

TRANSFORMER

Deep learning architecture

딥러닝 프레임워크[00]

최인엽 교수님

201904236 산업데이터사이언스학부 전병준

202184042 데이터 사이언스 학과 이재성



목차

01

트랜스포머 모델이란

02

모델의 구조

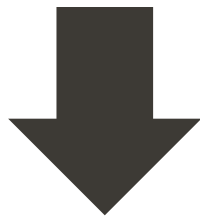
03

모델의 응용

Part 1 트랜스포머 모델이란

Attention is All You Need

Vaswani et al.



TRANSFORMER

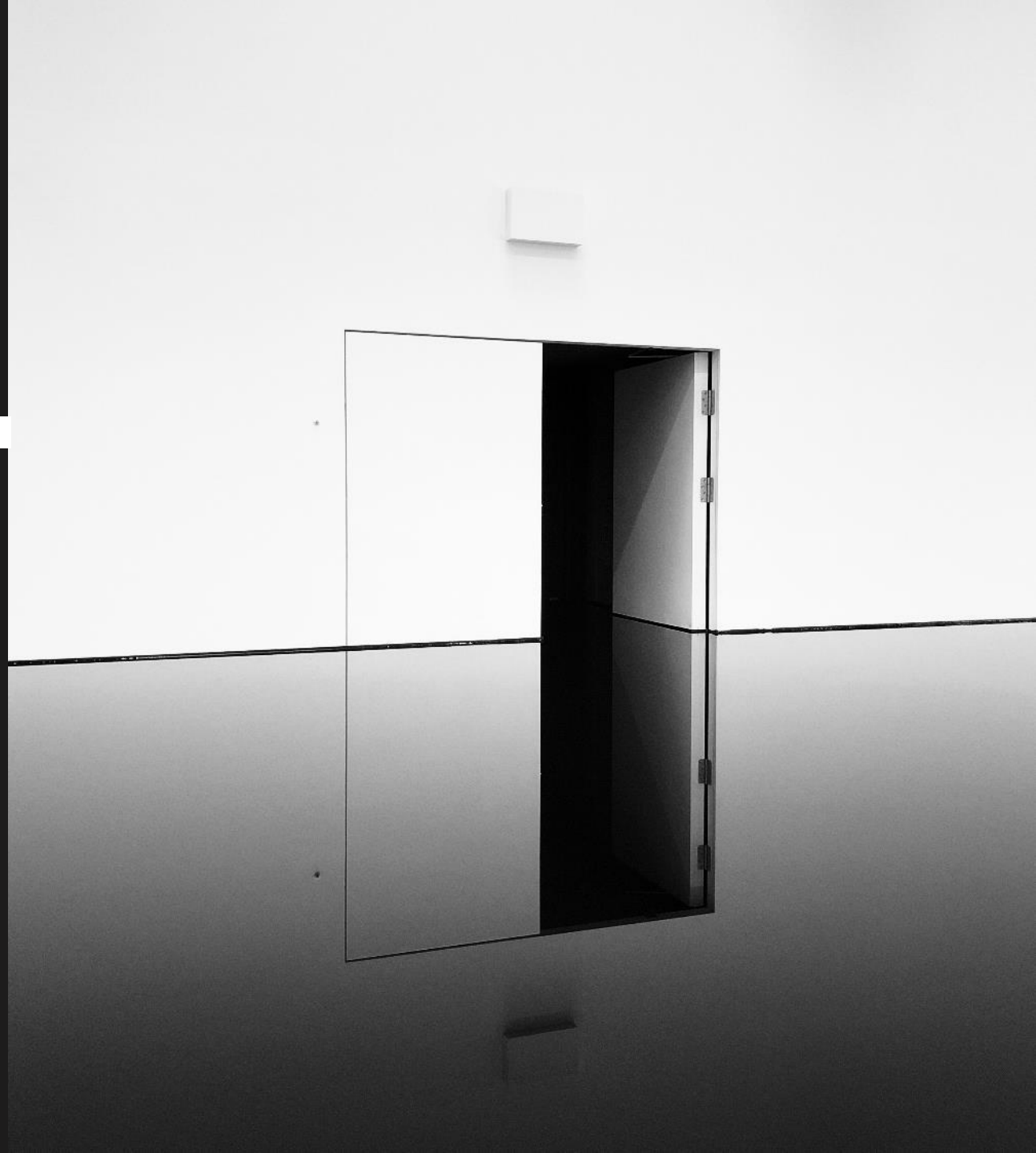


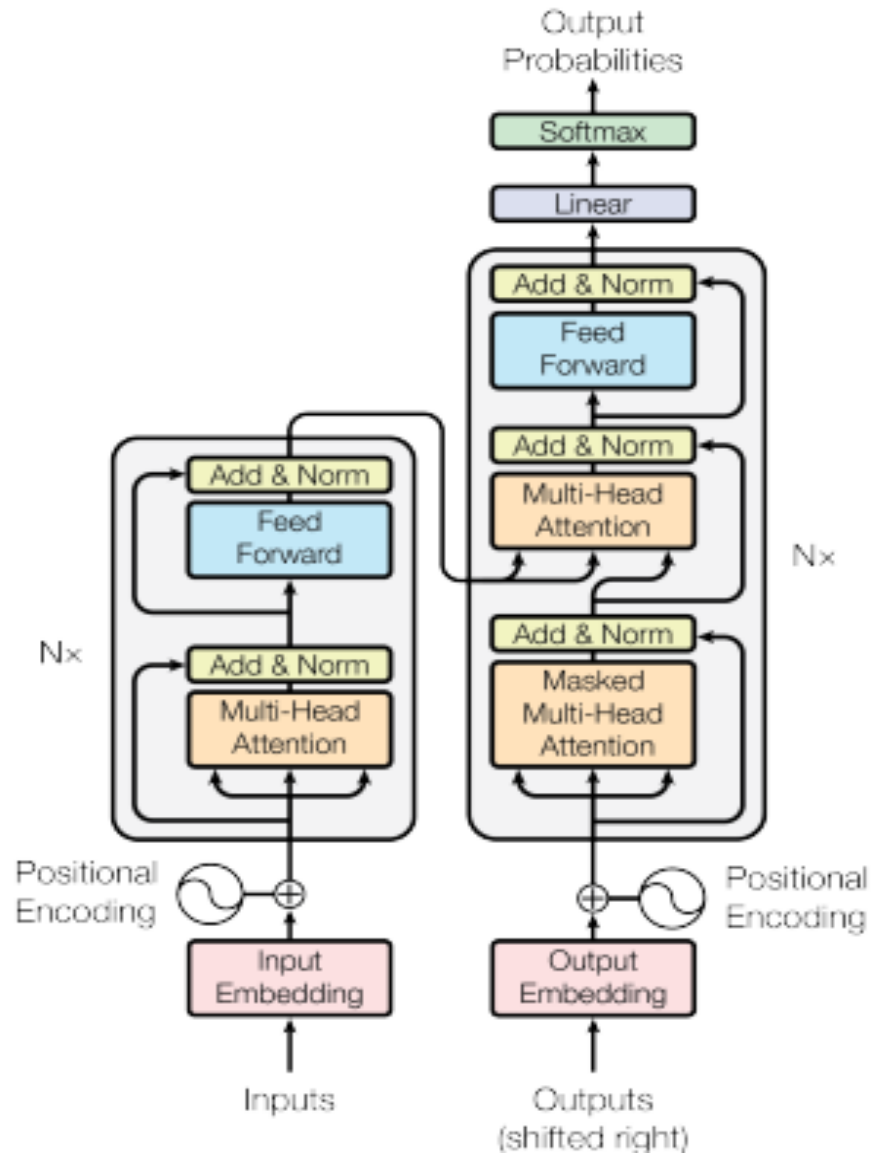
Transformer

GPT

BERT

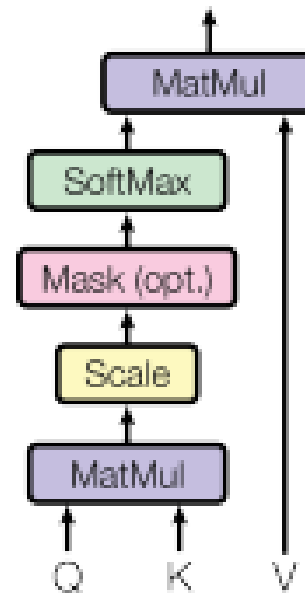
Part 2 모델의 구조



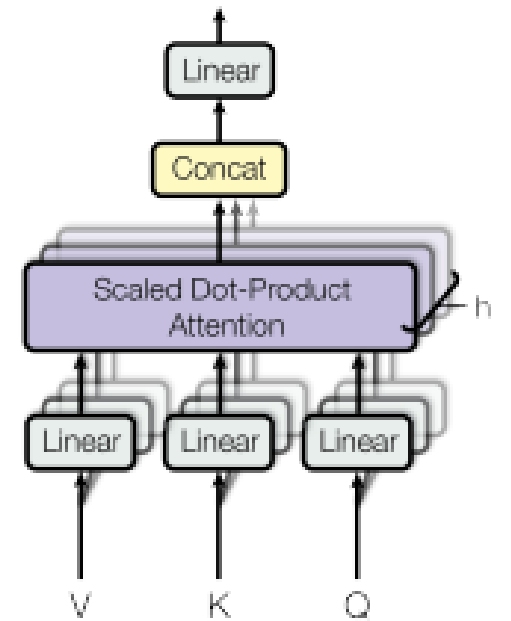


전체적인 아키텍처

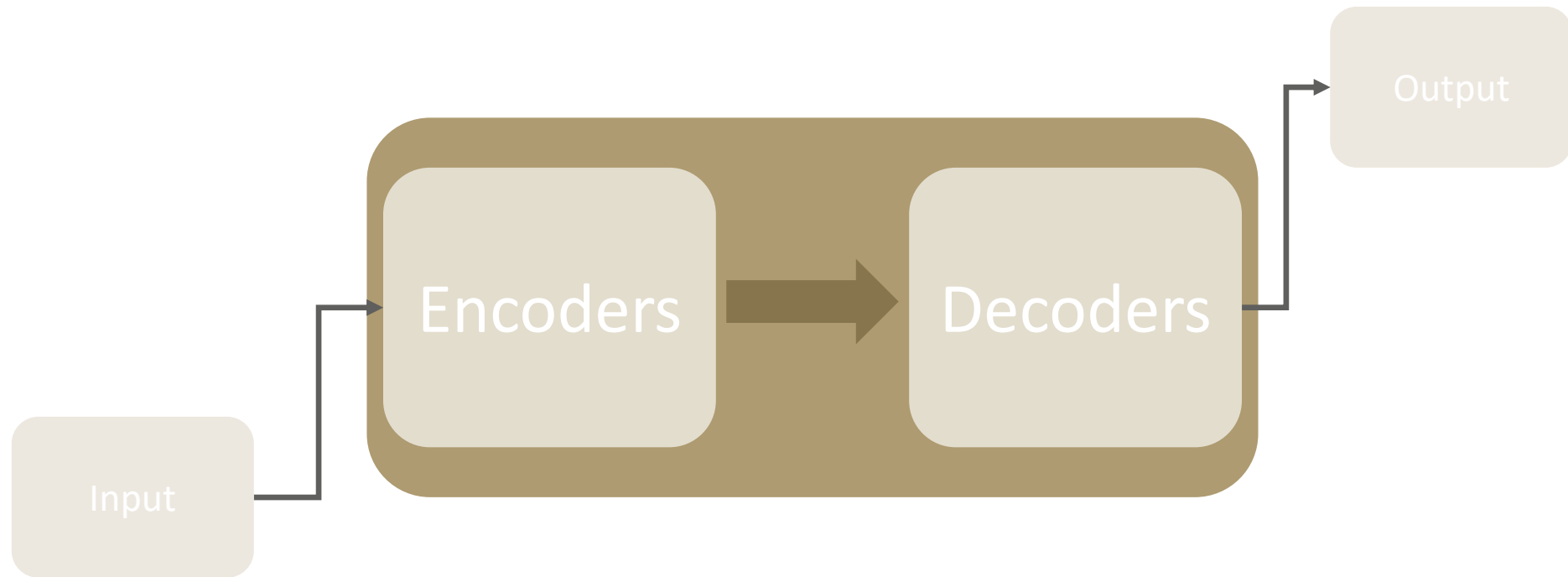
Scaled Dot-Product Attention

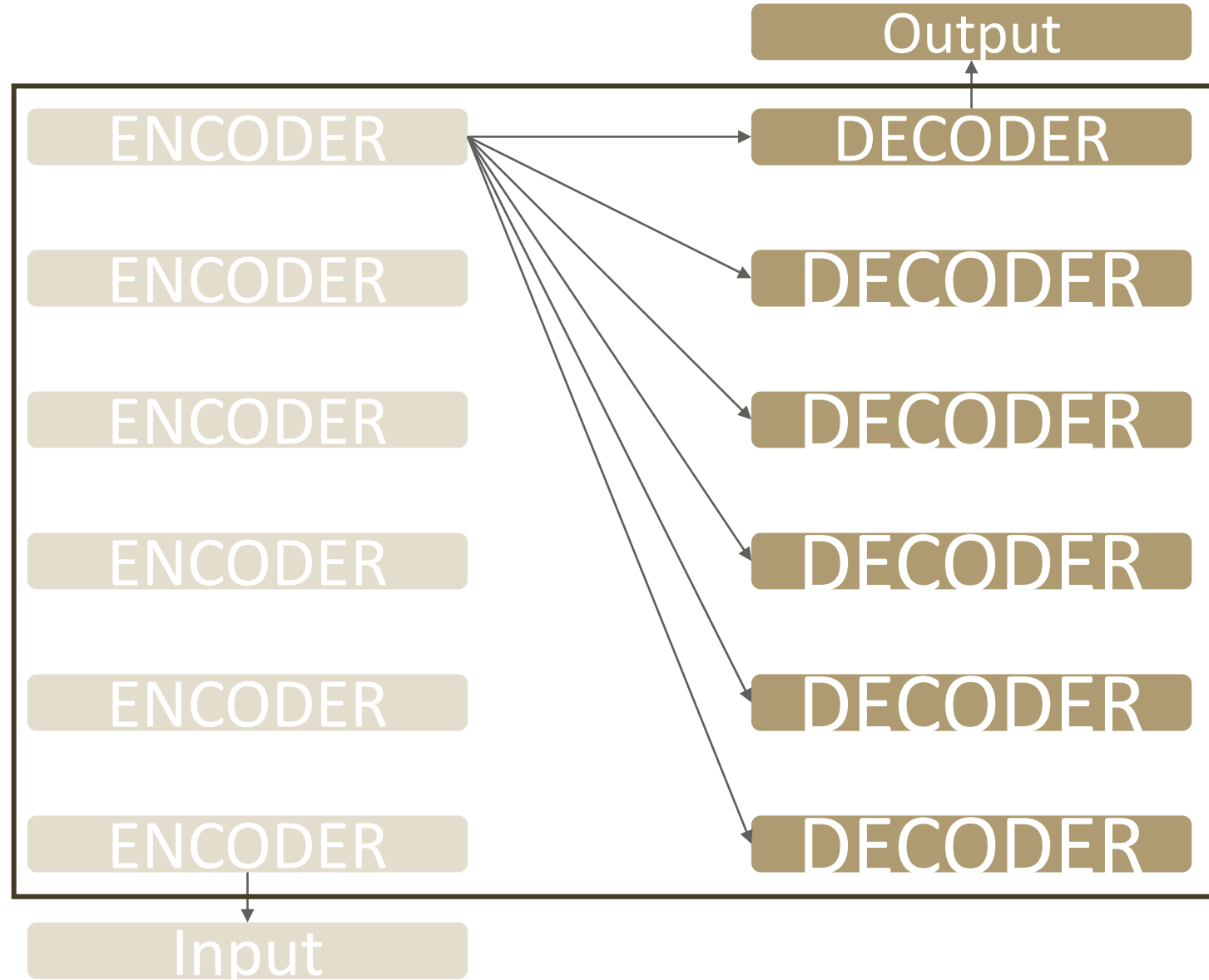


