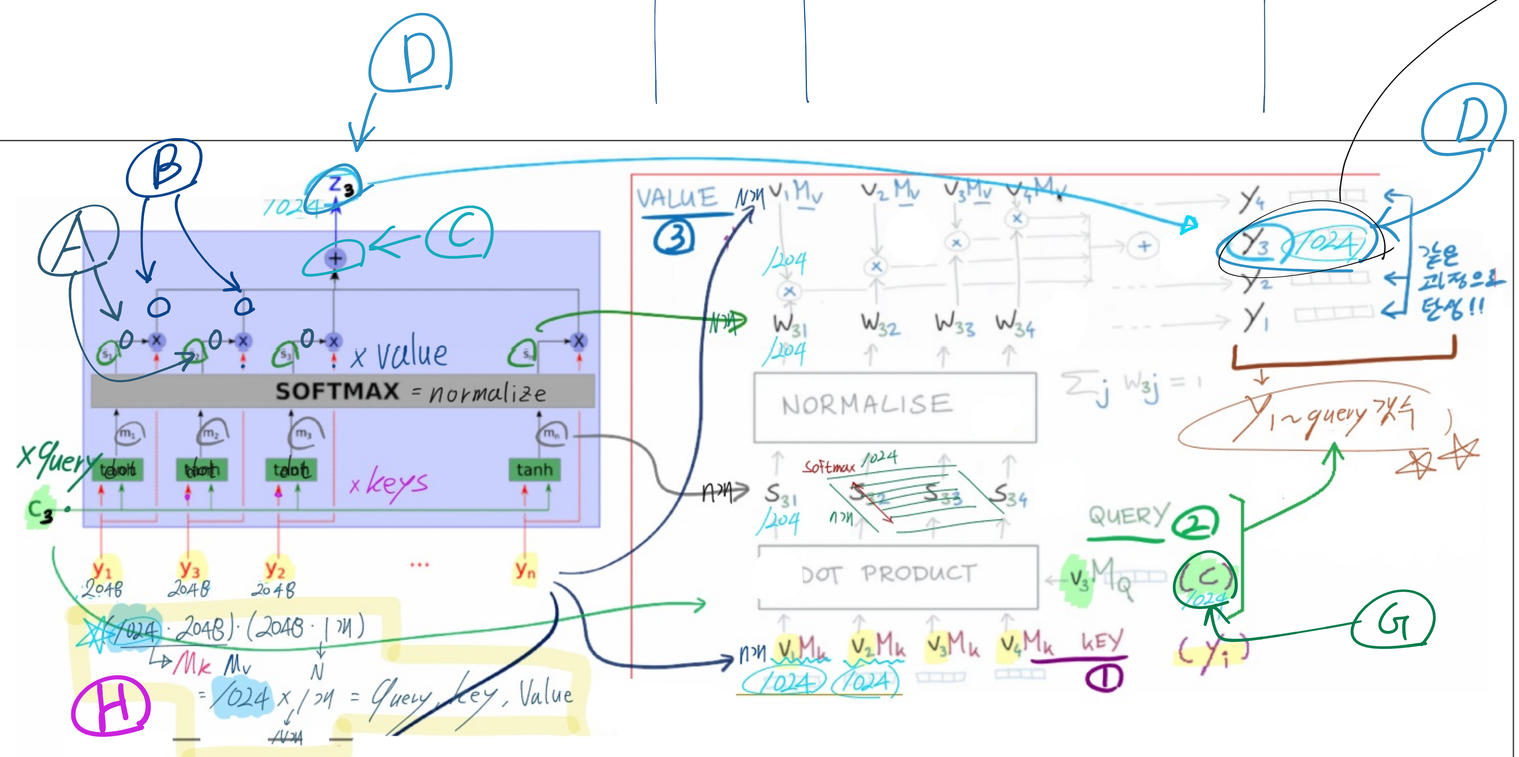
아래 그림처럼
이게 query에 대해서 만!

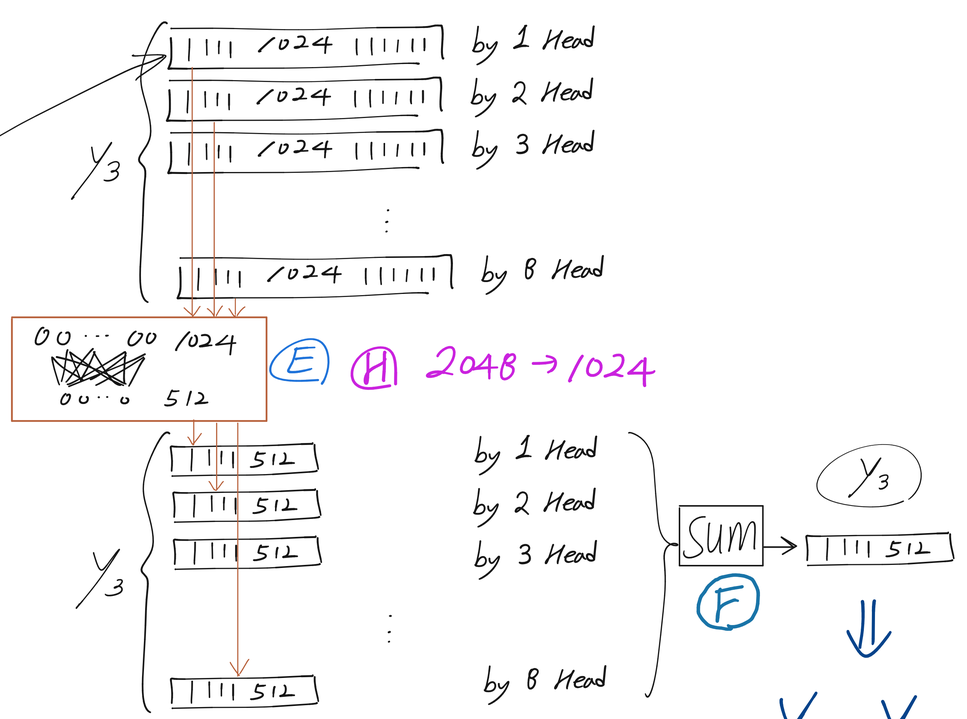
$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \Big], \tag{1}$$

