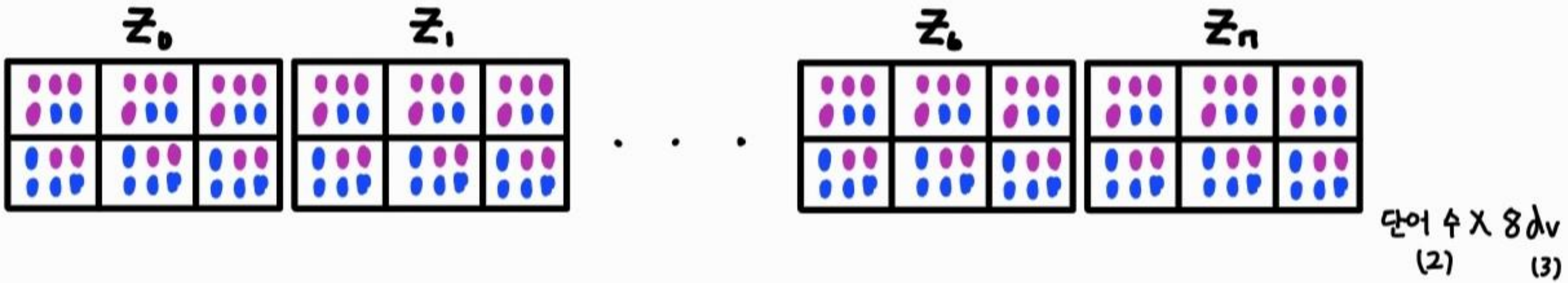
Softmax 통과



1. Multi-Head Attention(self attention)

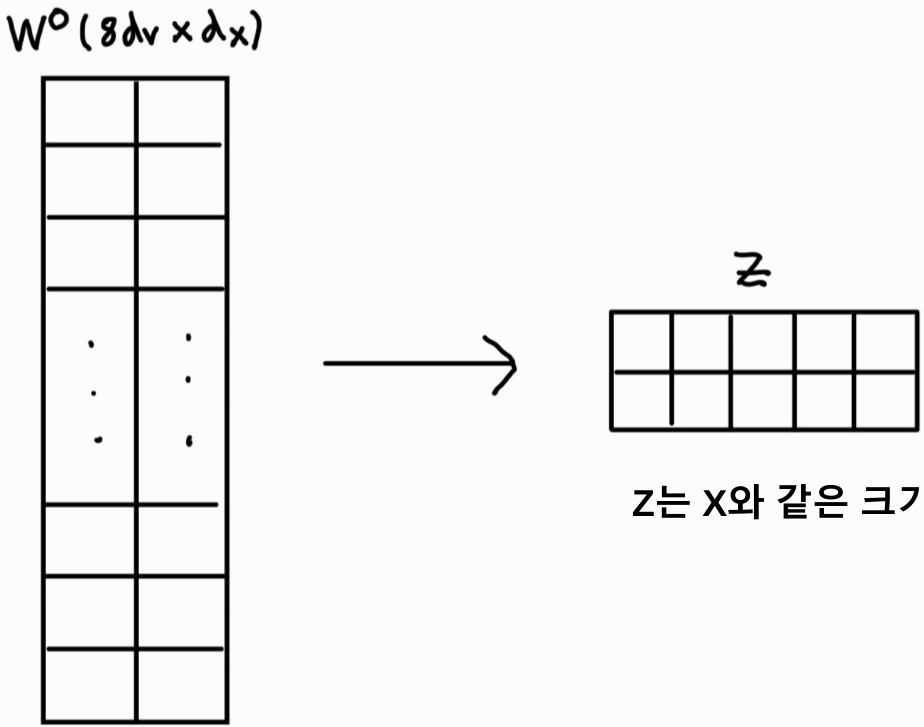


1. Multi-Head Attention(self attention)



이런 과정을 8번 반복(Multi-head)