Multi-Head Attention

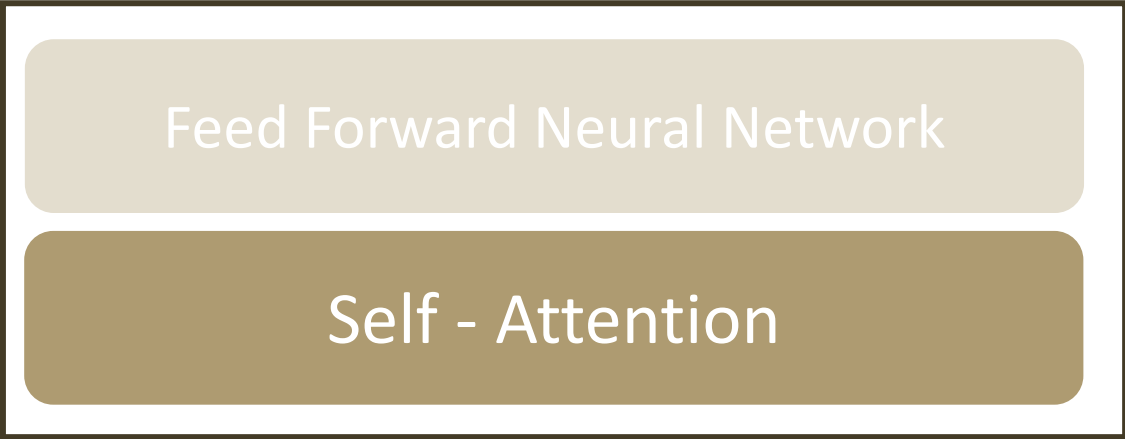


멀티 헤드 어텐션



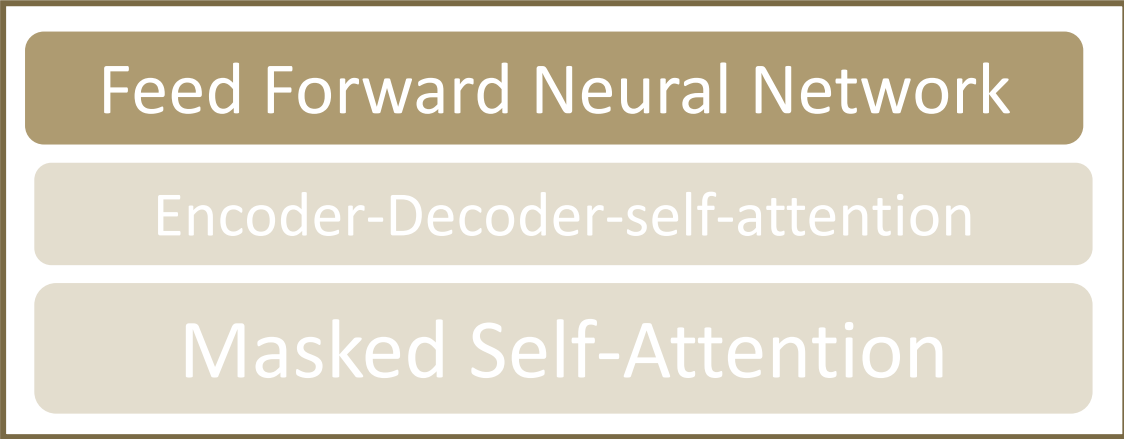


ENCODER BLOCK

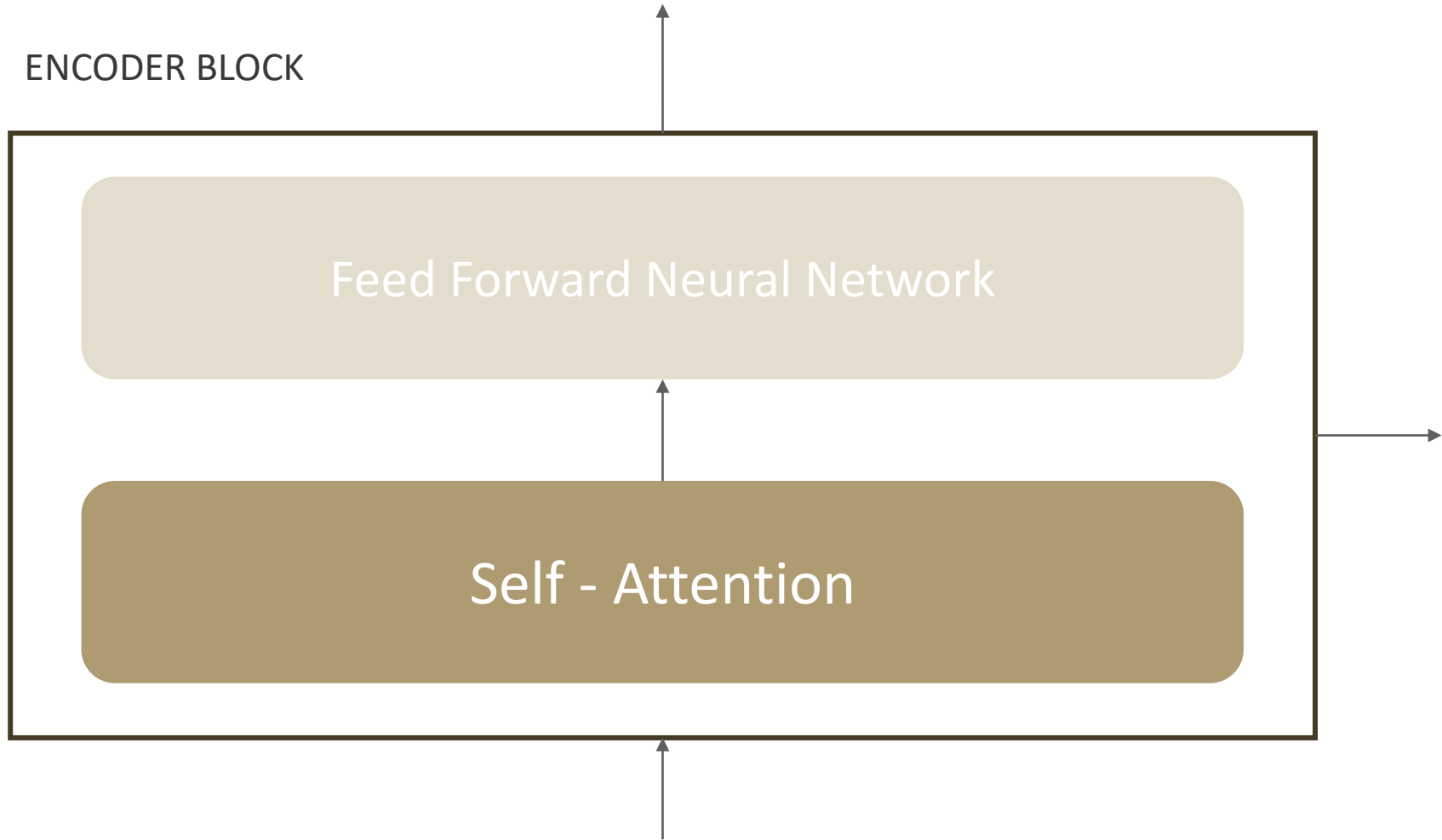


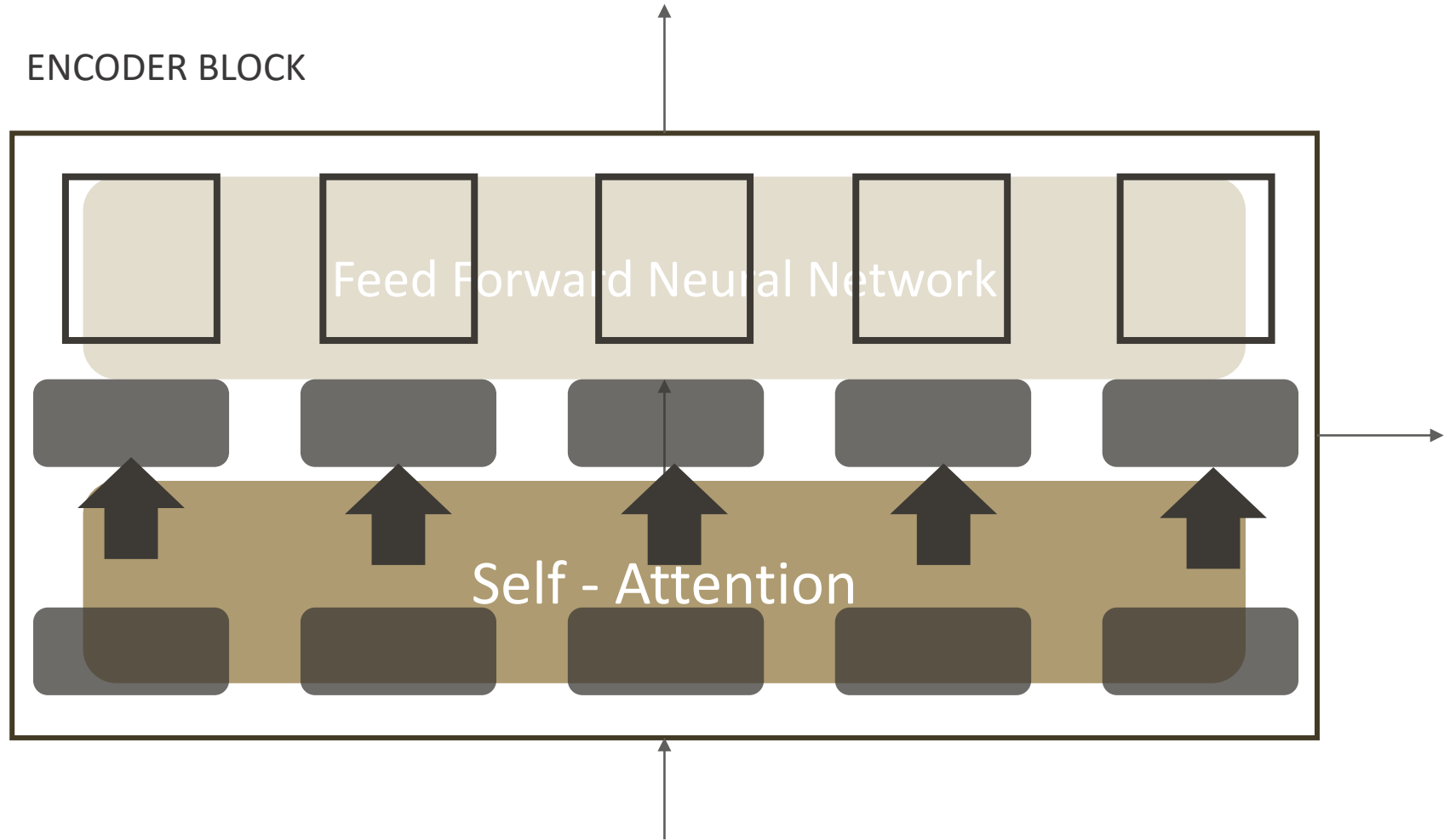
word1	word2	word3	<eos>	<pad>	...	<pad>
1	2	3	4	5	...	512

