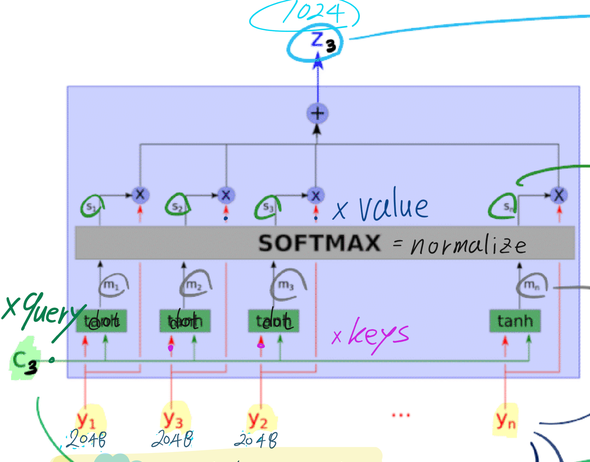
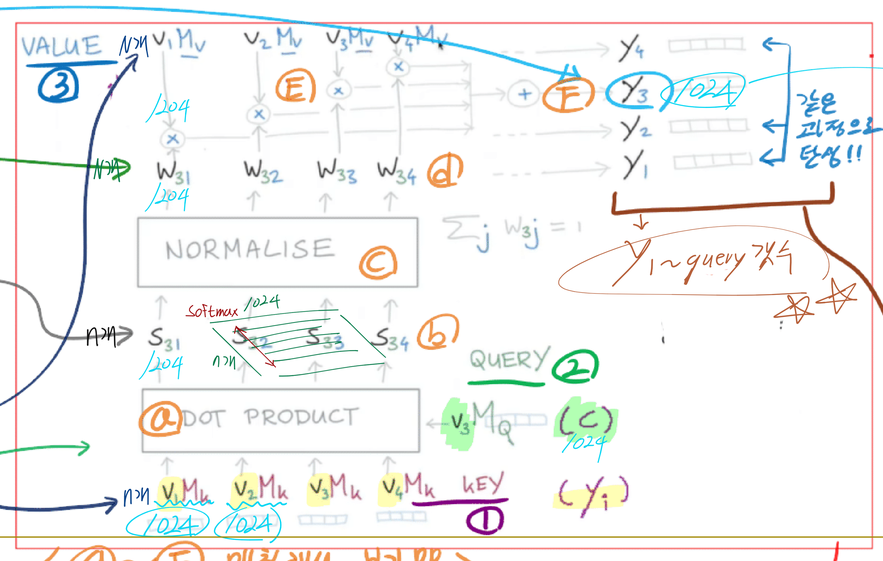
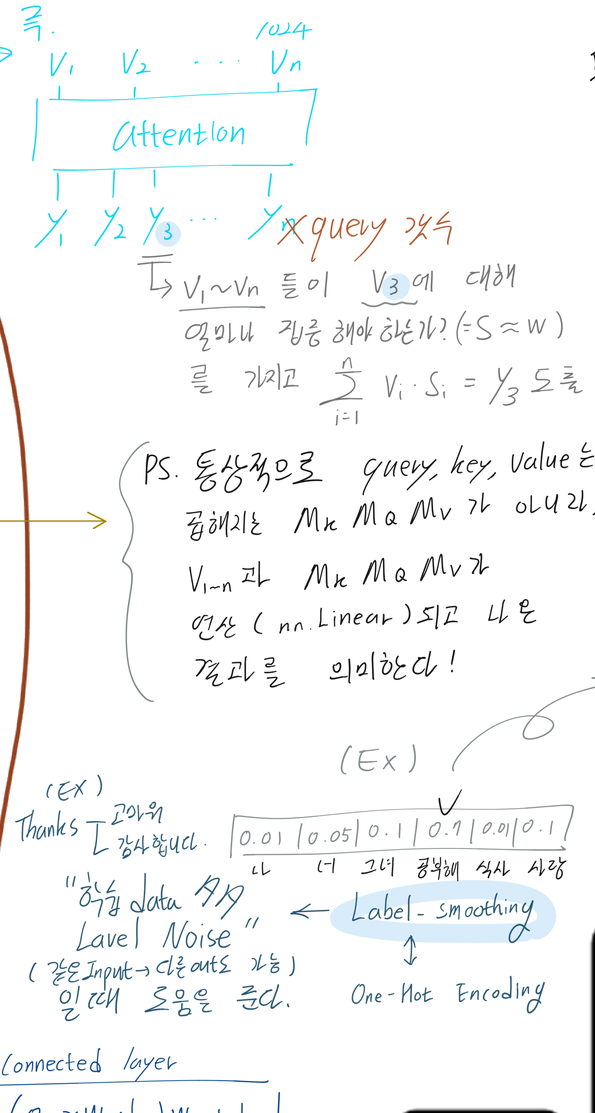


