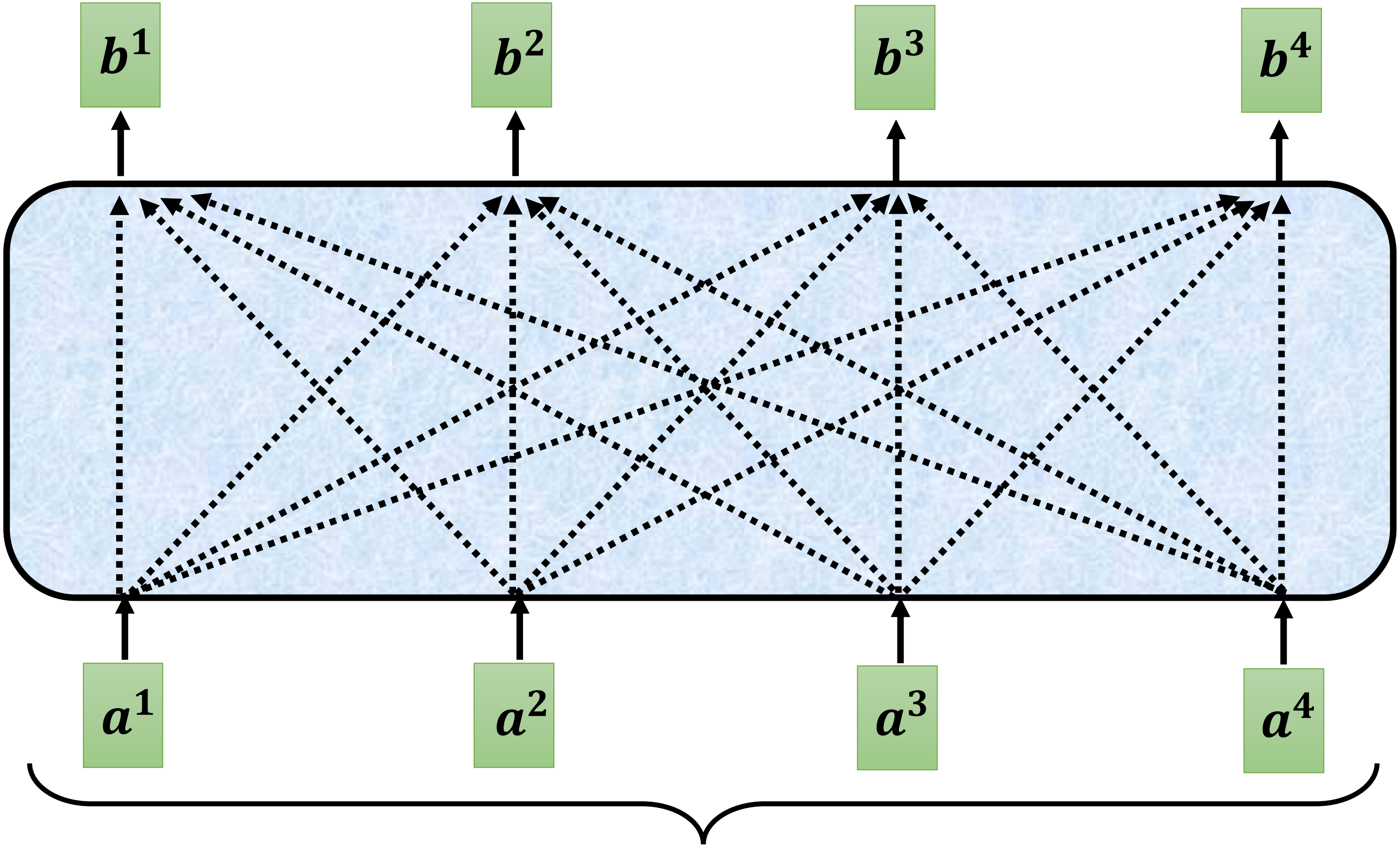
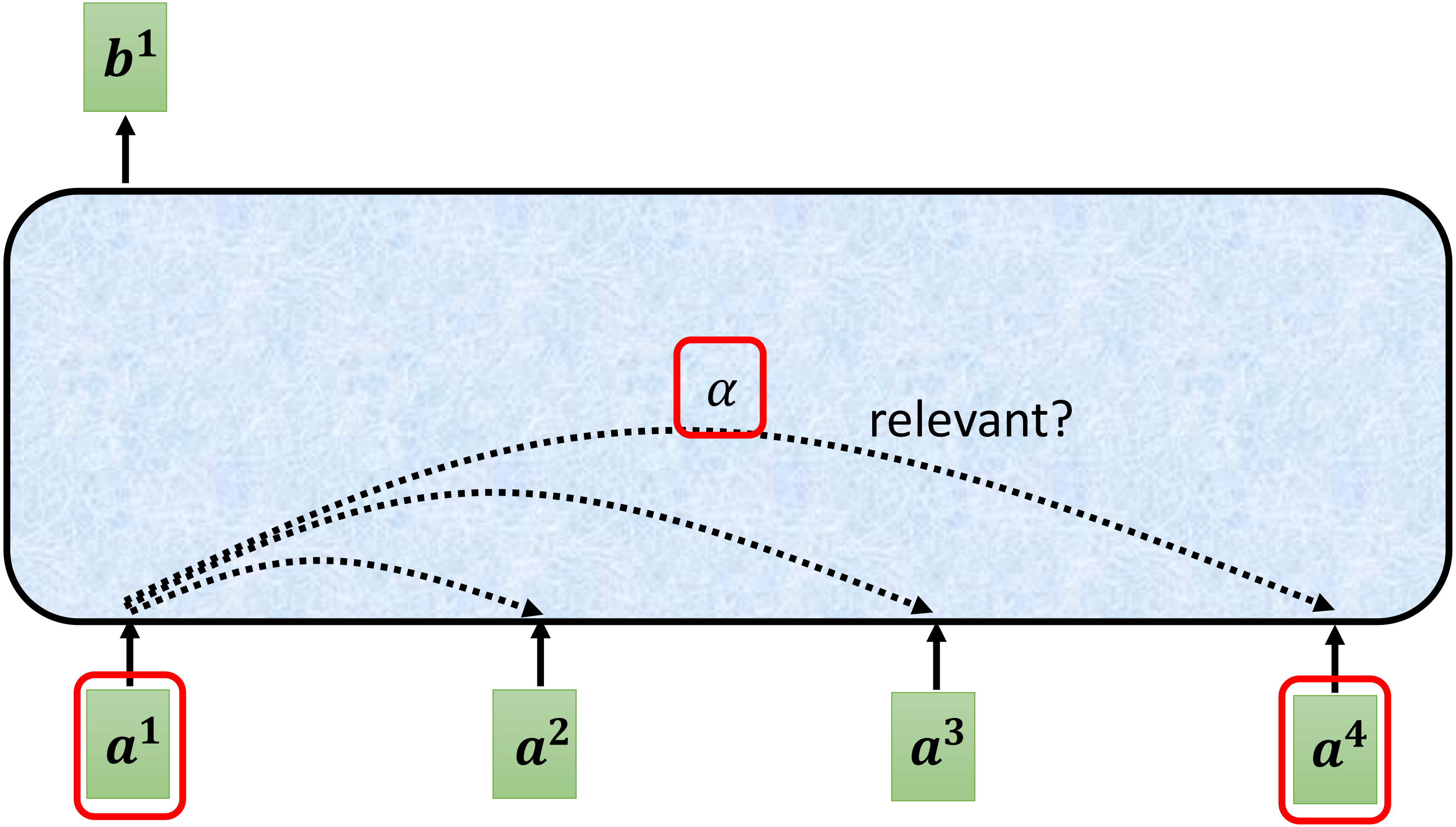
Self-attention