1. Multi-Head Attention(self attention)



Z는 X와 같은 크기

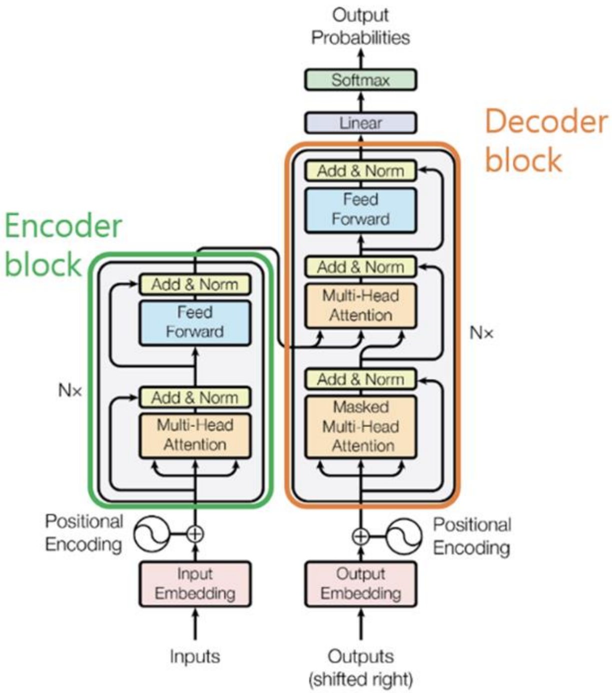
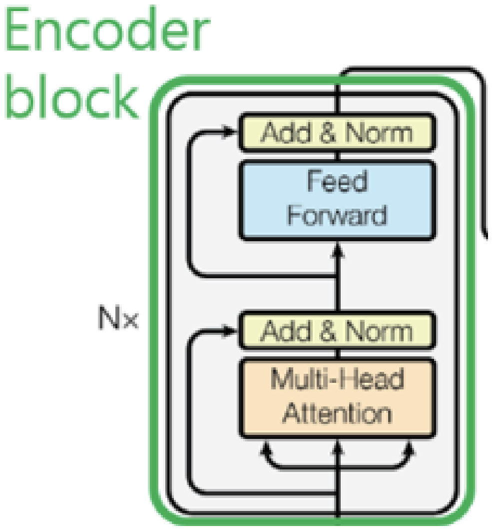


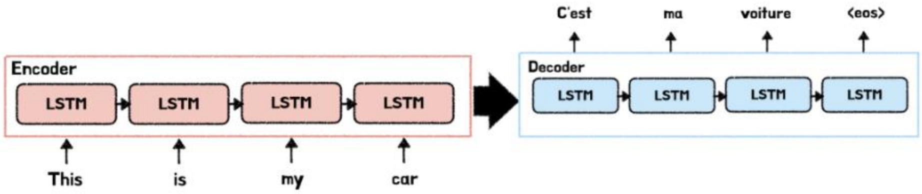
Figure 1: The Transformer - model architecture.

