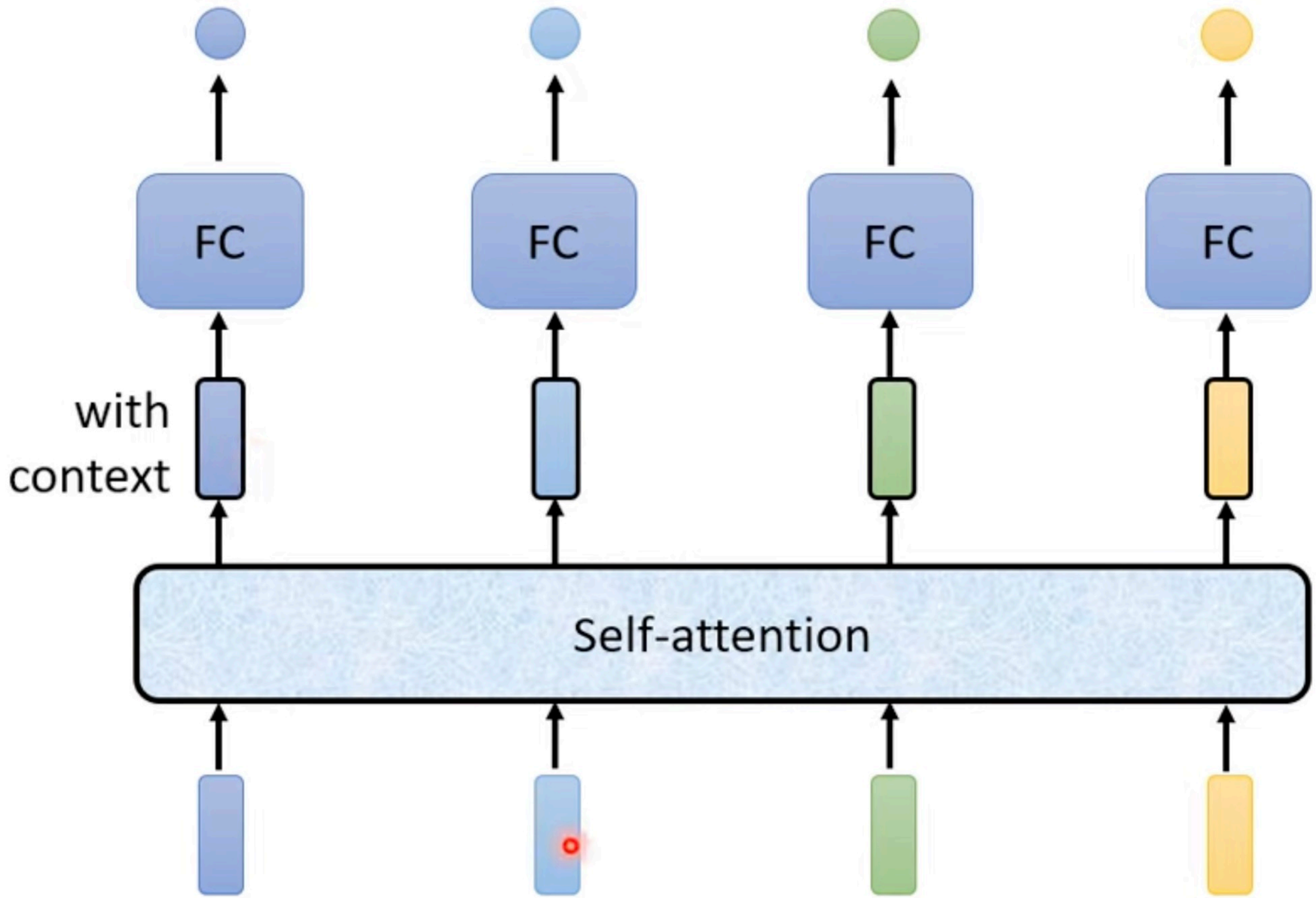