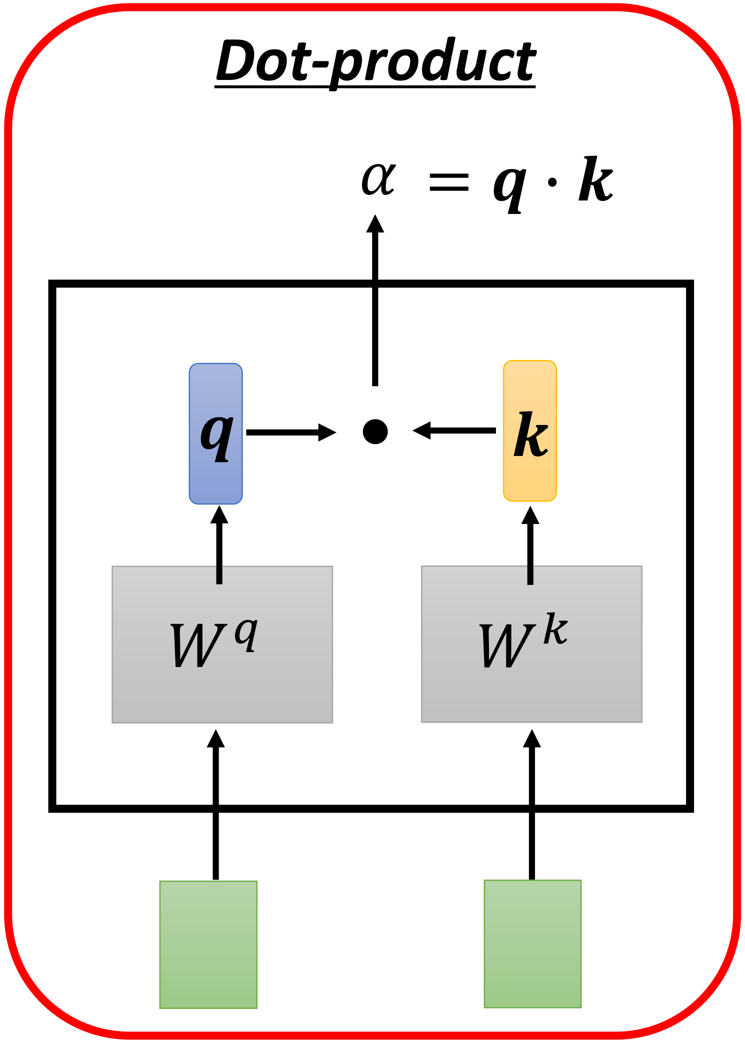
Can be either **input** or **a hidden layer**