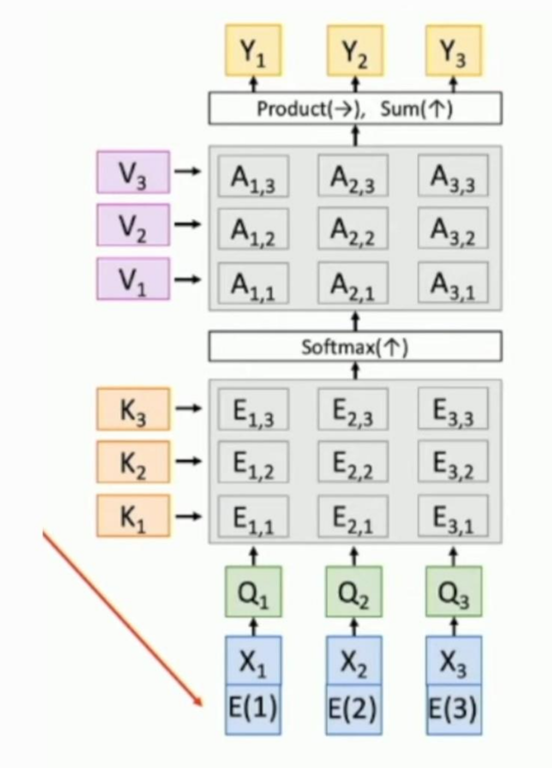
2. Feed-Forward Networks

$$FFN(x) = max(0, xW_1, +b_1)W_2 + b_2$$

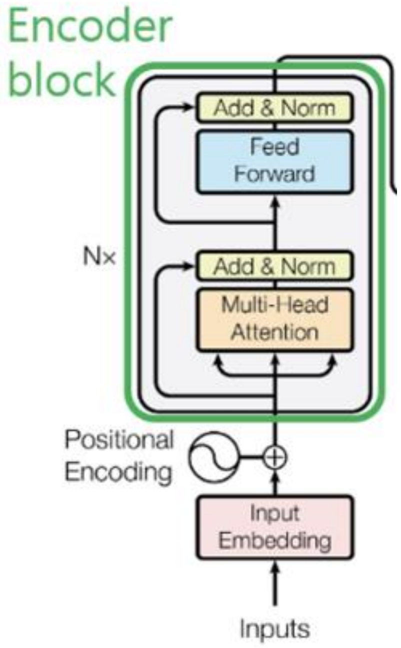


3. Positional Encoding





3. Positional Encoding

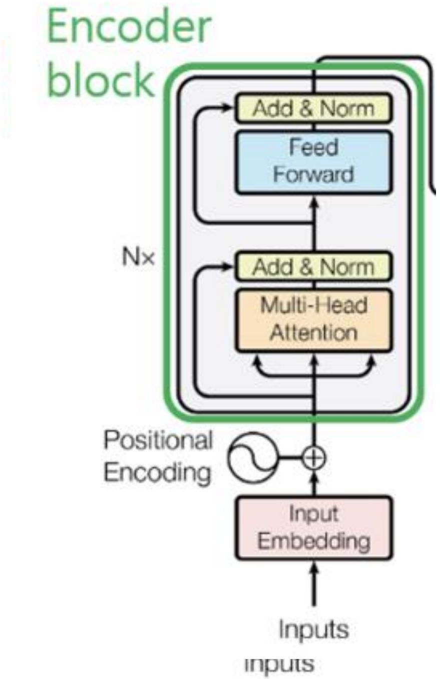


$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

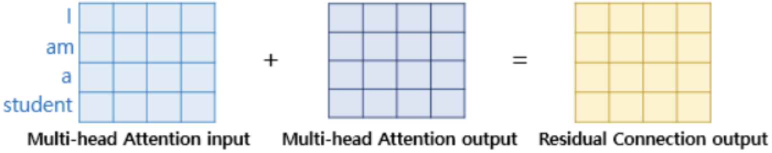
[트랜스포머(Transformer) 파헤치기—1. Positional Encoding (blossominkyung.com)]

Deep Sync

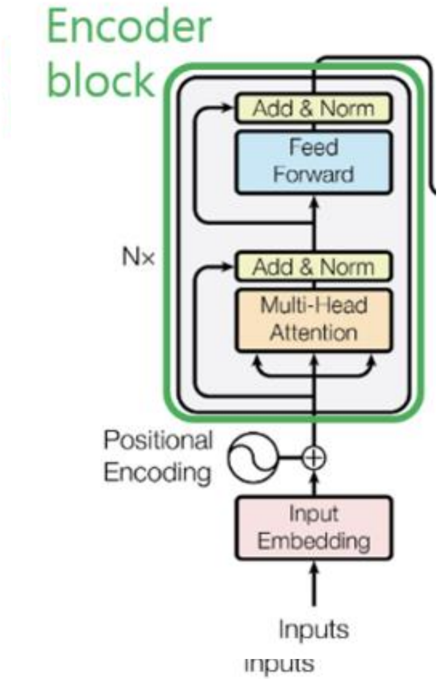
4. Add & Norm- Residual connection

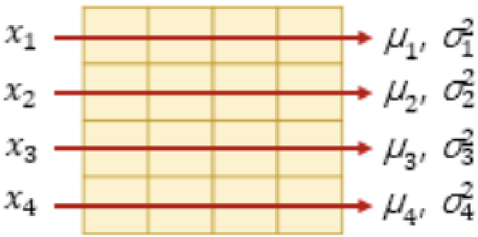


$$H(x) = x + Multi - head\ Attention(x)$$



4. Add & Norm – layer normalization



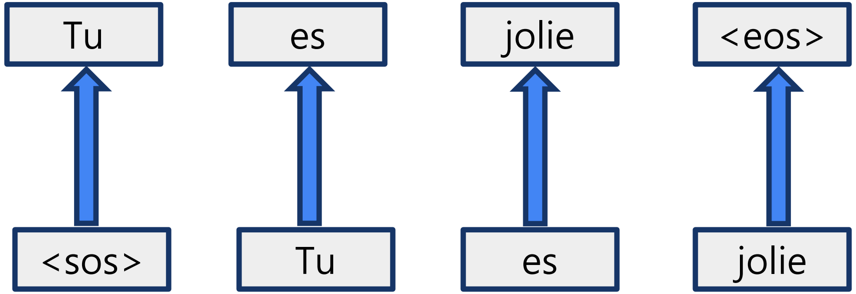


Residual Connection output

γ	1	1	1	1
β	0	0	0	0

$$\hat{x}_{i,k} = \frac{x_{i,k} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