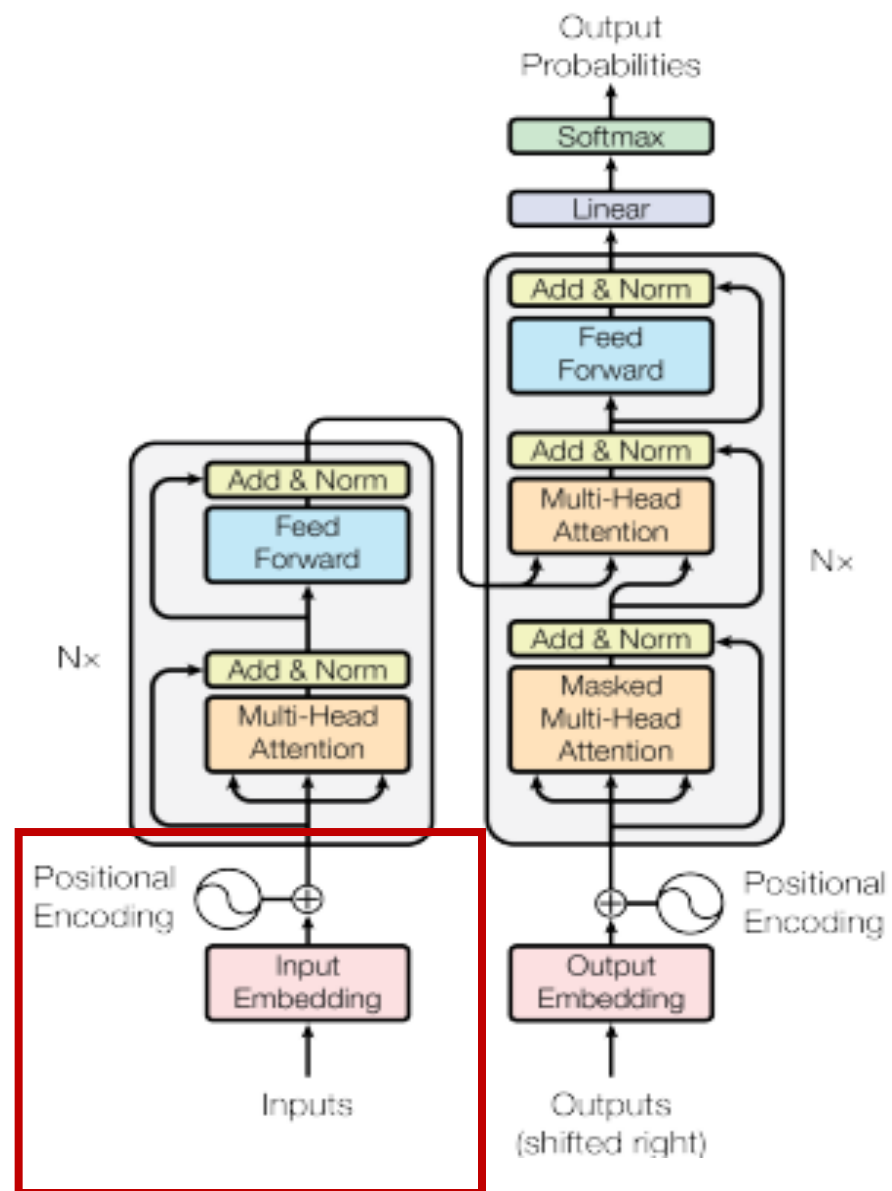
DECODER BLOCK



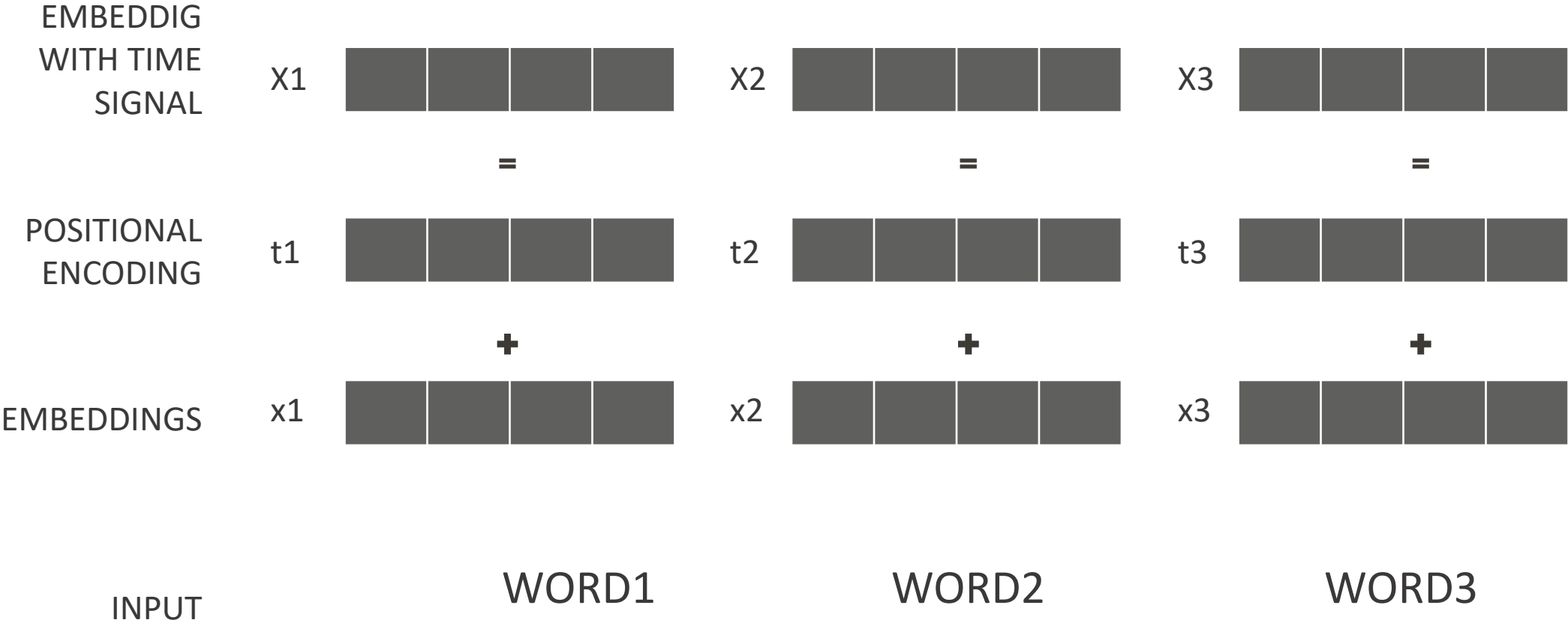
	word1	word2	word3			
1	2	3	4	5		512

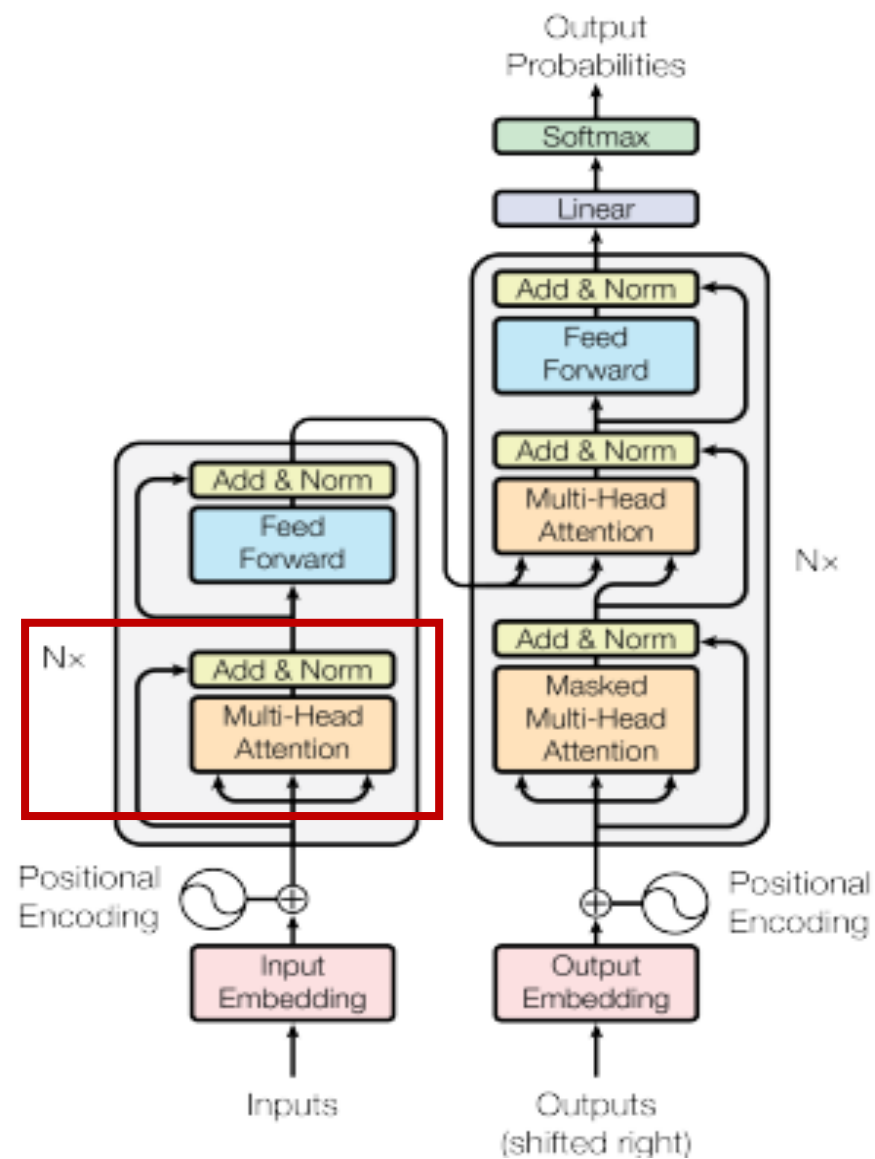


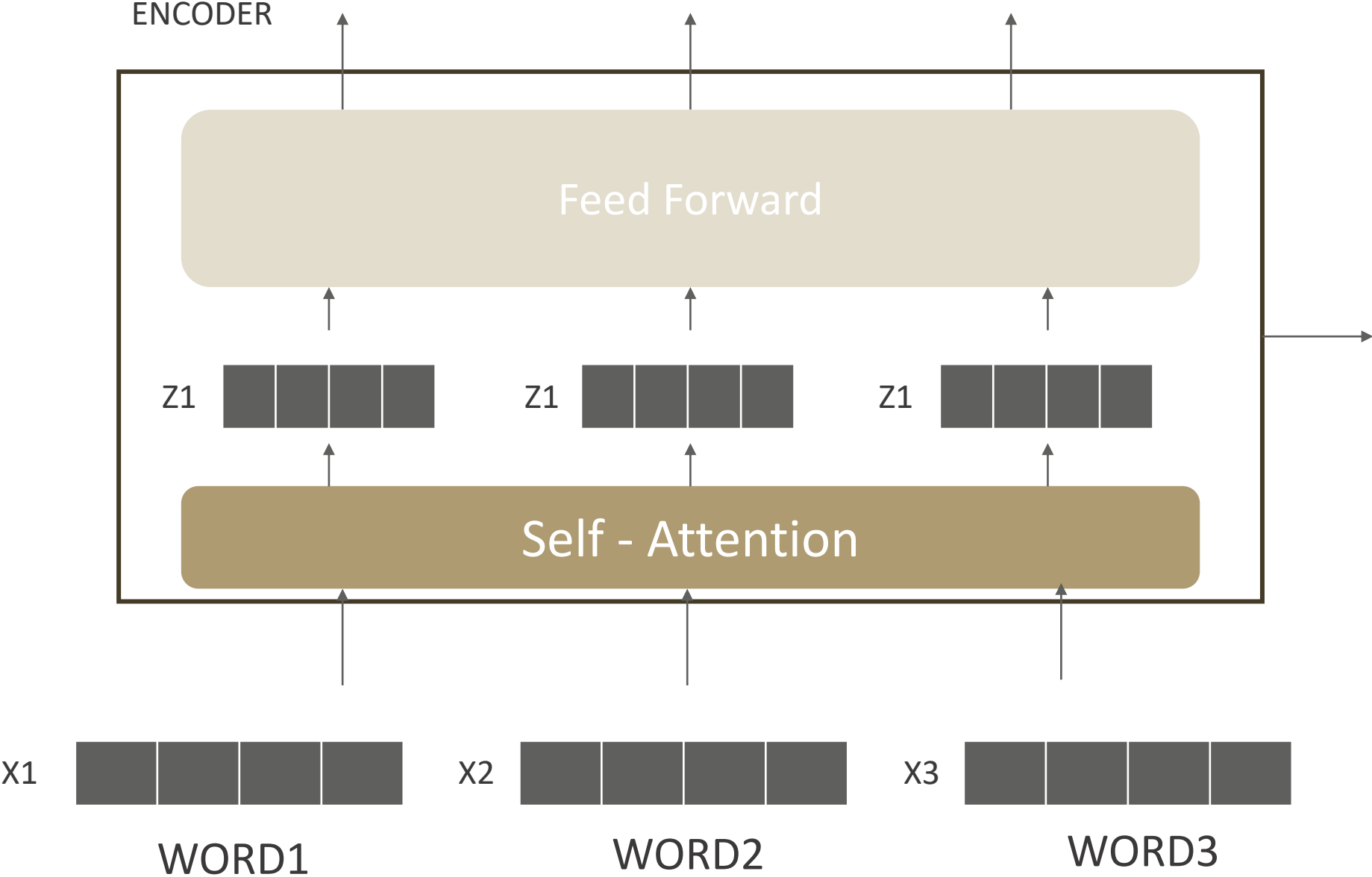




Positional Encoding







Attention

EMBEDDIG

X1



X2



Queries

Q1



Q2



Keys

K1



K2