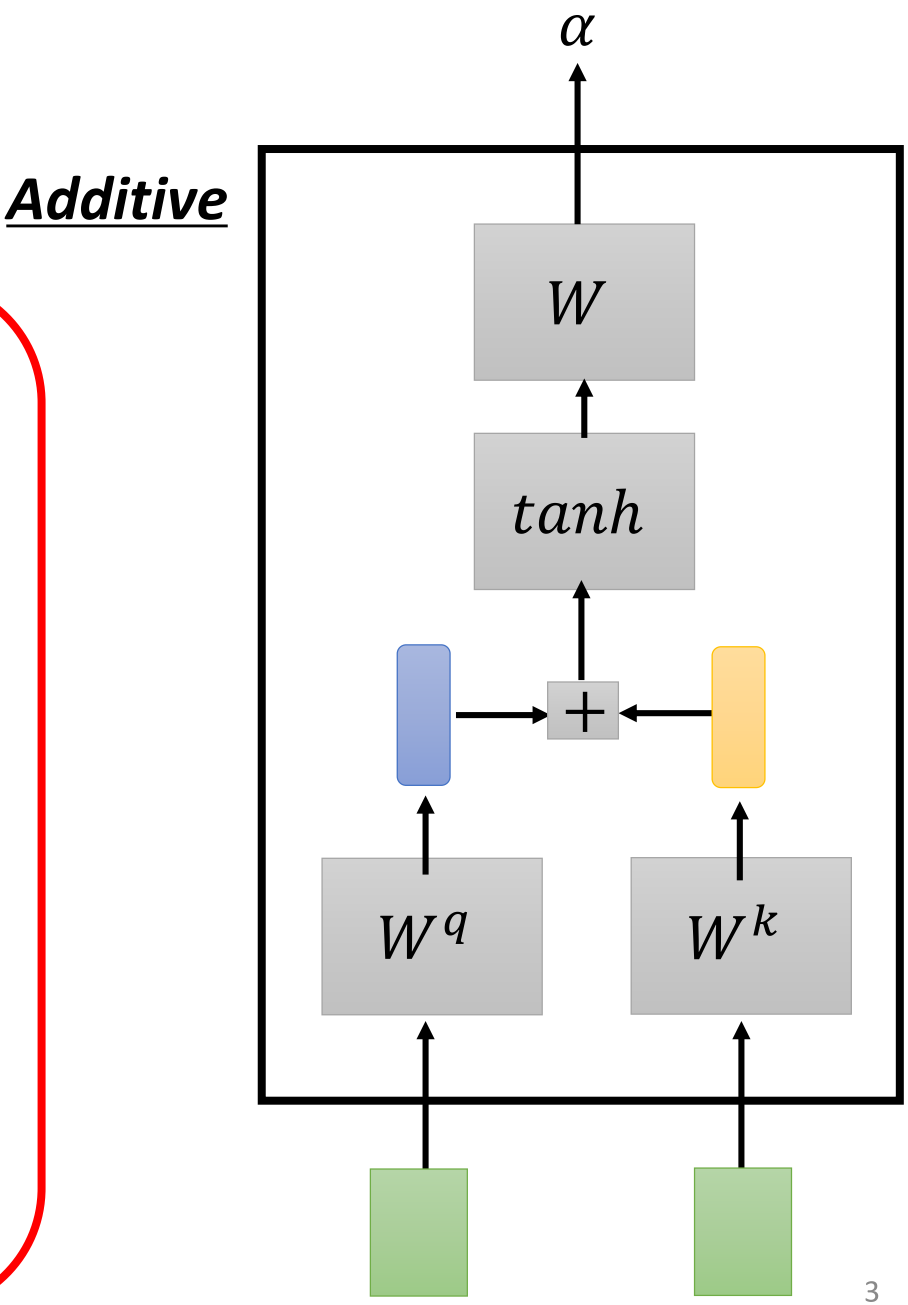
Self-attention



Find the relevant vectors in a sequence

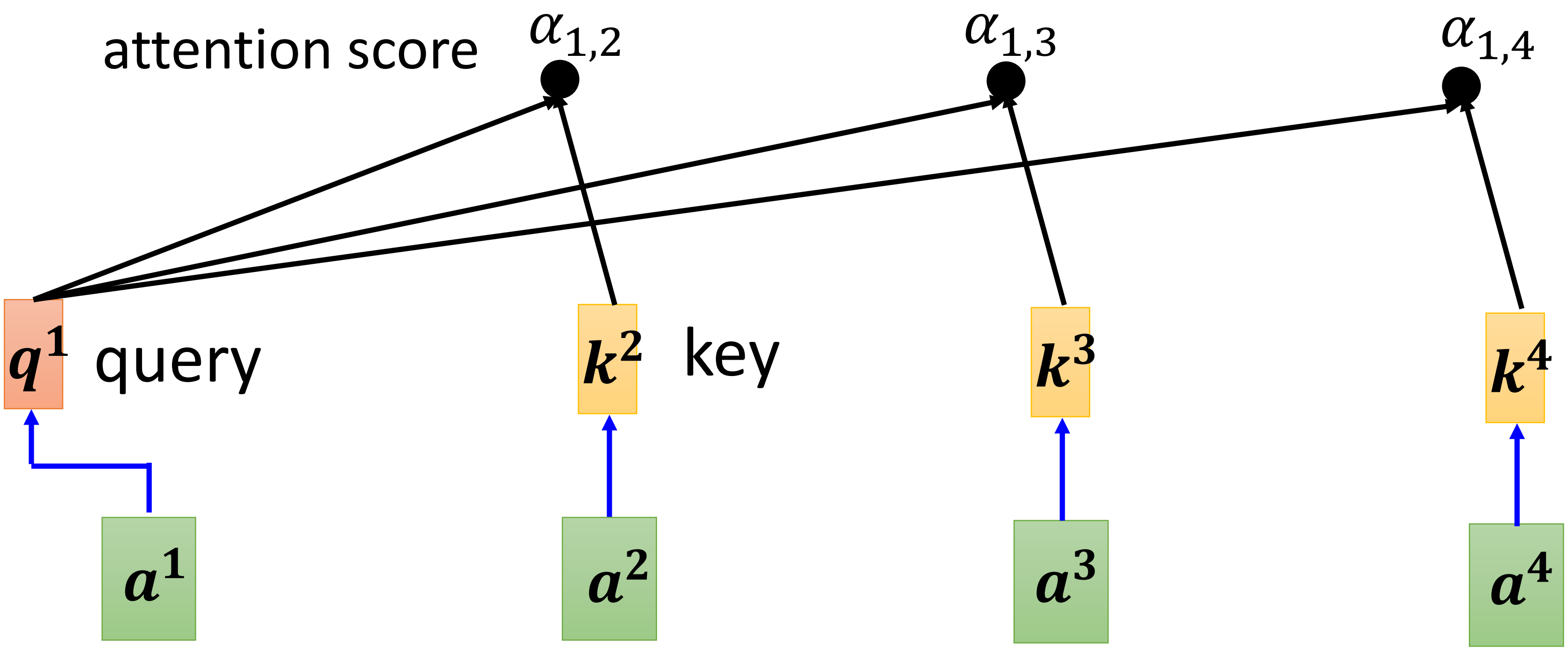
Self-attention





Self-attention

$\alpha_{1,2} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^2$ $\alpha_{1,3} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^3$ $\alpha_{1,4} = \boldsymbol{q}^1 \cdot \boldsymbol{k}^4$



attention score $\alpha_{1,2}$ $\alpha_{1,3}$ $\alpha_{1,4}$

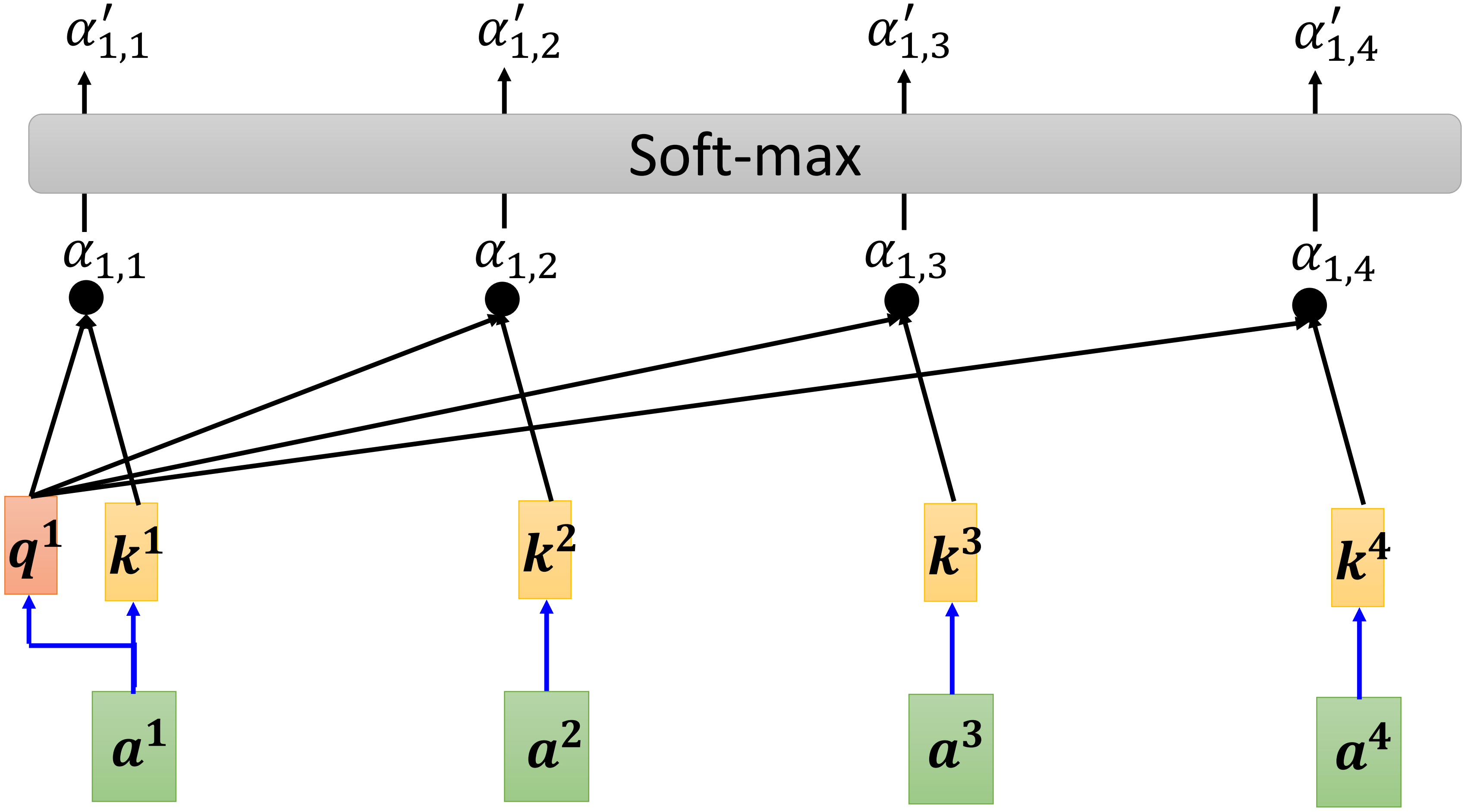
$\boldsymbol{q}^1$ query $\boldsymbol{k}^2$ key $\boldsymbol{k}^3$ $\boldsymbol{k}^4$

$\boldsymbol{a}^1$ $\boldsymbol{a}^2$ $\boldsymbol{a}^3$ $\boldsymbol{a}^4$

$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1$ $\boldsymbol{k}^2 = W^k \boldsymbol{a}^2$ $\boldsymbol{k}^3 = W^k \boldsymbol{a}^3$ $\boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