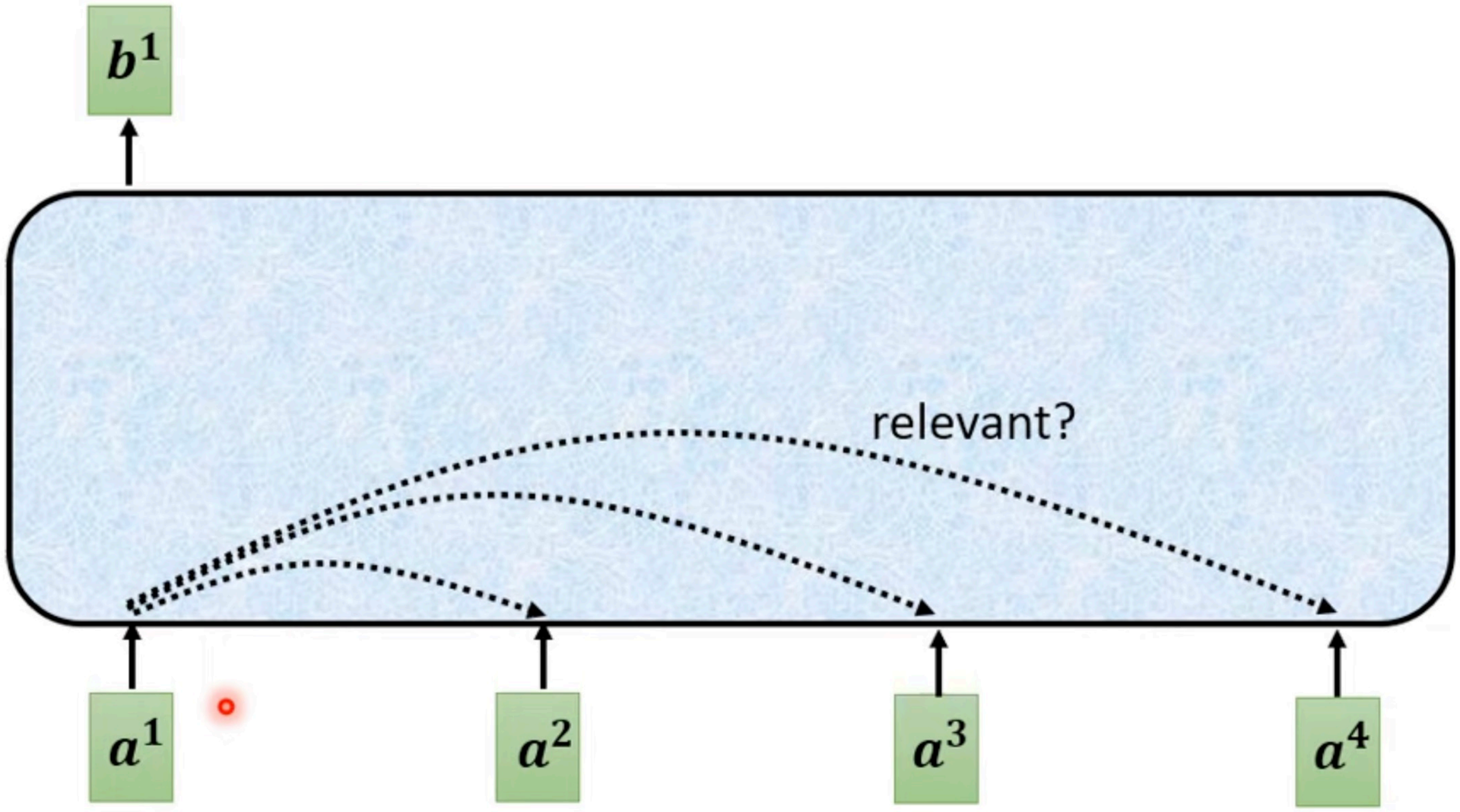
Self-attention