~~(1024 · 2048)~~ · (2048 · 1개)

↳ M_k M_v N

= 1024 x 1/개 = Query, Key, Value

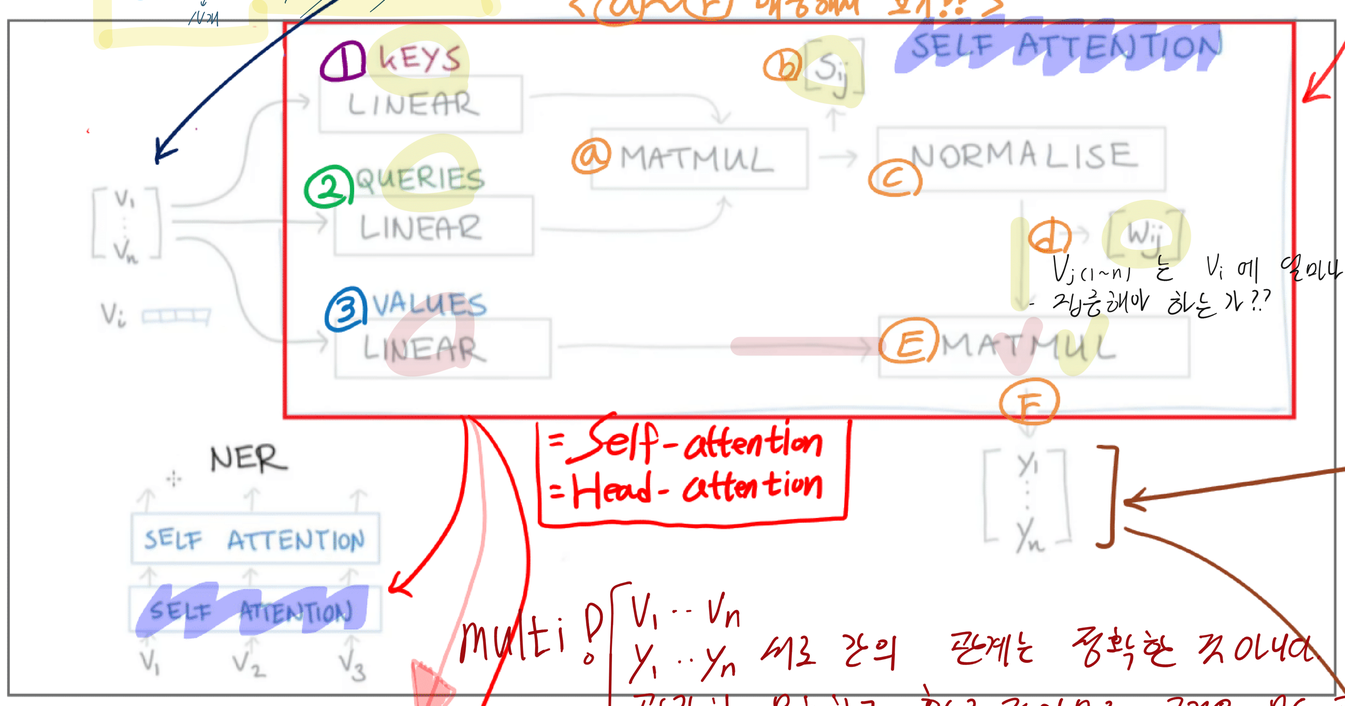




↳ $V_1 \sim V_n$ 들이 V_3에 대해 얼마나 집중 해야 하는가? (= S ≈ W) 를 가지고 $\sum_{i=1}^{n} V_i \cdot S_i = y_3$ 도출

