

이제 MultiHeadAttn을 살펴보자.
 MultiHeadAttn은 query와 key, value로 이루어진 여러 헤드를 사용하는 구조이다.

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

$\sum_{m=1}^M \overline{W}_m \left[\sum_{k \in \Omega_k} A_{mqk} \cdot \overline{W}'_m x_k \right],$

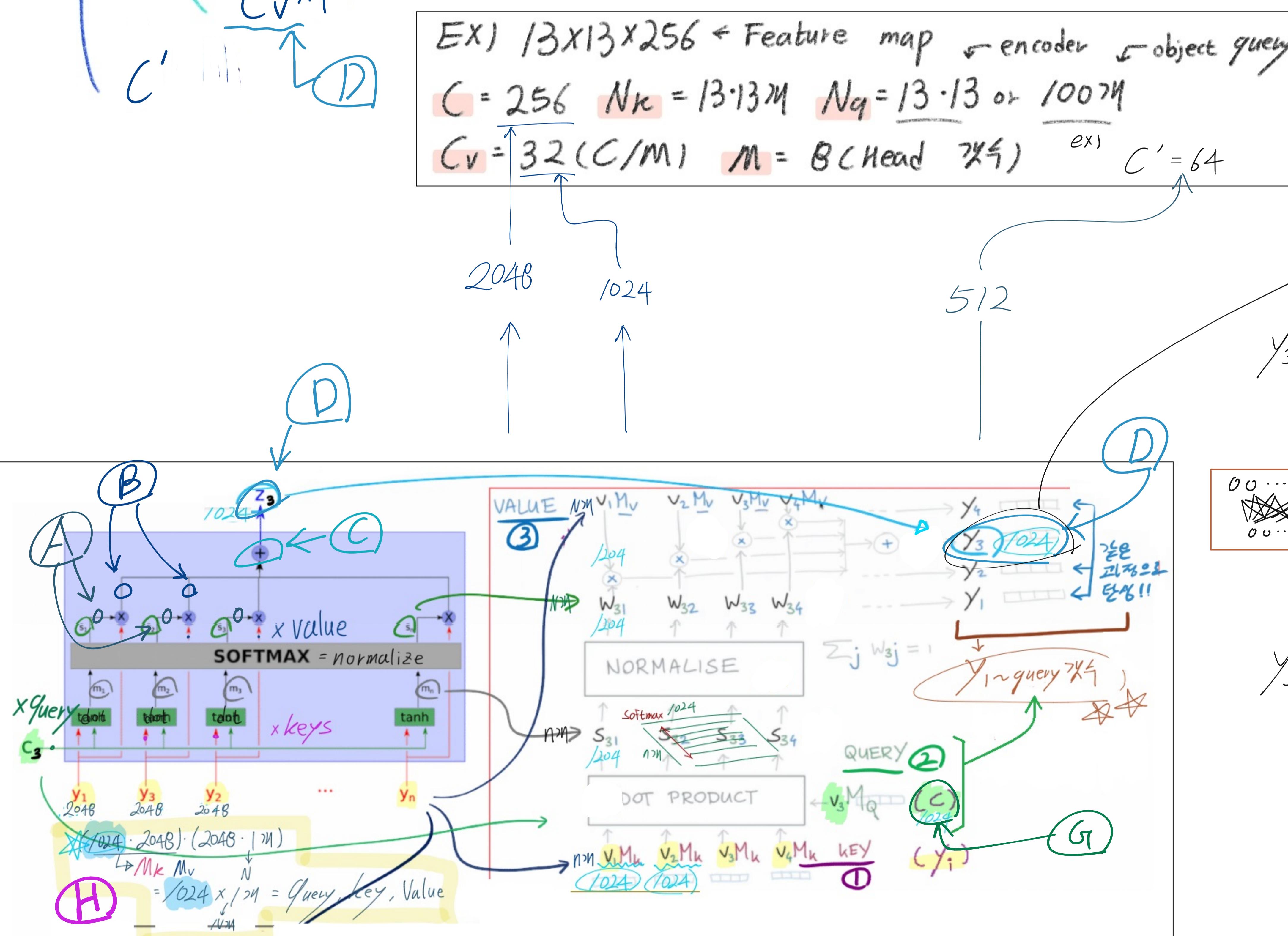
= $C_v ((C_v \cdot C) (C \times 1))$ element-wise multiple
 = $C_v (C_v \cdot 1) \leftarrow \text{Value}$

$C_v \leftarrow \text{key } 1 \text{ in one channel}$
 $\rightarrow N_k \times (HW^2)$