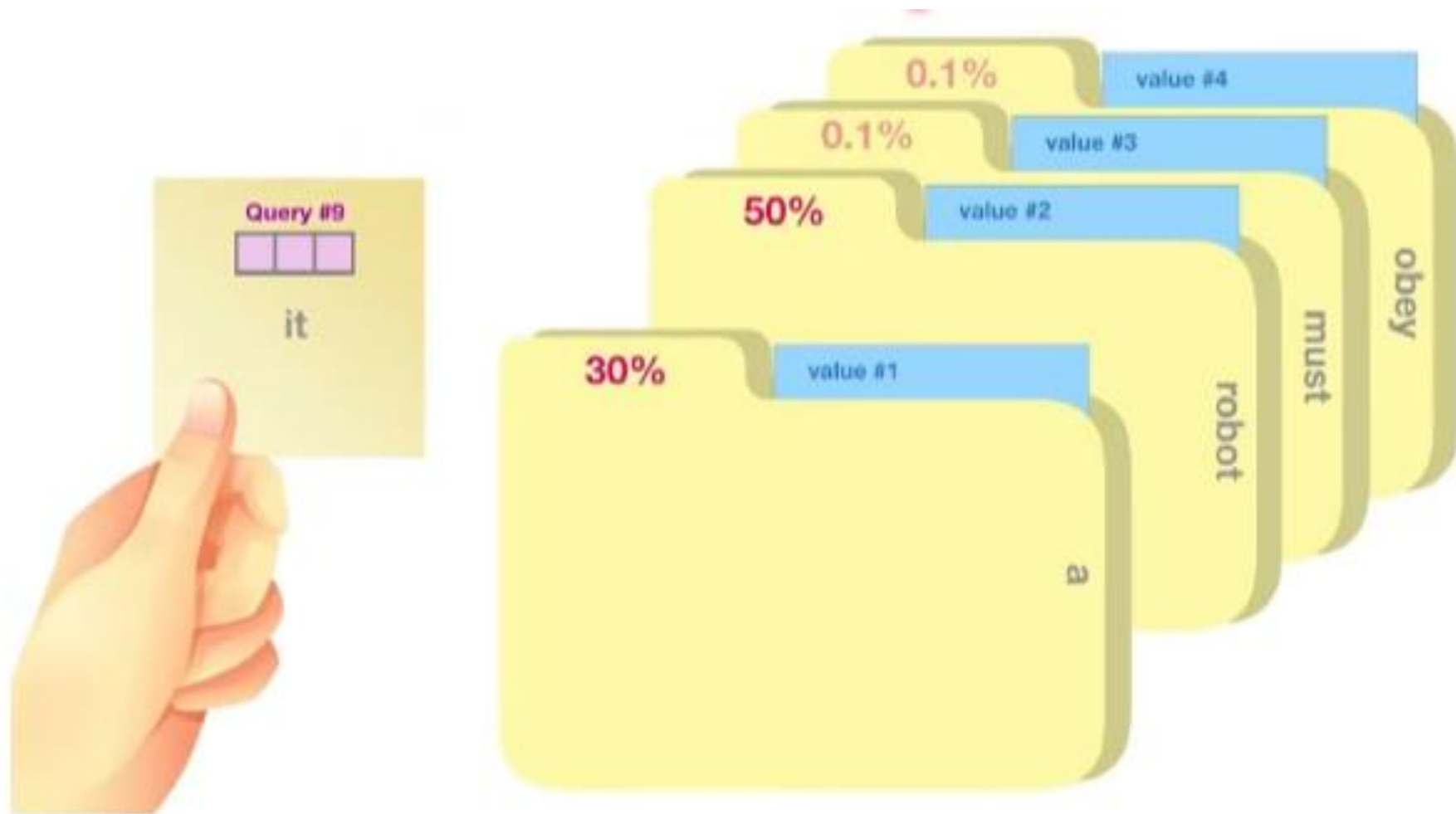
Values

V1



V2

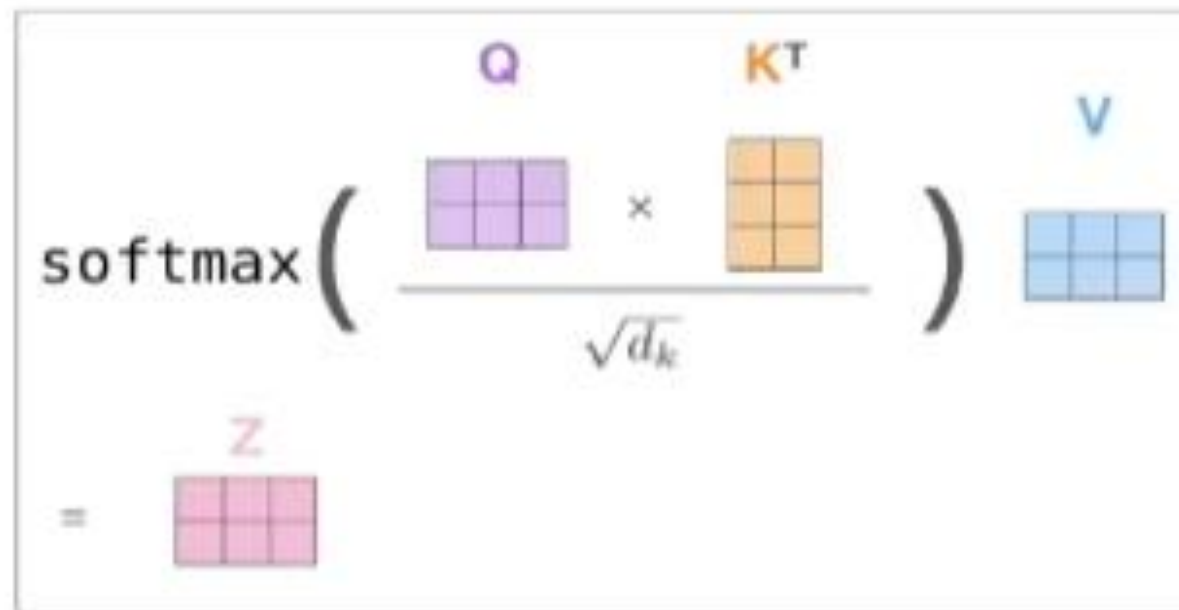
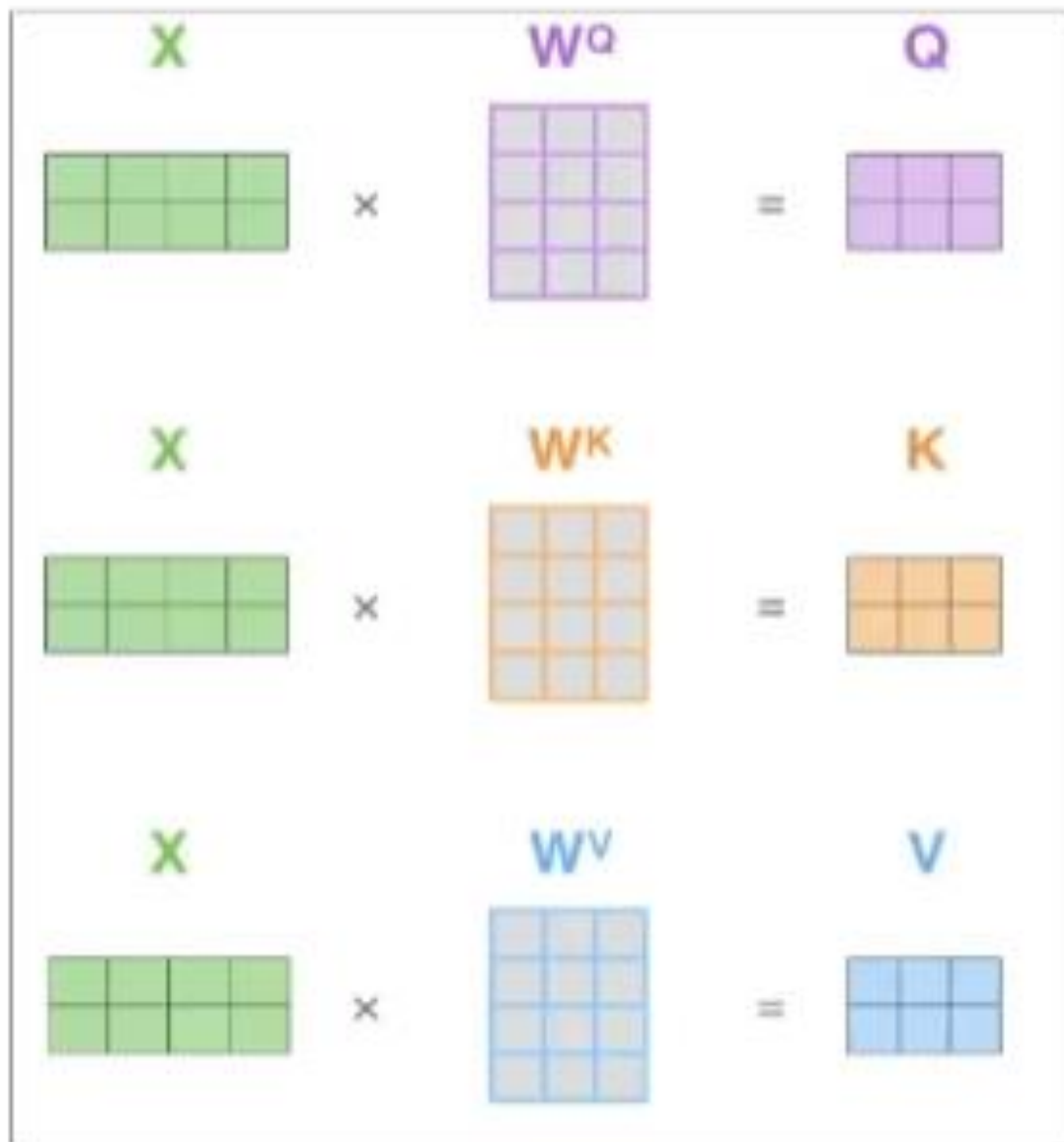




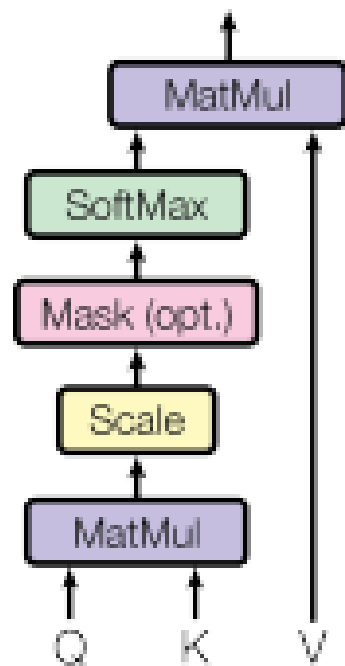
Attention



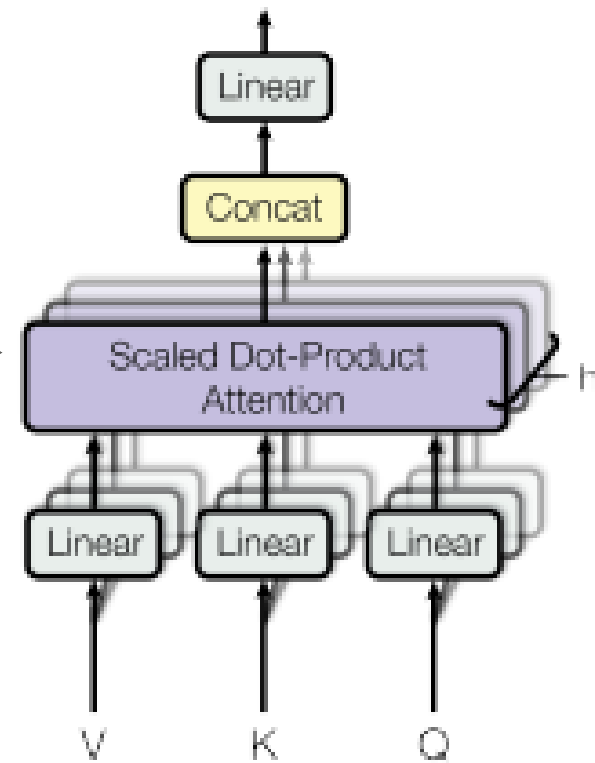
score	Q1 *K1 =				Q1 *K2 =			
Normalize	14				12			
Softmax	0.88				0.12			
✖ Value	V1				V2			
Sum	Z1				Z2			