Self-attention

$$\alpha'_{1,i} = exp(\alpha_{1,i})/\sum_j exp(\alpha_{1,j})$$



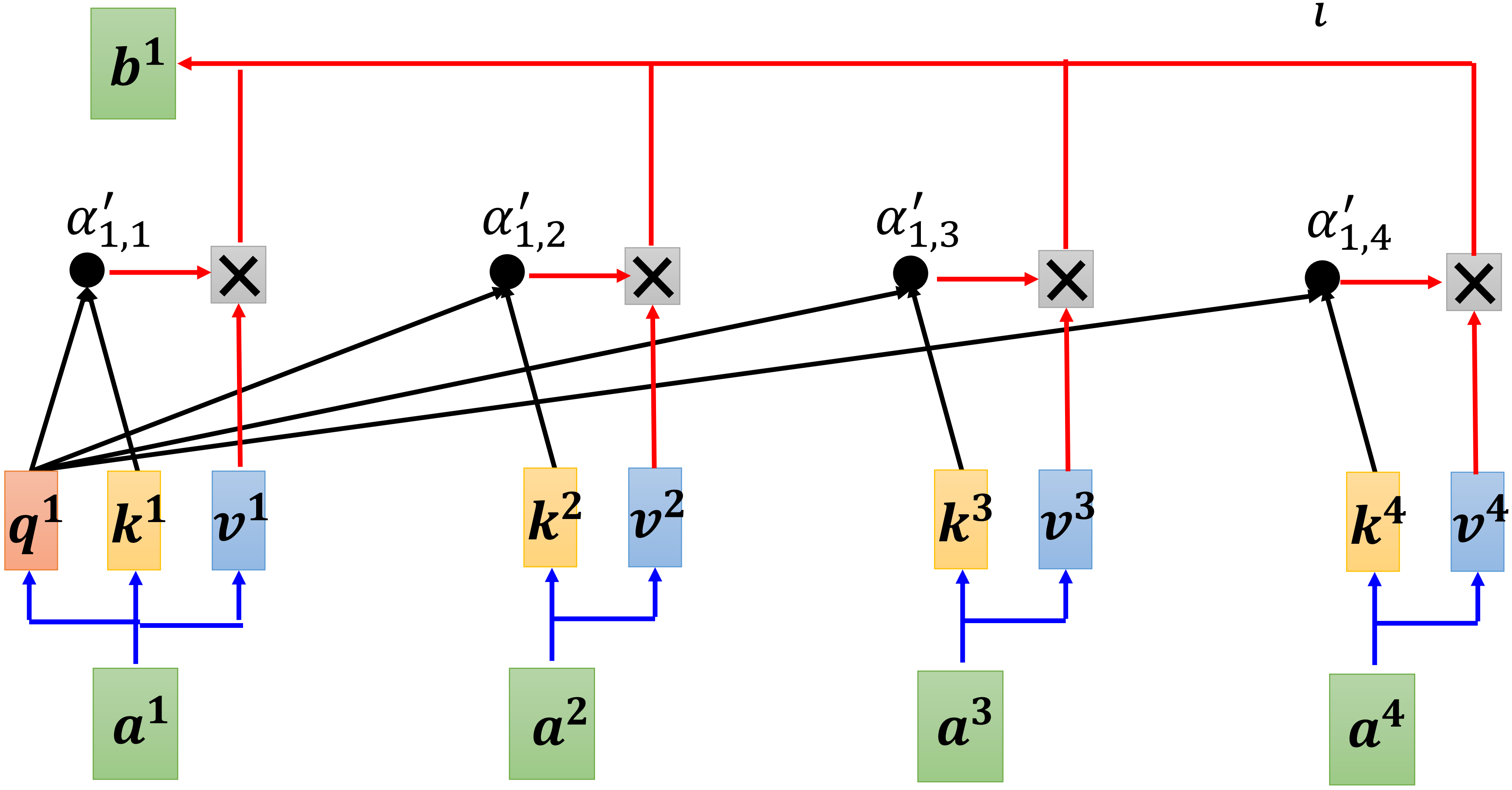
$\boldsymbol{q}^1 = W^q \boldsymbol{a}^1$ $\boldsymbol{k}^2 = W^k \boldsymbol{a}^2$ $\boldsymbol{k}^3 = W^k \boldsymbol{a}^3$ $\boldsymbol{k}^4 = W^k \boldsymbol{a}^4$

$\boldsymbol{k}^1 = W^k \boldsymbol{a}^1$

Self-attention

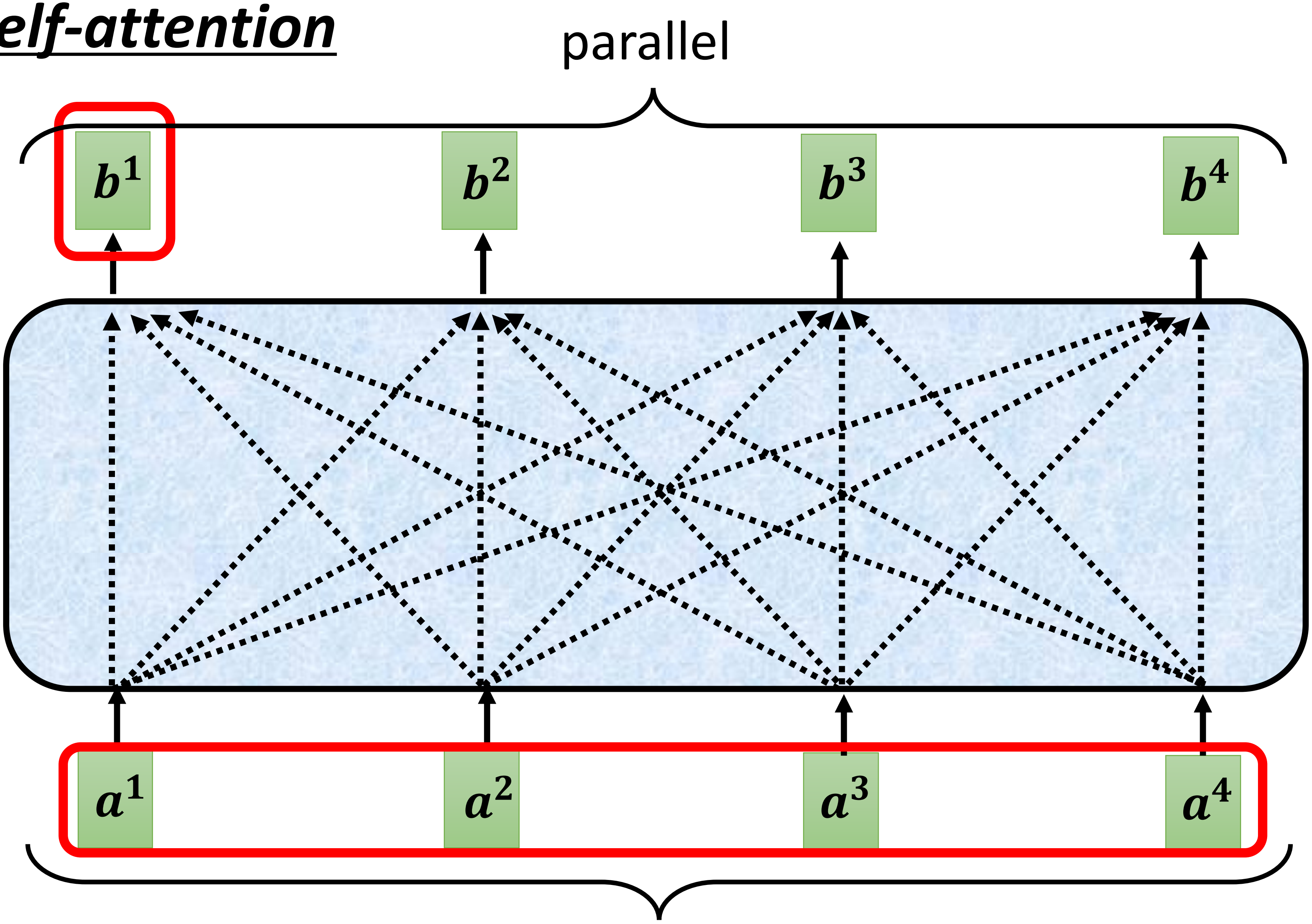
Extract information based on attention scores

$$\boldsymbol{b}^1 = \sum_i \alpha'_{1,i} \boldsymbol{v}^i$$



$\boldsymbol{v}^1 = W^v \boldsymbol{a}^1$ $\quad$ $\boldsymbol{v}^2 = W^v \boldsymbol{a}^2$ $\quad$ $\boldsymbol{v}^3 = W^v \boldsymbol{a}^3$ $\quad$ $\boldsymbol{v}^4 = W^v \boldsymbol{a}^4$