① $\rightarrow \text{Softmax}$
 3
 ③ shape (A)
 ⑤ $B \in \text{key} \otimes \text{value}$

EX) $13 \times 13 \times 256 \leftarrow \text{Feature map} \rightarrow \text{encoder} \rightarrow \text{softmax}$
 $C_v \rightarrow 13 \times 13 \times 256 \rightarrow 13 \times 13 \times 1 \rightarrow 13 \times 13 \times 256$

$\text{head} \leftarrow m=1$ $k \in \Omega_k$
 $z_q \in \mathbb{R}^{C_v \times C}$ and $W_m \in \mathbb{R}^{C \times C_v}$ are of learnable weights
weights $A_{mqk} \propto \exp\left\{\frac{z_q^T U_m^T \cdot V_m x_k}{\sqrt{C_v}}\right\}$ are normalized as
① \rightarrow Softmax
② query
③ Learnable Weight key
④ E^m query β^{Σ} key
⑤ Shape(A_{mqk}) = $(C^q \times 1)^T \times (C_v \cdot C)^T \times ((C_v \cdot C) \times N_k)$
= $(1 \times C_v)$
 $\times (C_v \times N_k)$
dot
dot
⑥ $G_1 = C_v \times N_k$
⑦ β^{Σ} key of size $A_{mqk} = C_v \times N_k \xrightarrow{\text{Softmax}} C_v \times N_k$



K_q 를 고려한 $\text{MultiHeadAttn}(z_q, x) = \sum_{m=1}^M W_m \left[\sum_k A_{mqk} \cdot W'_m x_k \right]$, (1)

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

$\sum_{m=1}^M \mathbf{W}_m \left[\sum_{k \in \Omega_k} \frac{A_{mqk}}{\sqrt{C_v}} \cdot \mathbf{W}'_m \mathbf{x}_k \right],$

element-wise multiple

value

key 1개의 차원

key 개수

$C \times C_v$

$C_v \times K_q$

$C' \times K_q$

$C' \times C_v$

$C \times C_v$

$C_v \times C$

$C_v \times 1$

$K_q \times C_v$

$K_q \cdot C_v$

$K_q \cdot C_v \cdot N_k \cdot (HW)^2$

$N_k = 13 \cdot 13 \cdot 256$

$N_q = 13 \cdot 13 \text{ or } 100 \cdot 256$

① \rightarrow Softmax

② \rightarrow 3

③ Shape (A)

④ Shape (B)

⑤ Shape (C)

EX) $13 \times 13 \times 256 \leftarrow$ Feature map \leftarrow encoder \leftarrow 0

③ $\text{shape}(A_{mqk}) = ((C^q \times 1)^T \times (C_v \cdot C^k)^T \times (C_v \cdot C) \times C \times N_k)$