Scaled Dot-Product Attention



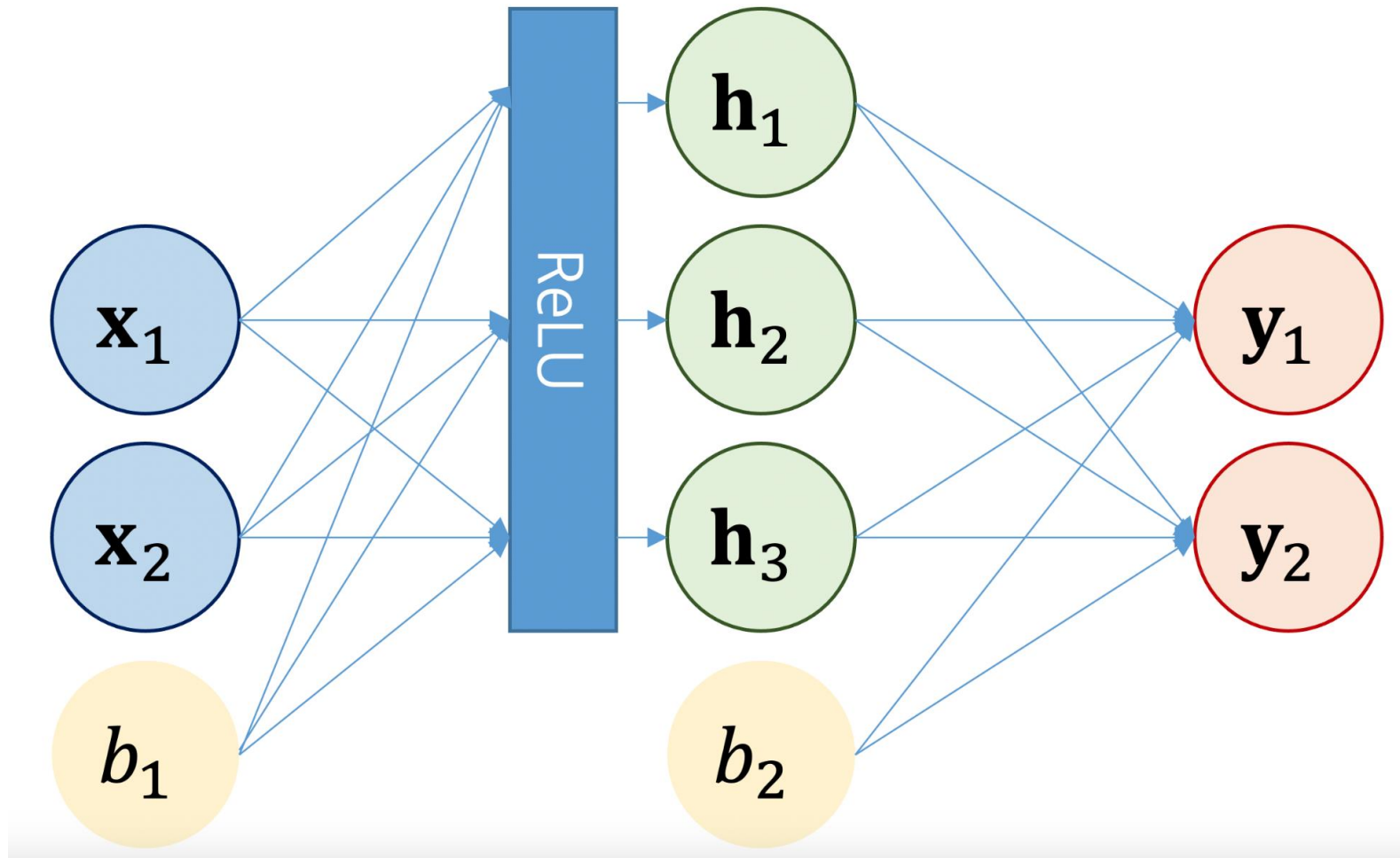
Multi-Head Attention

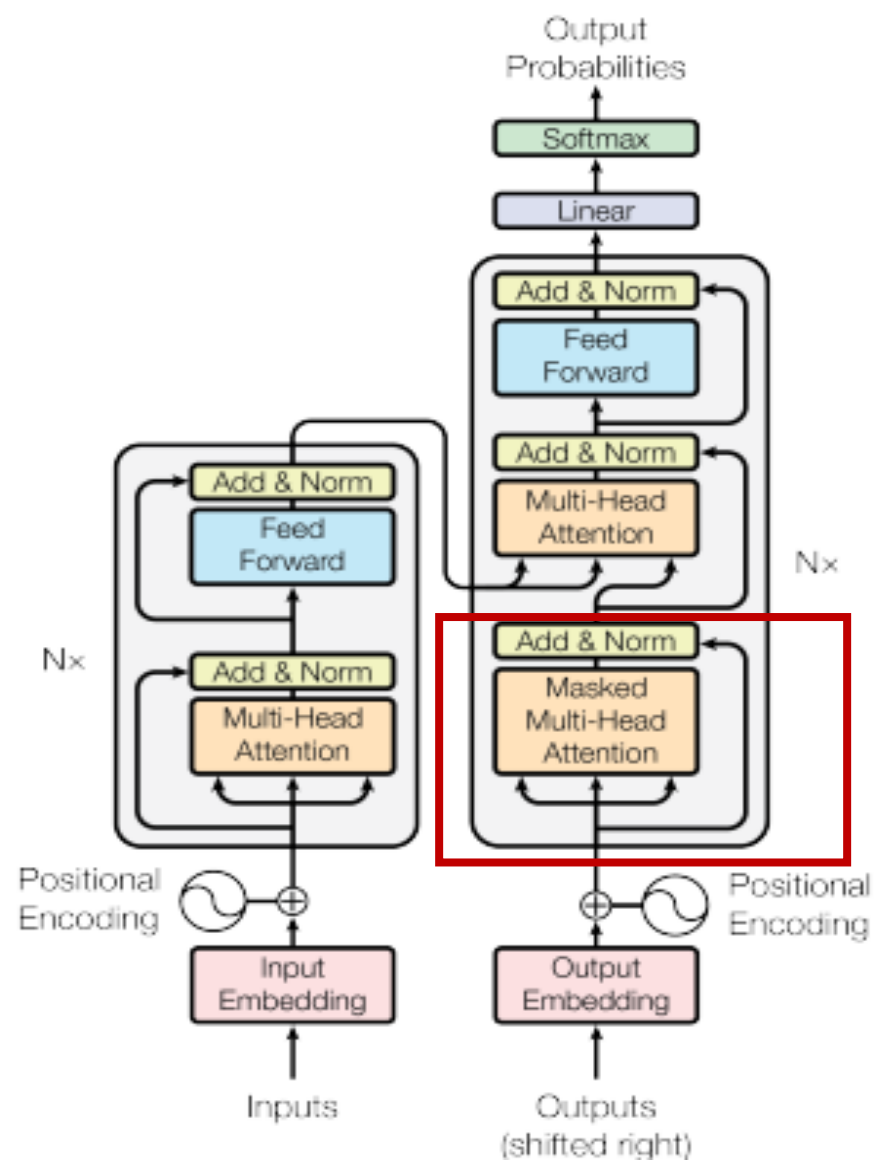


동시에 여러 번 수행

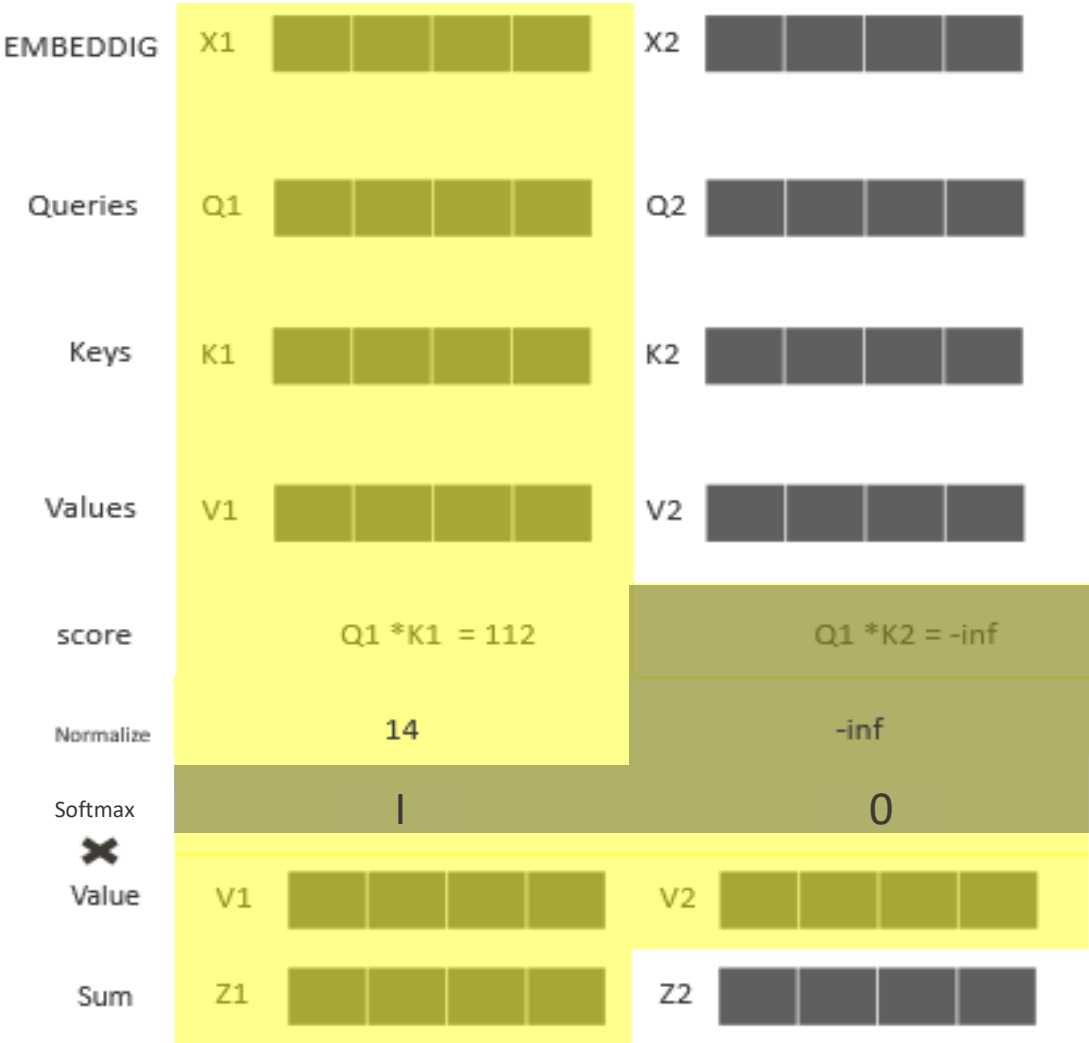
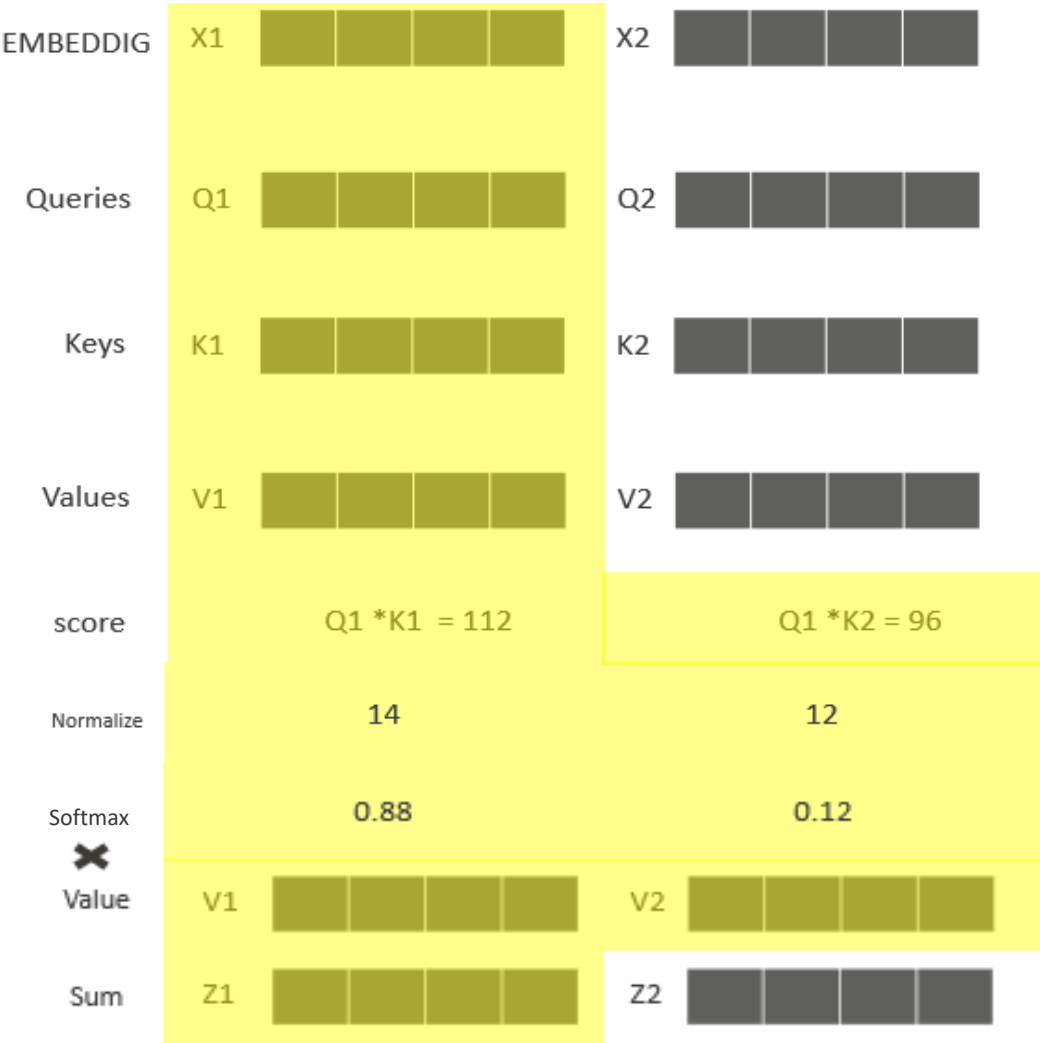