$$ln_i = \gamma \hat{x}_i + \beta = LayerNorm(x_i)$$



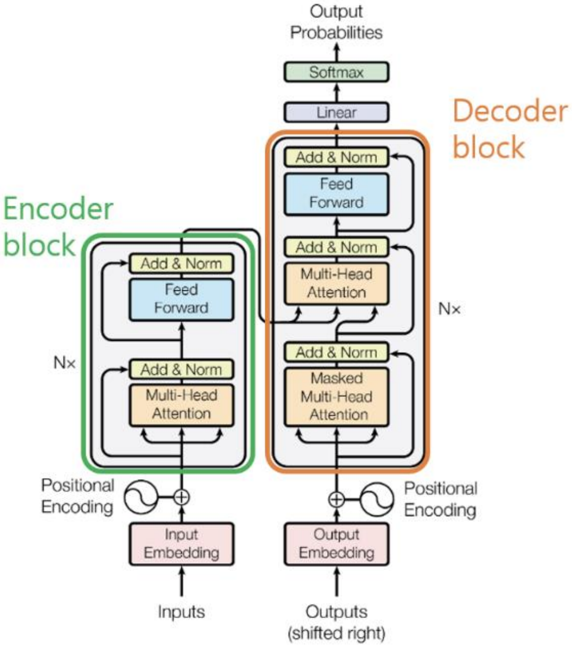
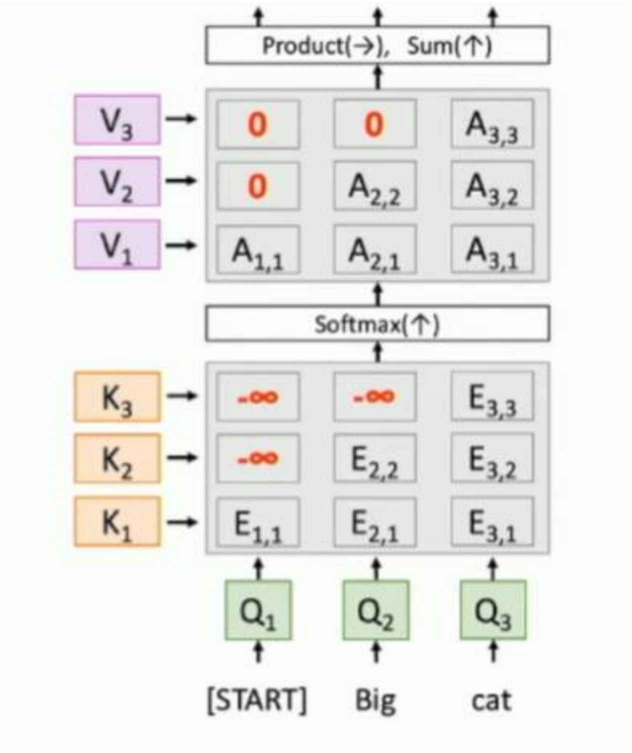


Figure 1: The Transformer - model architecture.