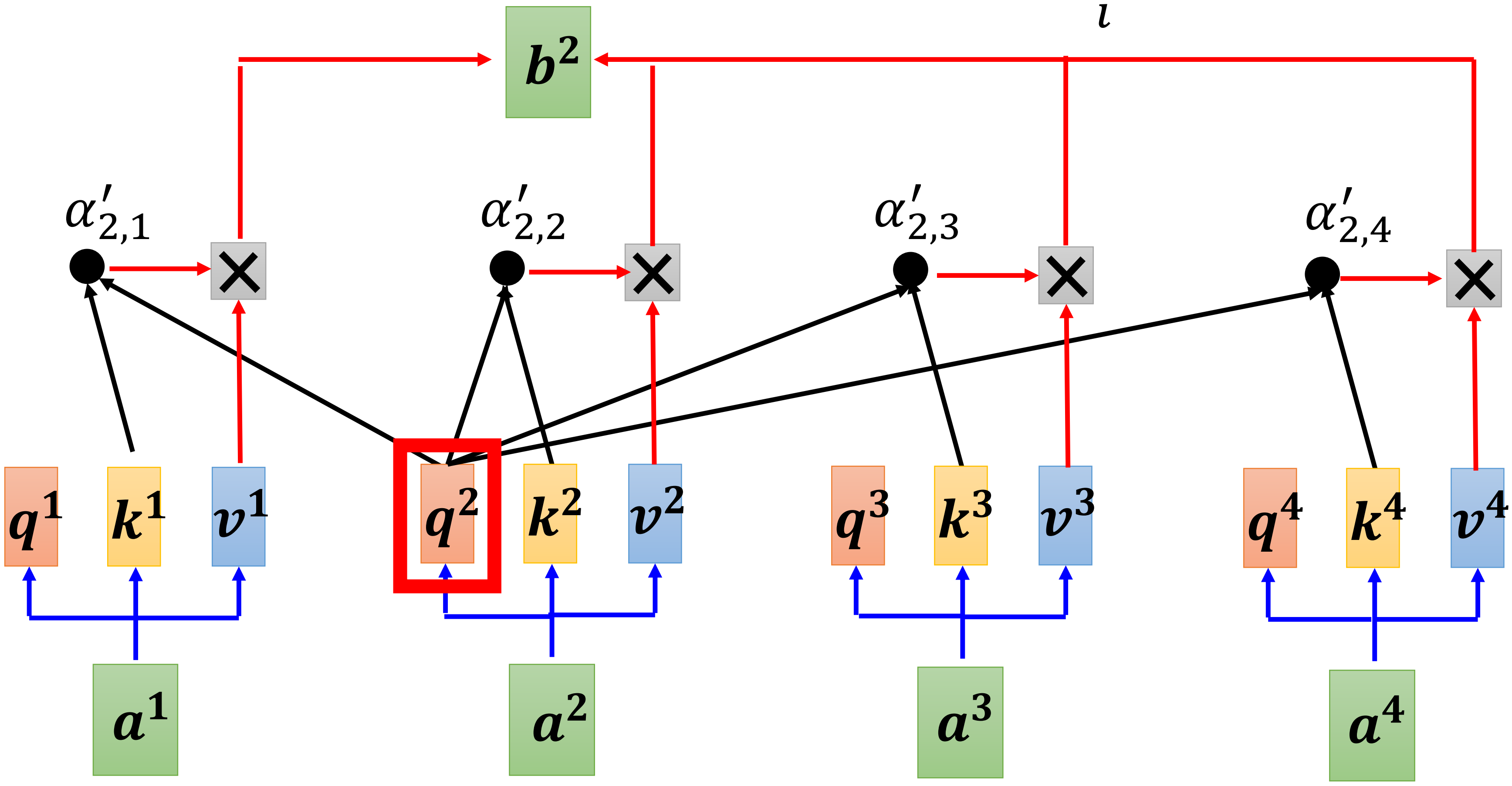
Self-attention



Can be either **input** or **a hidden layer**

Self-attention

$$\boldsymbol{b}^2 = \sum_i \alpha'_{2,i} \boldsymbol{v}^i$$



Self-attention

$$q^i = W^q a^i$$

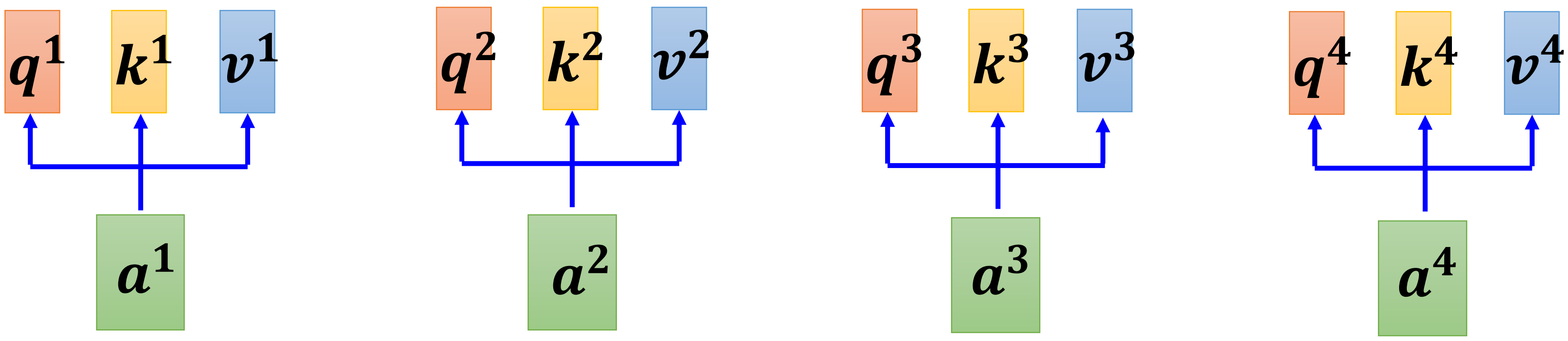
$$\underbrace{\begin{bmatrix} q^1 & q^2 & q^3 & q^4 \end{bmatrix}}_{Q} = W^q \underbrace{\begin{bmatrix} a^1 & a^2 & a^3 & a^4 \end{bmatrix}}_{I}$$

$$k^i = W^k a^i$$

$$\underbrace{\begin{bmatrix} k^1 & k^2 & k^3 & k^4 \end{bmatrix}}_{K} = W^k \underbrace{\begin{bmatrix} a^1 & a^2 & a^3 & a^4 \end{bmatrix}}_{I}$$

$$v^i = W^v a^i$$

$$\underbrace{\begin{bmatrix} v^1 & v^2 & v^3 & v^4 \end{bmatrix}}_{V} = W^v \underbrace{\begin{bmatrix} a^1 & a^2 & a^3 & a^4 \end{bmatrix}}_{I}$$

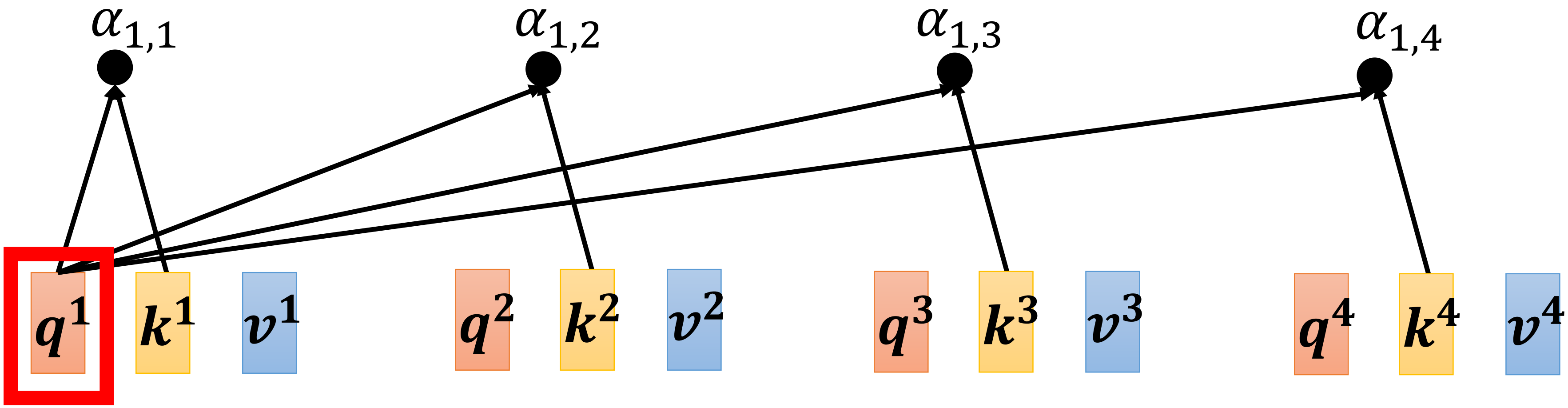


Self-attention

$\alpha_{1,1} = k^1 \; q^1$ $\quad \alpha_{1,2} = k^2 \; q^1$

$\alpha_{1,3} = k^3 \; q^1$ $\quad \alpha_{1,4} = k^4 \; q^1$

$$\begin{bmatrix} \alpha_{1,1} \\ \alpha_{1,2} \\ \alpha_{1,3} \\ \alpha_{1,4} \end{bmatrix} = \begin{bmatrix} k^1 \\ k^2 \\ k^3 \\ k^4 \end{bmatrix} q^1$$