$$FFN(z) = \max(0, zW_1 + b_1)W_2 + b_2$$

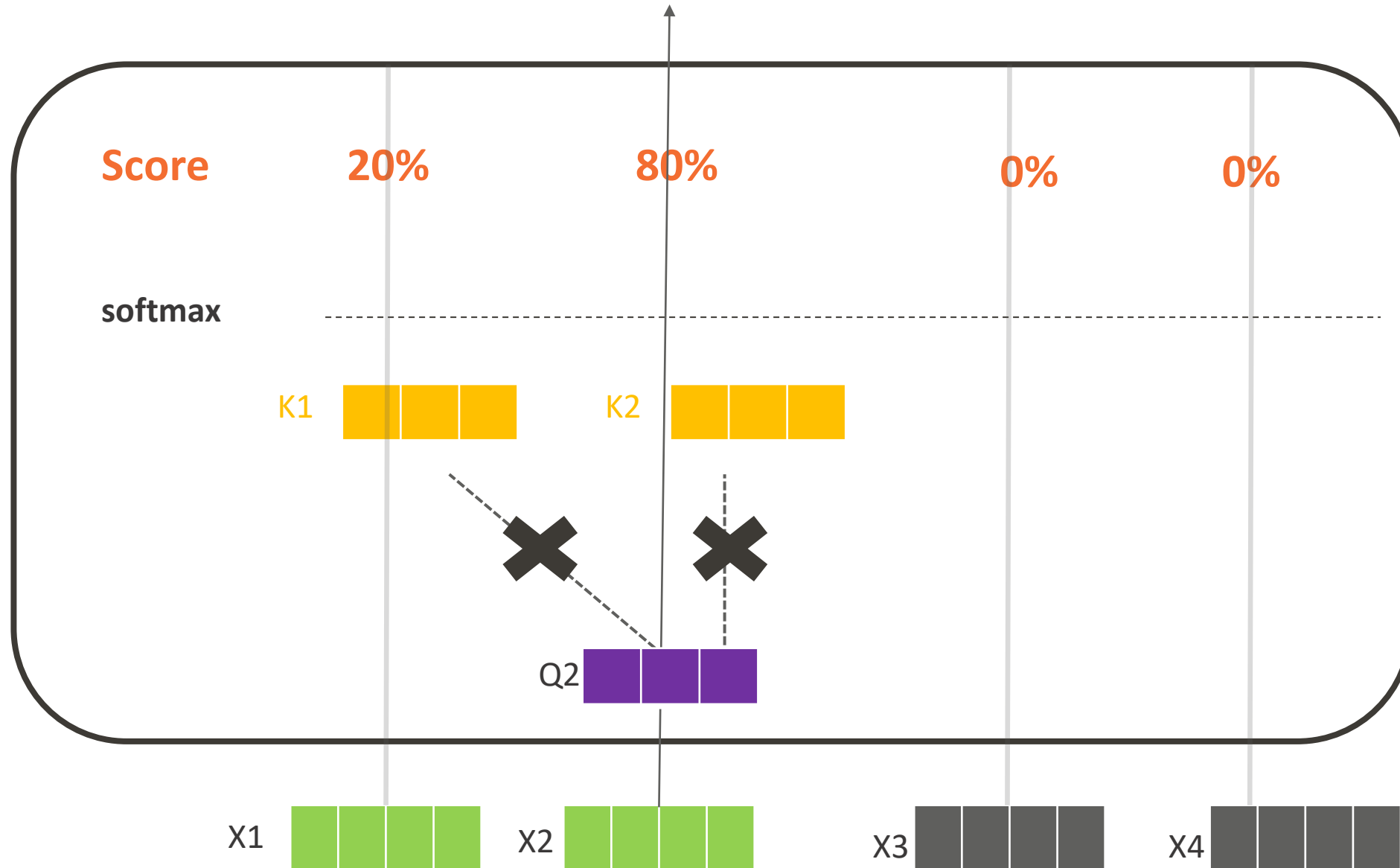




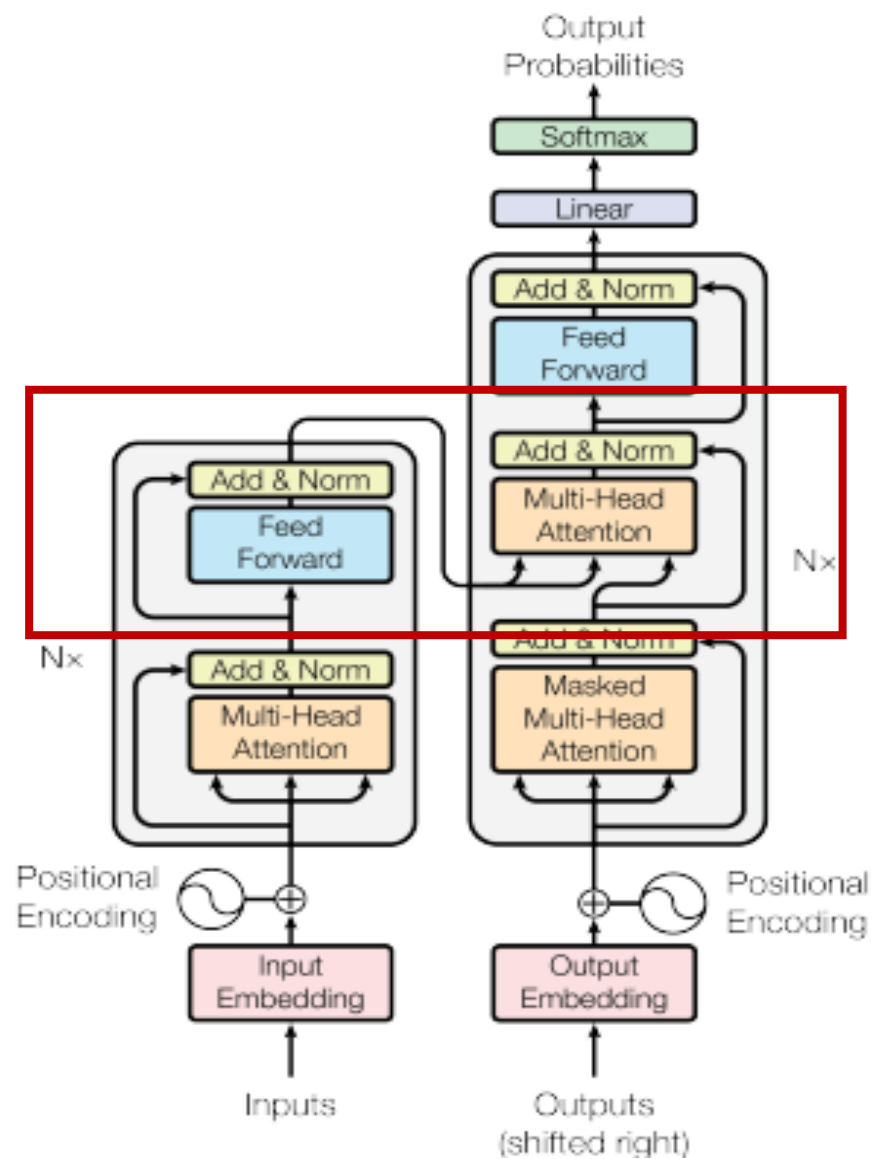
Masked Multi Head Attention

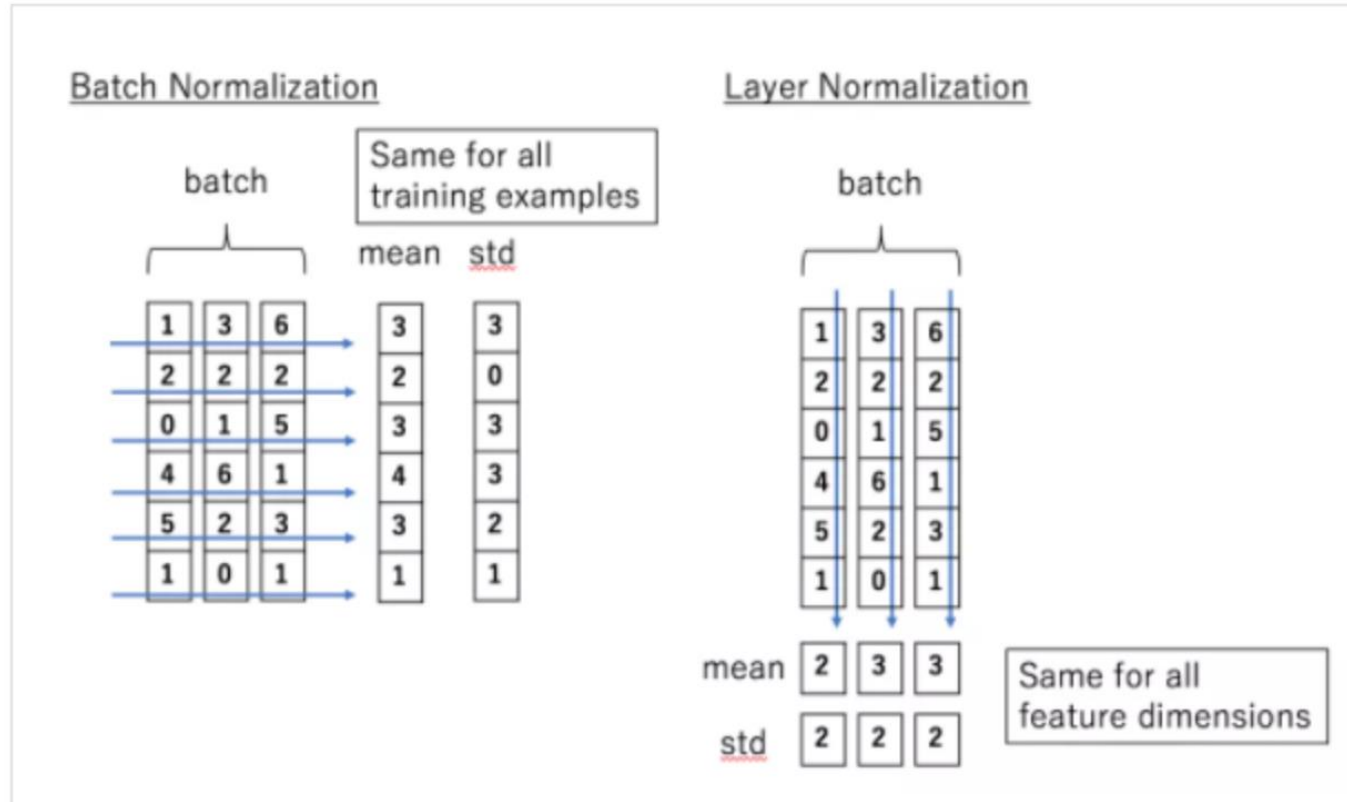


Masked Multi Head Attention



이유





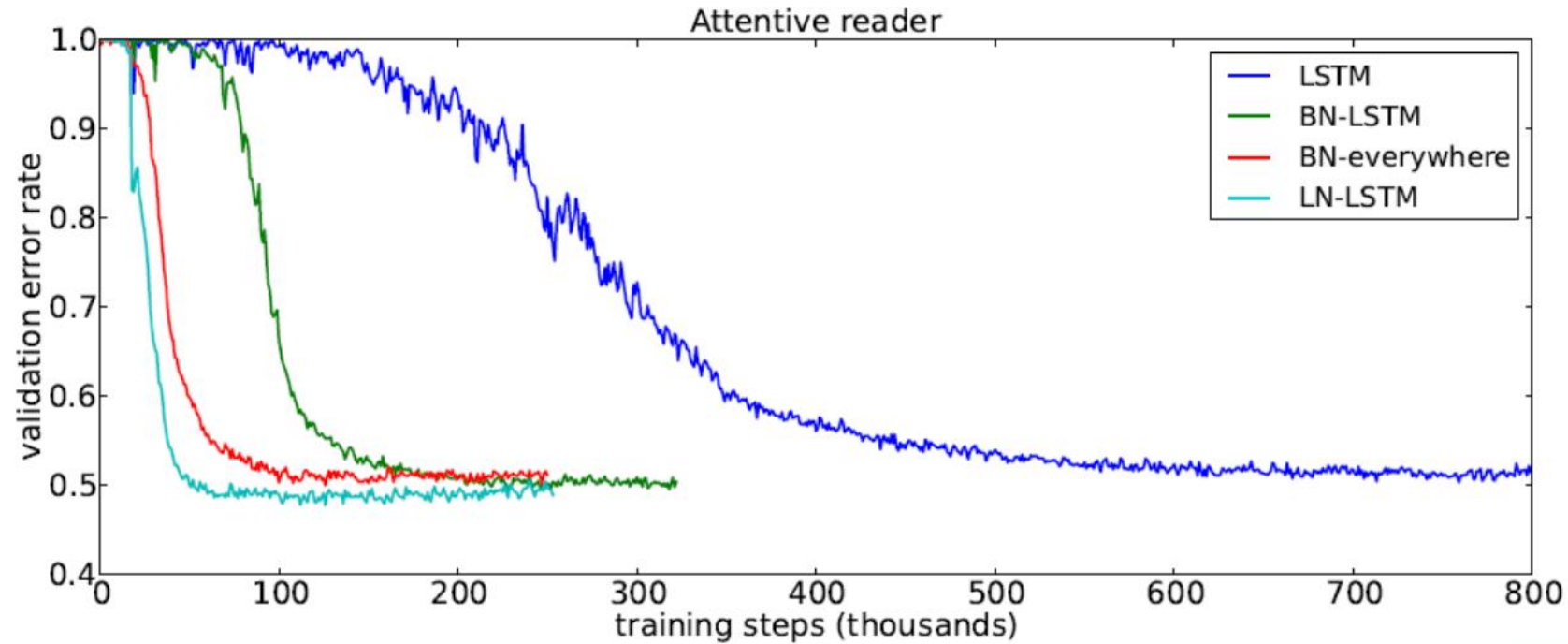