5. Masked Multi-Head attention



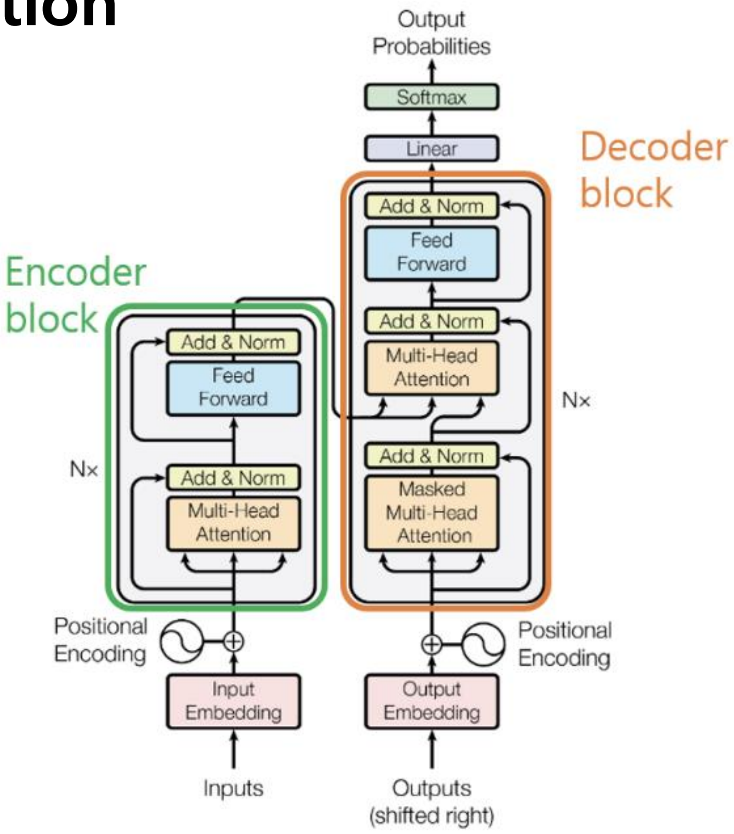
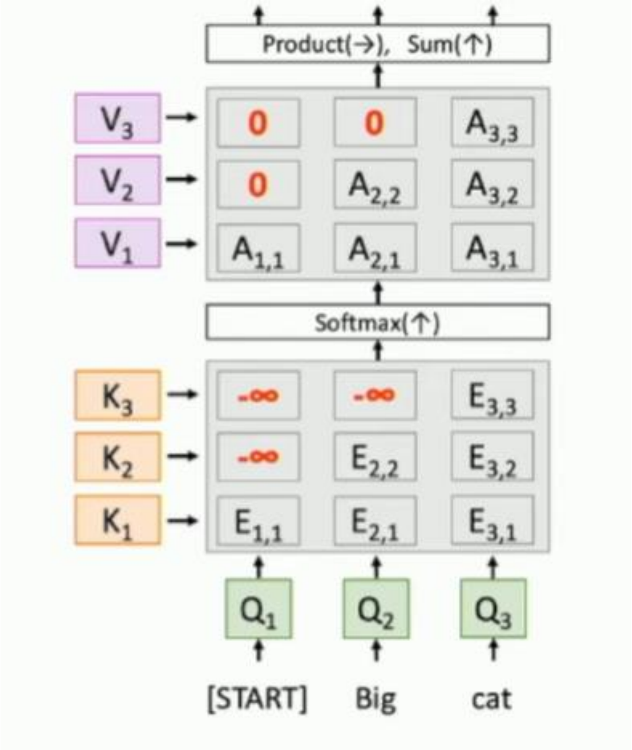


Figure 1: The Transformer - model architecture.

6. Decoder's Multi-Head Attention



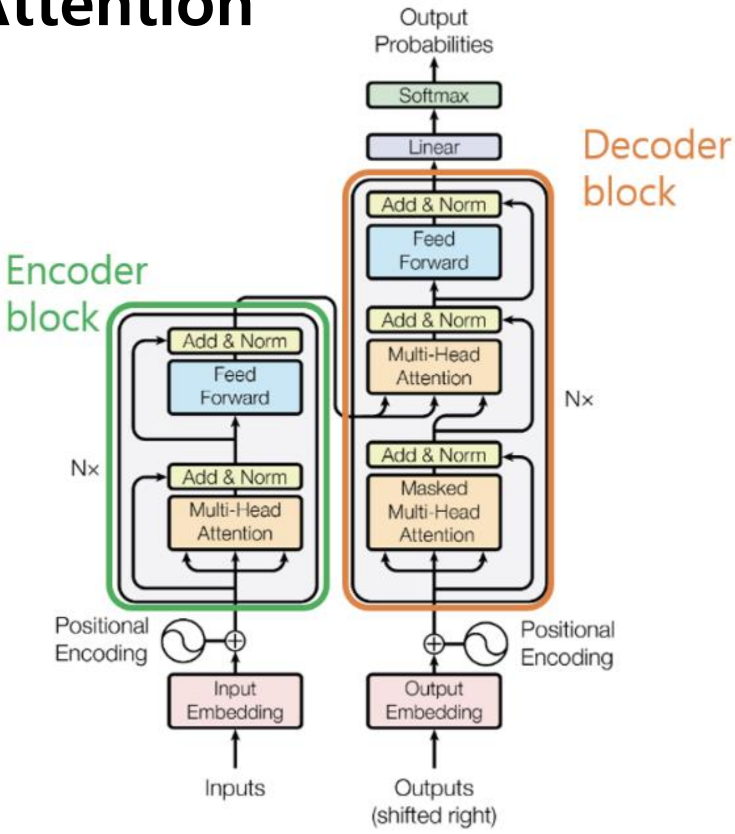


Figure 1: The Transformer - model architecture.

모델 평가

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

모델 평가

Deep Sync

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

모델 평가

Deep Sync

	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	
		head		4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16		d_k			5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4							$d \uparrow$		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	dropout		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

Thank you for your time