특징
① $V_1 \cdots V_n \rightarrow y_1 \cdots y_n$ ~~똑같이 n개 나옴~~ query 갯수
② V_1. shape = 2048
y_1. shape = 1024
shape 유지 될 필요 없음
이전 Post ViT 코드 확인하기

PS. 통상적으로 query, key, value는 곱해지는 M_k M_Q M_V 가 아니라, $V_{1\sim n}$과 M_k M_Q M_V가 연산 (nn.Linear)되고 나온 결과를 의미한다!

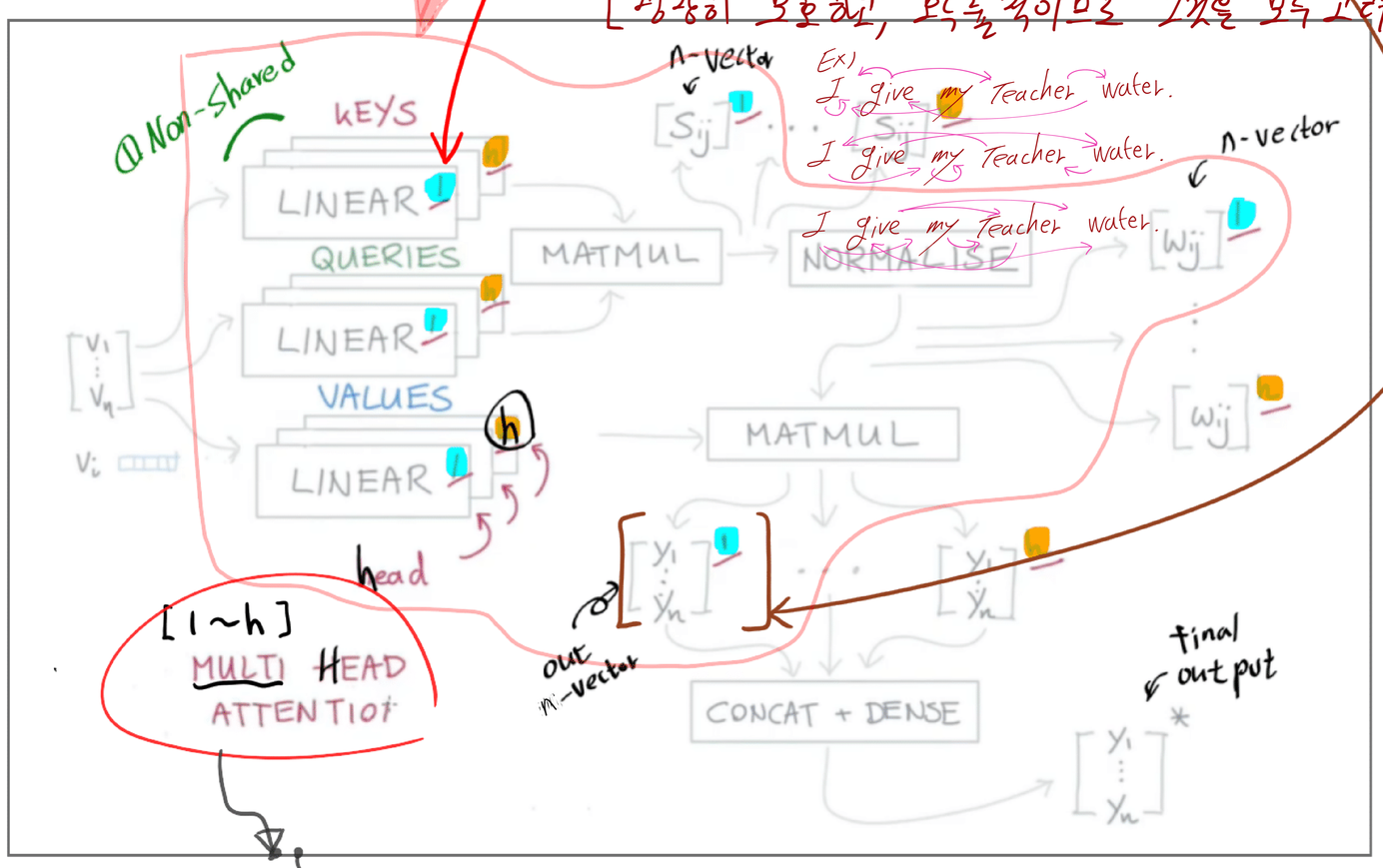


<(a)~(F) 매칭해서 보기!!>

$V_{j(1\sim n)}$는 V_i에 얼마나 집중해야 하는가??