Figure 2: Validation curves for the attentive reader model. BN results are taken from [Cooijmans et al., 2016].

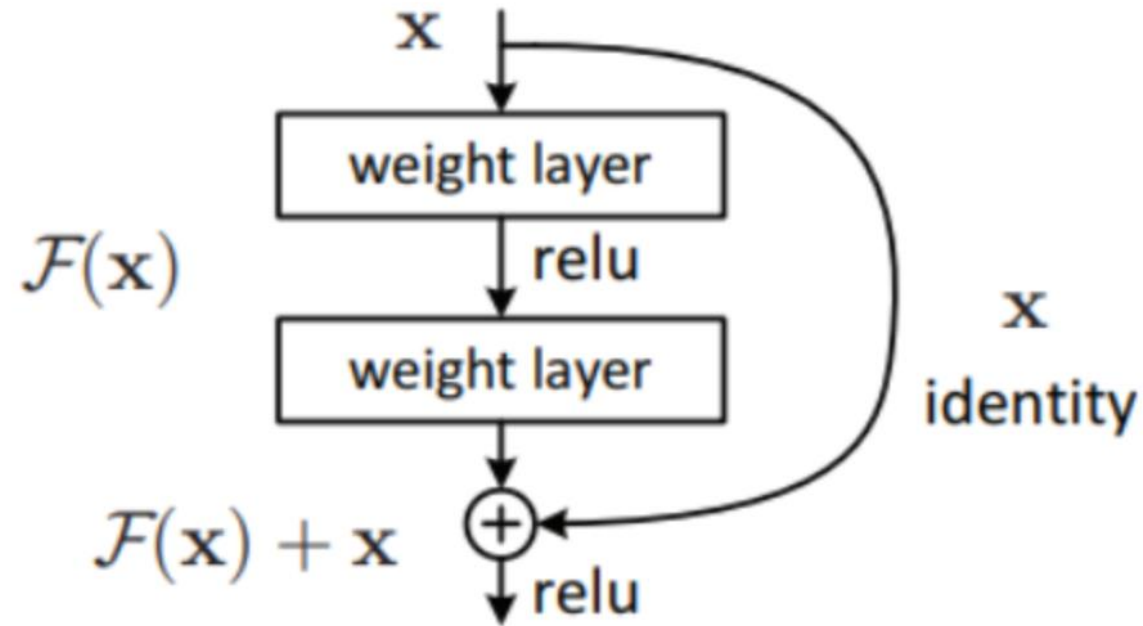


Figure 2. Residual learning: a building block.

Part 3 모델의 응용



대표 모델

BERT

GPT

대표 사례

구글 번역

OpenAI
GPT

Q&A



EXIT



Fin.