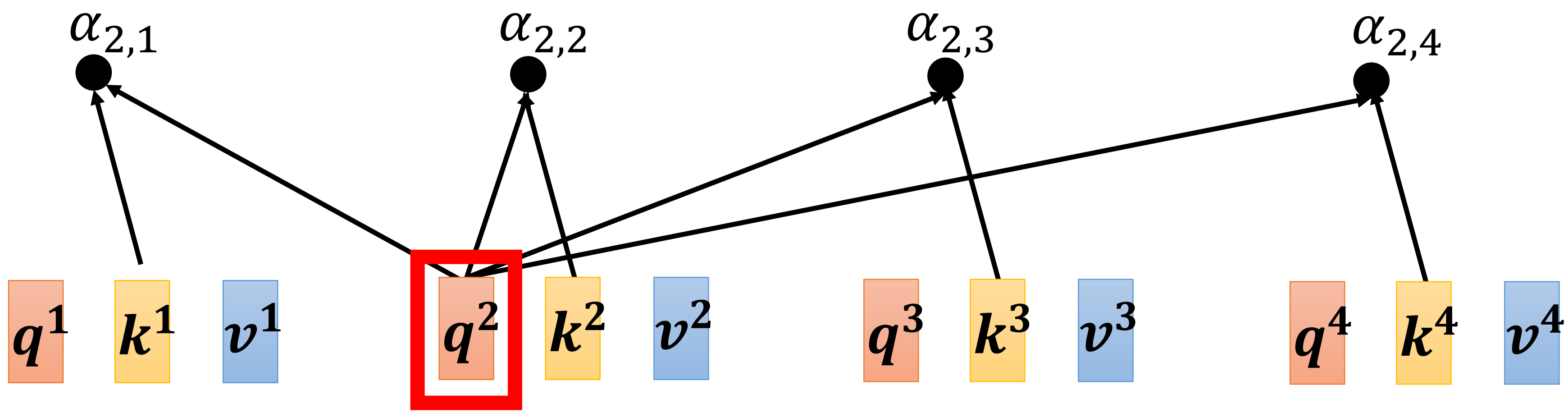


Self-attention

$$\alpha_{1,1} = \boldsymbol{k}^1 \boldsymbol{q}^1 \qquad \alpha_{1,2} = \boldsymbol{k}^2 \boldsymbol{q}^1$$

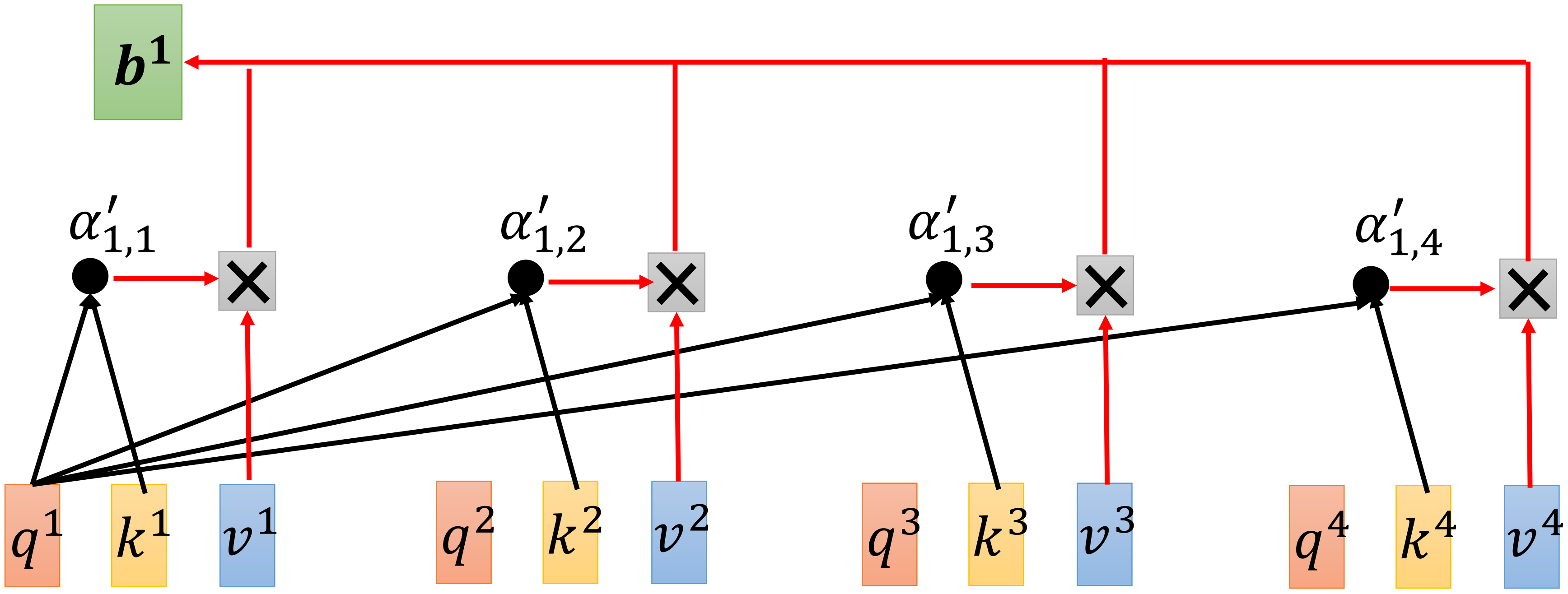
$$\alpha_{1,3} = \boldsymbol{k}^3 \boldsymbol{q}^1 \qquad \alpha_{1,4} = \boldsymbol{k}^4 \boldsymbol{q}^1$$

$$\begin{matrix} \alpha_{1,1} \\ \alpha_{1,2} \\ \alpha_{1,3} \\ \alpha_{1,4} \end{matrix} = \begin{matrix} \boldsymbol{k}^1 \\ \boldsymbol{k}^2 \\ \boldsymbol{k}^3 \\ \boldsymbol{k}^4 \end{matrix} \; \boldsymbol{q}^1$$


$\alpha_{2,1}$
$\alpha_{2,2}$
$\alpha_{2,3}$
$\alpha_{2,4}$
$\boldsymbol{q}^1$ $\boldsymbol{k}^1$ $\boldsymbol{v}^1$
$\boldsymbol{q}^2$ $\boldsymbol{k}^2$ $\boldsymbol{v}^2$
$\boldsymbol{q}^3$ $\boldsymbol{k}^3$ $\boldsymbol{v}^3$
$\boldsymbol{q}^4$ $\boldsymbol{k}^4$ $\boldsymbol{v}^4$