where m indexes the attention head, $\boldsymbol{W}'_m \in \mathbb{R}^{C_v \times C}$ and $\boldsymbol{W}_m \in \mathbb{R}^{C \times C_v}$ are of learnable weights ($C_v = C/M$ by default). The attention weights $A_{mqk} \propto \exp\{\frac{\boldsymbol{z}_q^T \boldsymbol{U}_m^T \cdot \boldsymbol{V}_m \boldsymbol{x}_k}{\sqrt{C_v}}\}$ are normalized as

EX) 13x13x256 ← Feature map ← encoder ← object query
C = 256 N_k = 13·13개 N_q = 13·13 or 100개 (ex)
C_v = 32 (C/M) M = 8 (Head 갯수) C' = 64

2048 1024 512





K_q 모두 고려할 때

$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \Big], \tag{1}$$

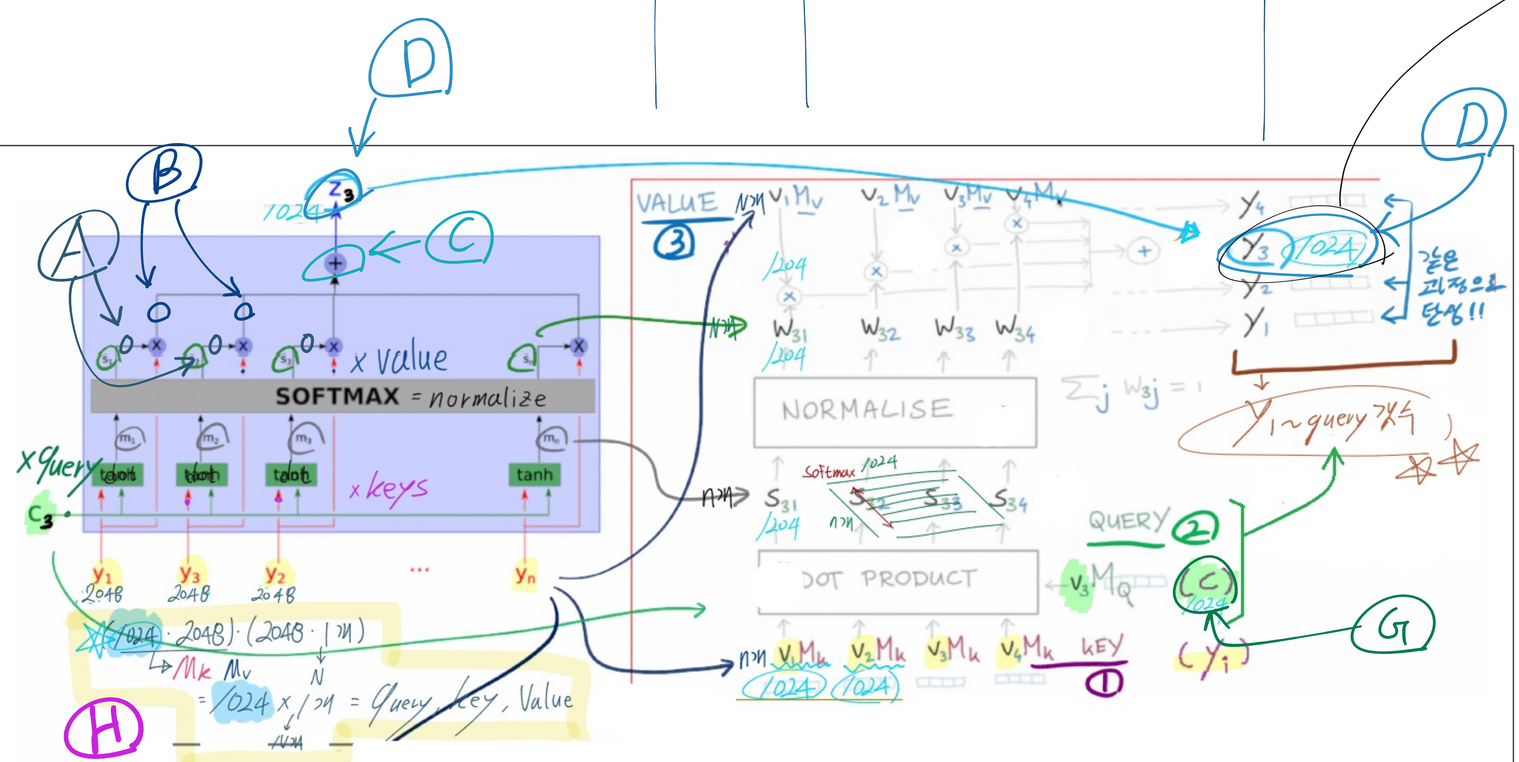
where m indexes the attention head, $\boldsymbol{W}'_m \in \mathbb{R}^{C_v \times C}$ and $\boldsymbol{W}_m \in \mathbb{R}^{C \times C_v}$ are of learnable weights ($C_v = C/M$ by default). The attention weights $A_{mqk} \propto \exp\{\frac{\boldsymbol{z}_q^T \boldsymbol{U}_m^T \cdot \boldsymbol{V}_m \boldsymbol{x}_k}{\sqrt{C_v}}\}$ are normalized as