= Self-attention
= Head-attention

multi! $V_1 \cdots V_n$, $y_1 \cdots y_n$ 서로 간의 관계는 정확한 것 아니어도 광범위 포함하고, 확률적이므로 그것을 모두 고려한다!



"공부해" ③ decoder의 다음 단어에 가장 적합한 단어는??

Output Probabilities

(Ex)

Thanks I 고마워 감사합니다.

나	너	그녀	공부해	식사	사랑
0.01	0.05	0.1	0.7	0.01	0.1

"학습 data 多多 Level Noise" ← Label-smoothing

(같은 Input → 다른 outs 가능) 일때 도움을 준다.

↕ One-Hot Encoding

Fully Connected layer

$\max(0, xW_1+b_1)W_2+b_2$

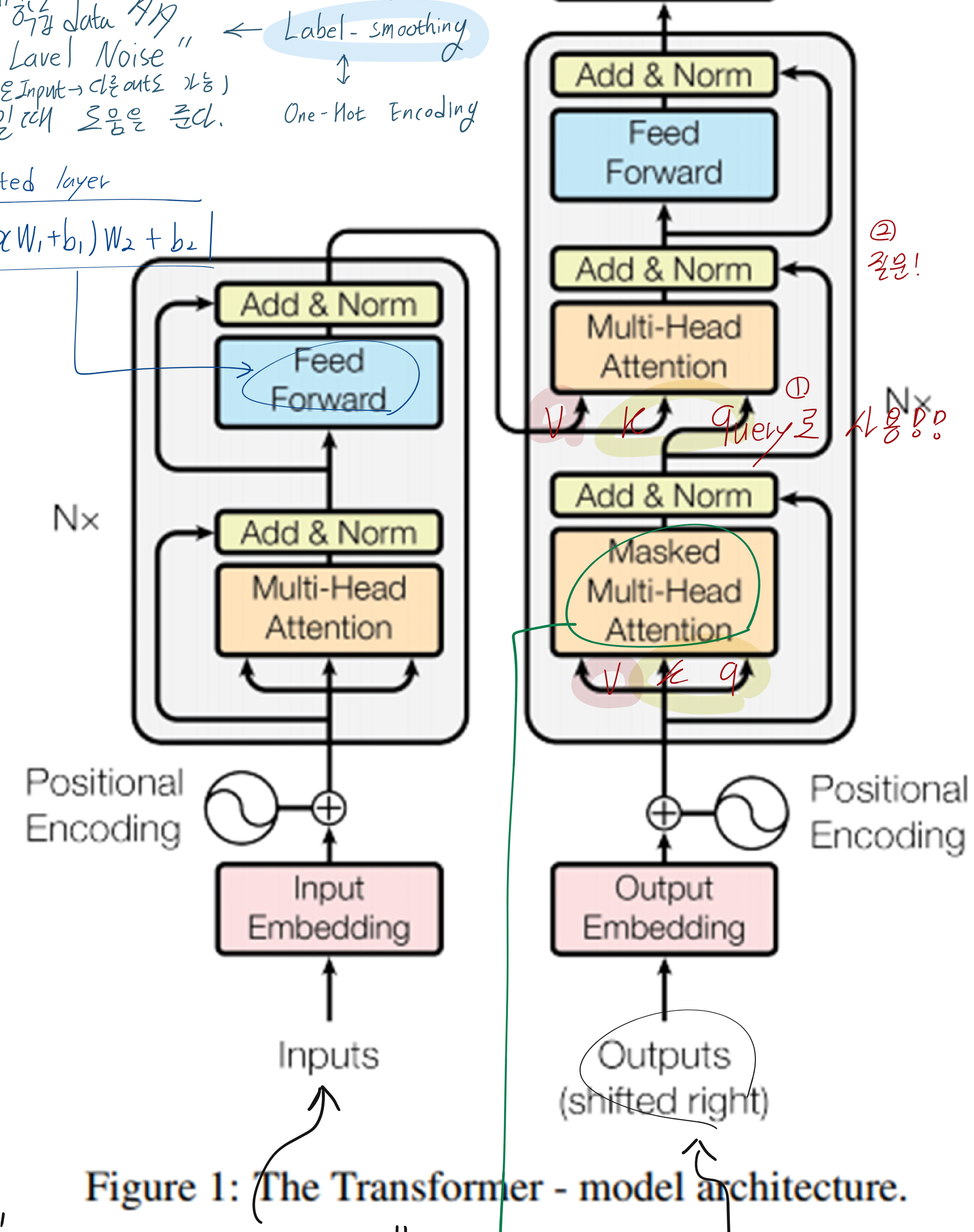
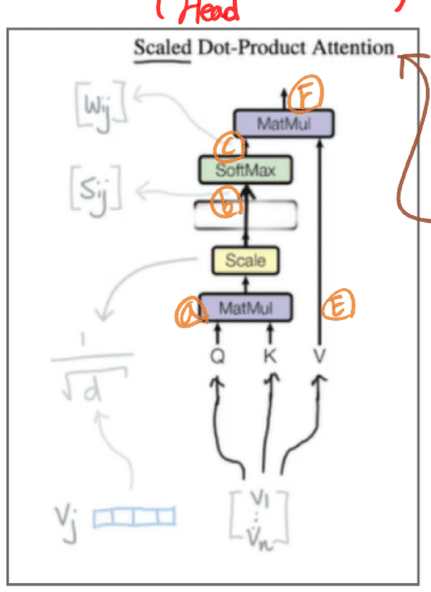
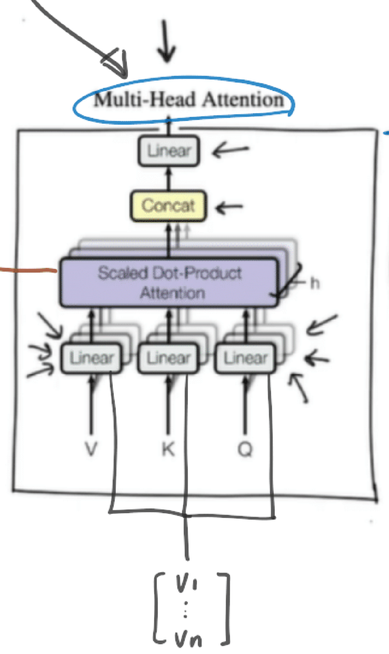
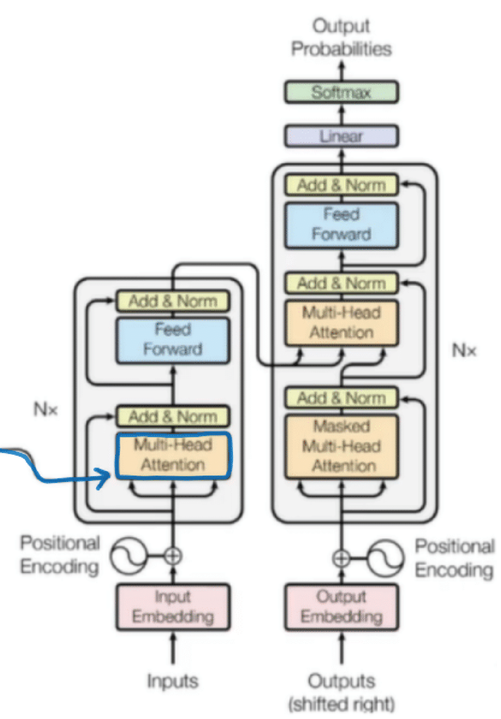


Figure 1: The Transformer - model architecture.

"I study at school"

"난 학교 에서"

지금까지 decoder에서 들어온 값들에만 Attention을 적용함

Transformer

= { Self Head Attention }