$= (1 \times C_v)$

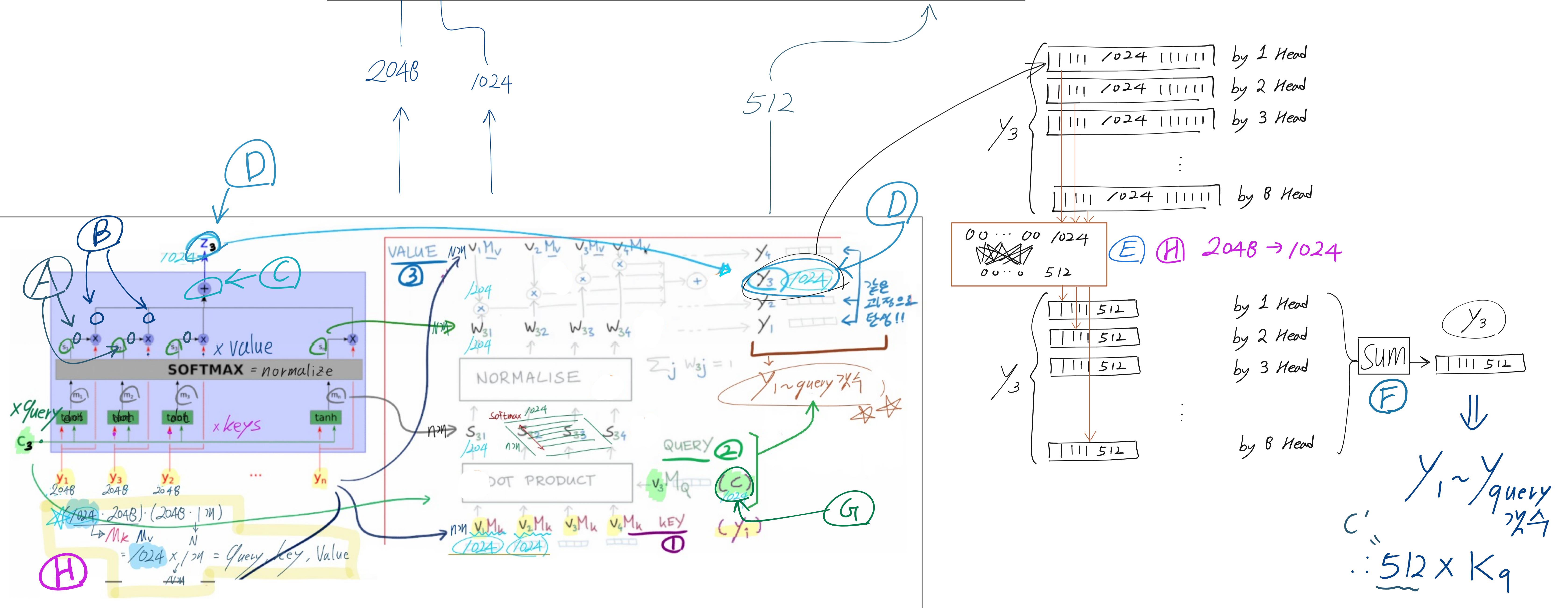
(G) $\rightarrow C_v \times N_k$

④ $\exists \text{는 key에 대해 } A_{mqk} = C_v \times N_k \xrightarrow{\text{Softmax}} C_v \times N_k$

cap \leftarrow encoder \leftarrow object query
 $13 \cdot 13 \text{ or } 100\text{개}$
 ex) $C' = 64$