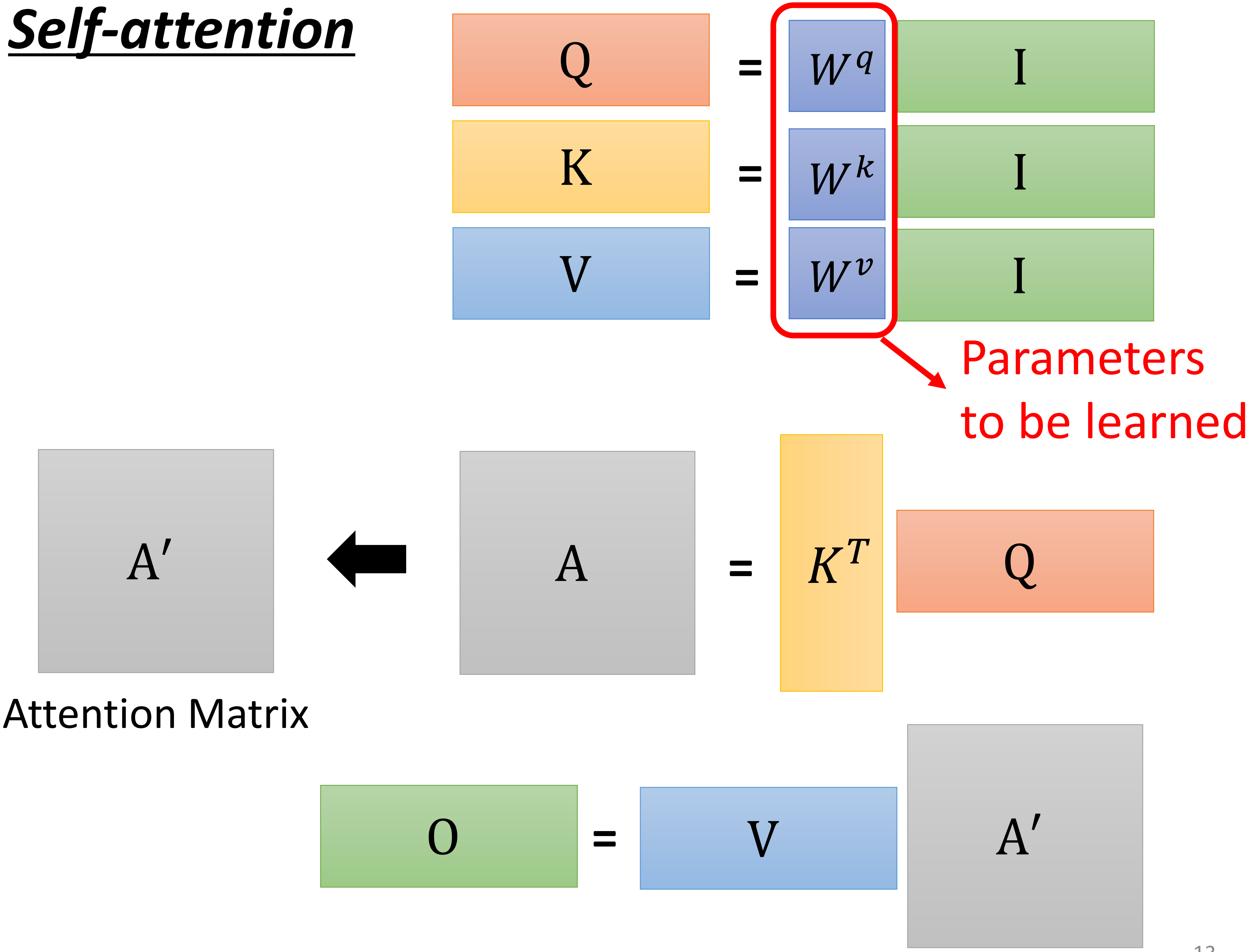
$$\underbrace{\begin{matrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \end{matrix}}_{A'} \; \underset{\text{softmax}}{\Leftarrow} \; \underbrace{\begin{matrix} \alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1} \\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2} \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3} \\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4} \end{matrix}}_{A} = \underbrace{\begin{matrix} \boldsymbol{k}^1 \\ \boldsymbol{k}^2 \\ \boldsymbol{k}^3 \\ \boldsymbol{k}^4 \end{matrix}}_{K^T} \; \underbrace{\begin{matrix} \boldsymbol{q}^1 & \boldsymbol{q}^2 & \boldsymbol{q}^3 & \boldsymbol{q}^4 \end{matrix}}_{Q}$$

Self-attention



$$\underbrace{\begin{bmatrix} b^1 & b^2 & b^3 & b^4 \end{bmatrix}}_{O} = \underbrace{\begin{bmatrix} v^1 & v^2 & v^3 & v^4 \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \end{bmatrix}}_{A'}$$