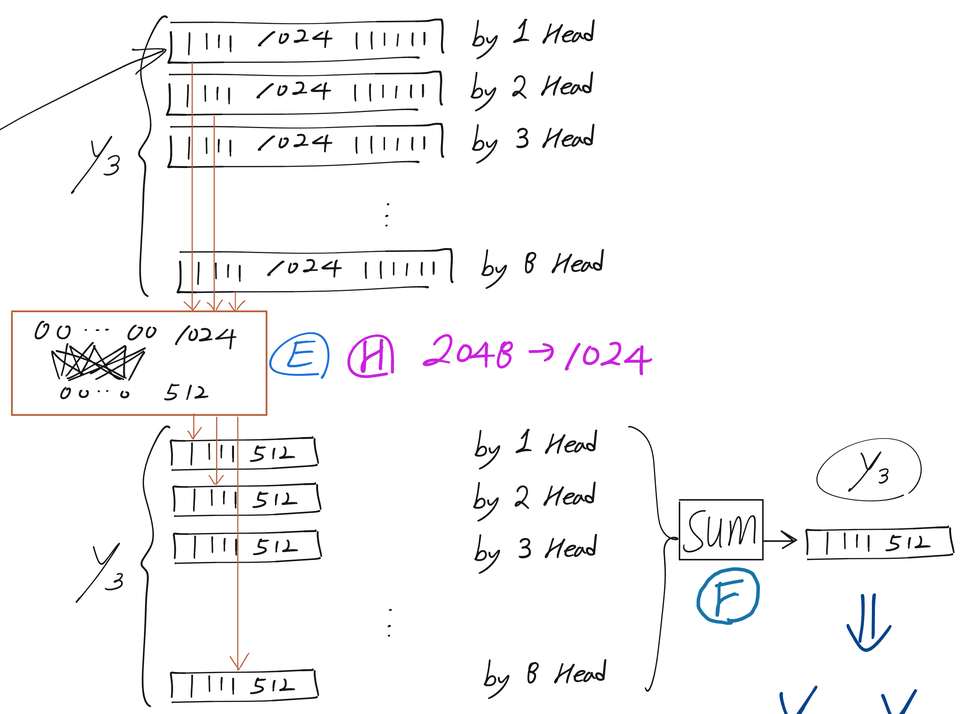
EX) 13×13×256 ← Feature map ← encoder ← object query

$C = 256$ $N_k = 13 \cdot 13$개 $N_q = 13 \cdot 13$ or 100개 (ex)

$C_v = 32$ (C/M) $M = 8$ (Head 개수) $C' = 64$

2048 1024 512





K_q 모두 고려할 때

$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \Big], \tag{1}$$

where m indexes the attention head, $\boldsymbol{W}'_m \in \mathbb{R}^{C_v \times C}$ and $\boldsymbol{W}_m \in \mathbb{R}^{C \times C_v}$ are of learnable weights ($C_v = C/M$ by default). The attention weights $A_{mqk} \propto \exp\{\frac{\boldsymbol{z}_q^T \boldsymbol{U}_m^T \cdot \boldsymbol{V}_m \boldsymbol{x}_k}{\sqrt{C_v}}\}$ are normalized as

EX) 13×13×256 ← Feature map ← encoder ← object query

C = 256 N_k = 13·13개 N_q = 13·13 or 100개 (ex)

C_v = 32 (C/M) M = 8 (Head 갯수)

① 1개의 key에게 "query에 대해 어떻게 생각해??"

② 1개의 key : "나는 query를 이만큼 집중해 = 관심있어"

③ Key ≈ Value : "그니까 난 딱 W_{qk}만큼만 관여할게~!"

④ 우리가 각각 관여한 정도를 다 모으면!

⑤ C'은 '우리의 생각 모음'을 여러번 고려해볼 걸괴야!

⑥ 이 짓을 (K_q번 반복해