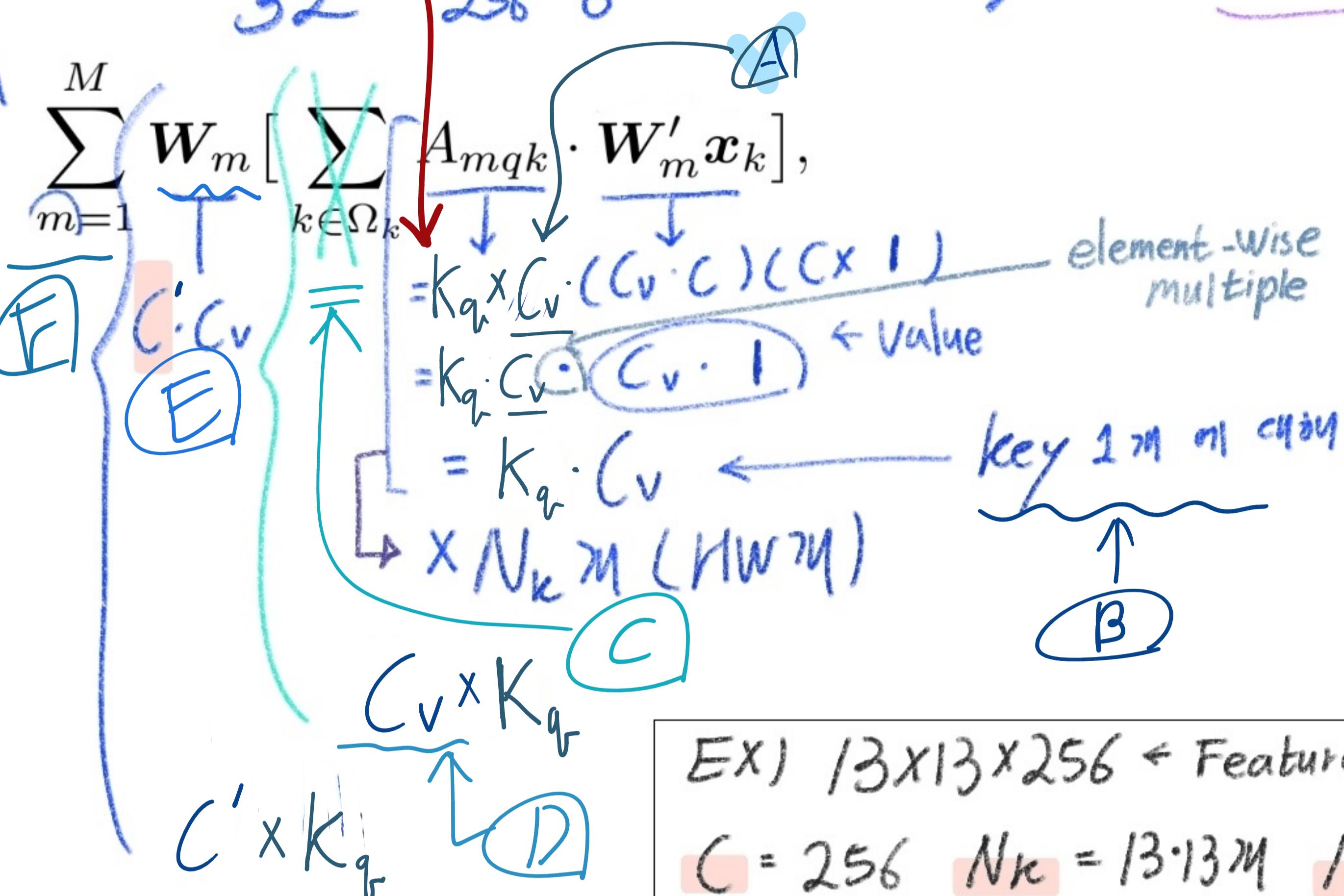
중에서 $12N_k$
 $= A$

$\begin{cases} C_v \\ N_k\text{개} \end{cases} \xrightarrow{\text{Softmax}}$



$$K_q \Leftarrow \text{MultiHeadAttn}(z_q, x) = \sum_{m=1}^M W_m \left[\sum_{k \in \Omega_k} A_{mqk} \cdot W'_m x_k \right], \quad (1)$$

where m indexes the attention head, $W'_m \in \mathbb{R}^{C_v \times C}$ and $W_m \in \mathbb{R}^{C \times C_v}$ are of learnable weights ($C_v = C/M$ by default). The attention weights $A_{mqk} \propto \exp\left\{\frac{z_q^T U_m^T \cdot V_m x_k}{\sqrt{C_v}}\right\}$ are normalized as

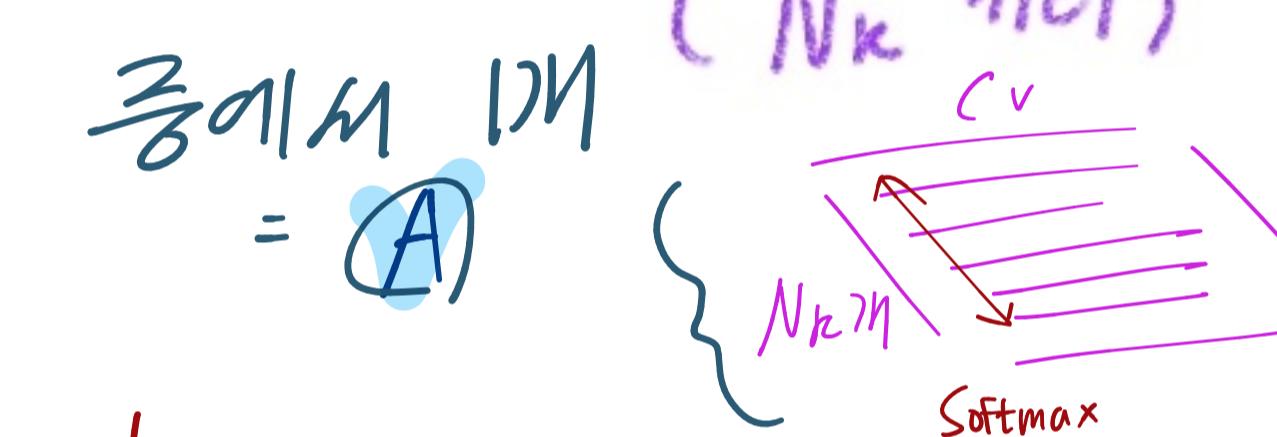


3

③ Shape (A_{mqk})

$$(G) = (C_v \times 1)^T \times (C_v \cdot C)^T \xrightarrow{\text{dot}} C_v \times N_k$$

⑤ $\exists \in \text{key} \text{에 } \text{값 } A_{mqk} = C_v \times N_k \xrightarrow{\text{Softmax}} C_v \times N_k$



① 1개의 key에게 "query에 대해 어떻게 생각해??"

② 1개의 key : "나는 query를 어떤 걸 학습해야 학습할까?"

③ Key \approx Value : "그냥 난 딱 W_{qk} 만큼만 관여할게~!"

④ 우리가 각각 관여한 정도를 다 모으면!
이게 물어본 하나의 q 에 대한 우리의 생각 모음이야!

⑤ C' 은 '우리의 생각 모음'을 여러번 고려해본 결과야!
하나의 q 에 대해