Self-attention



$$Q = W^q I$$

$$K = W^k I$$

$$V = W^v I$$

Parameters to be learned

$$A' \leftarrow A = K^T Q$$

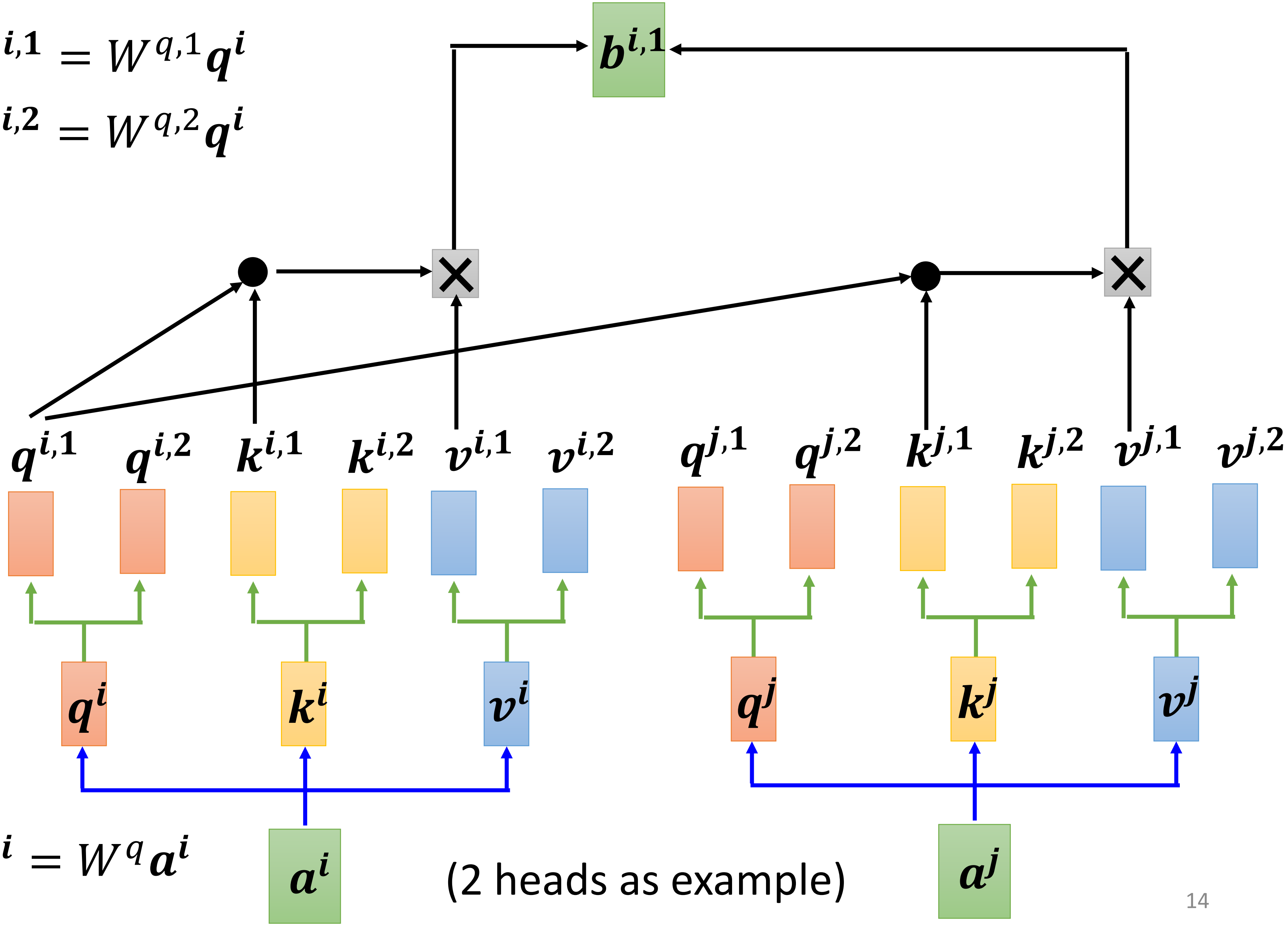
Attention Matrix

$$O = V A'$$

Multi-head Self-attention Different types of relevance

$$q^{i,1} = W^{q,1} q^i$$

$$q^{i,2} = W^{q,2} q^i$$



$$q^i = W^q a^i$$

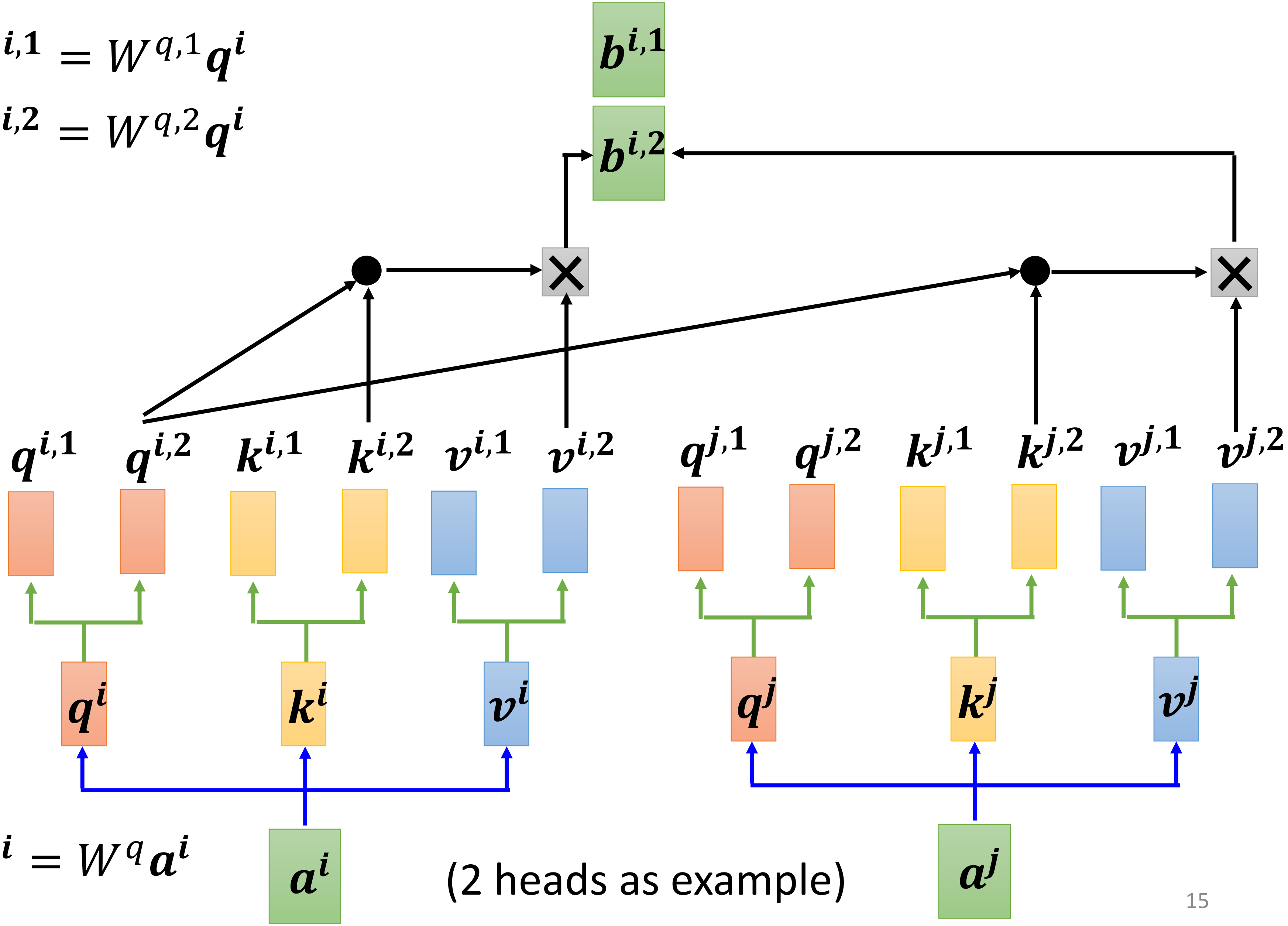
(2 heads as example)

Multi-head Self-attention

Different types of relevance

$$q^{i,1} = W^{q,1} q^i$$

$$q^{i,2} = W^{q,2} q^i$$



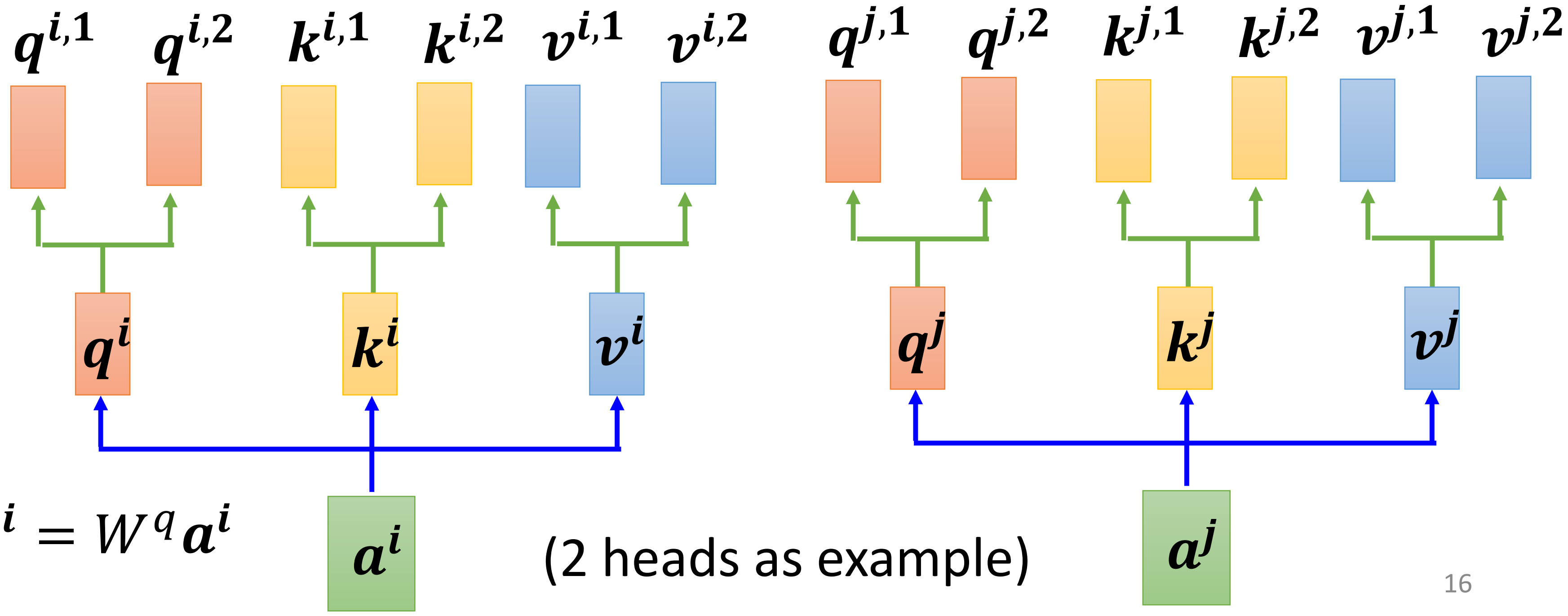
$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention

Different types of relevance

$$b^i = W^O \begin{bmatrix} b^{i,1} \\ b^{i,2} \end{bmatrix}$$



$$q^i = W^q a^i$$

(2 heads as example)