這邊就給你一個另外一個Vector

Self-attention

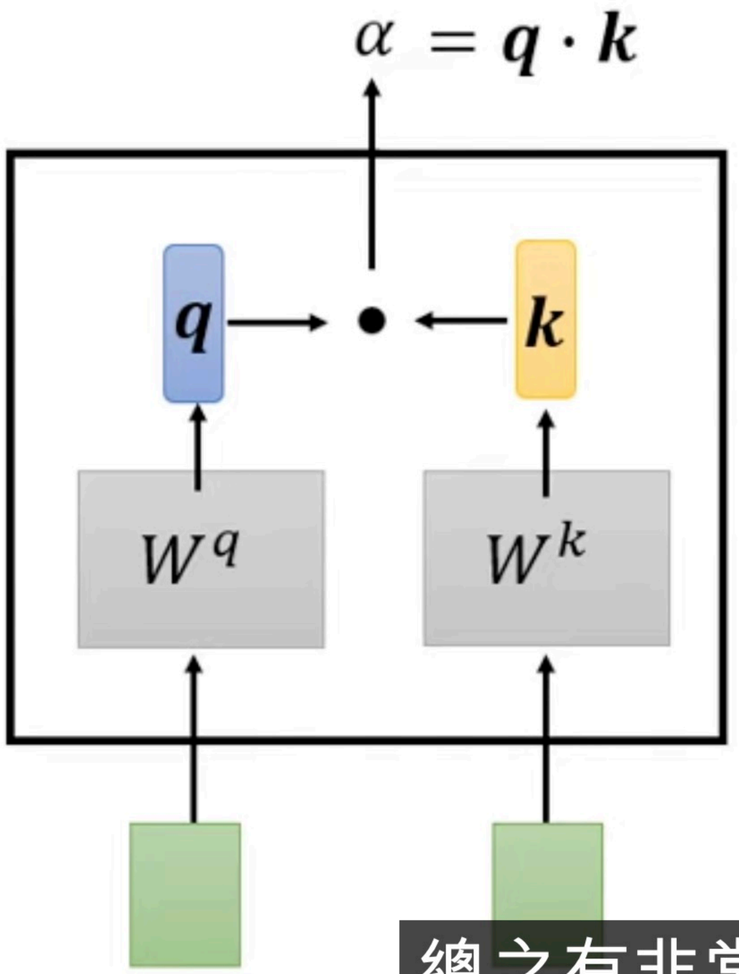


Find the relevant vectors in a sequence

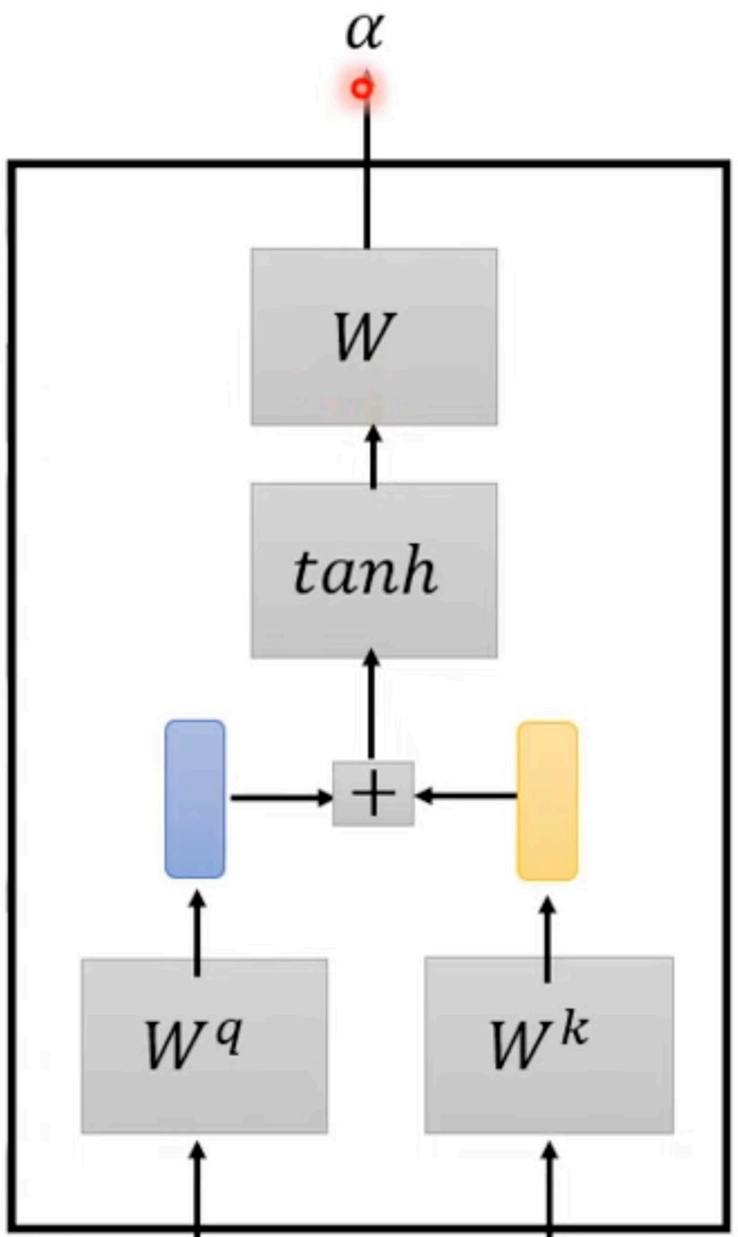
所需要用到的資訊

Self-attention

Dot-product



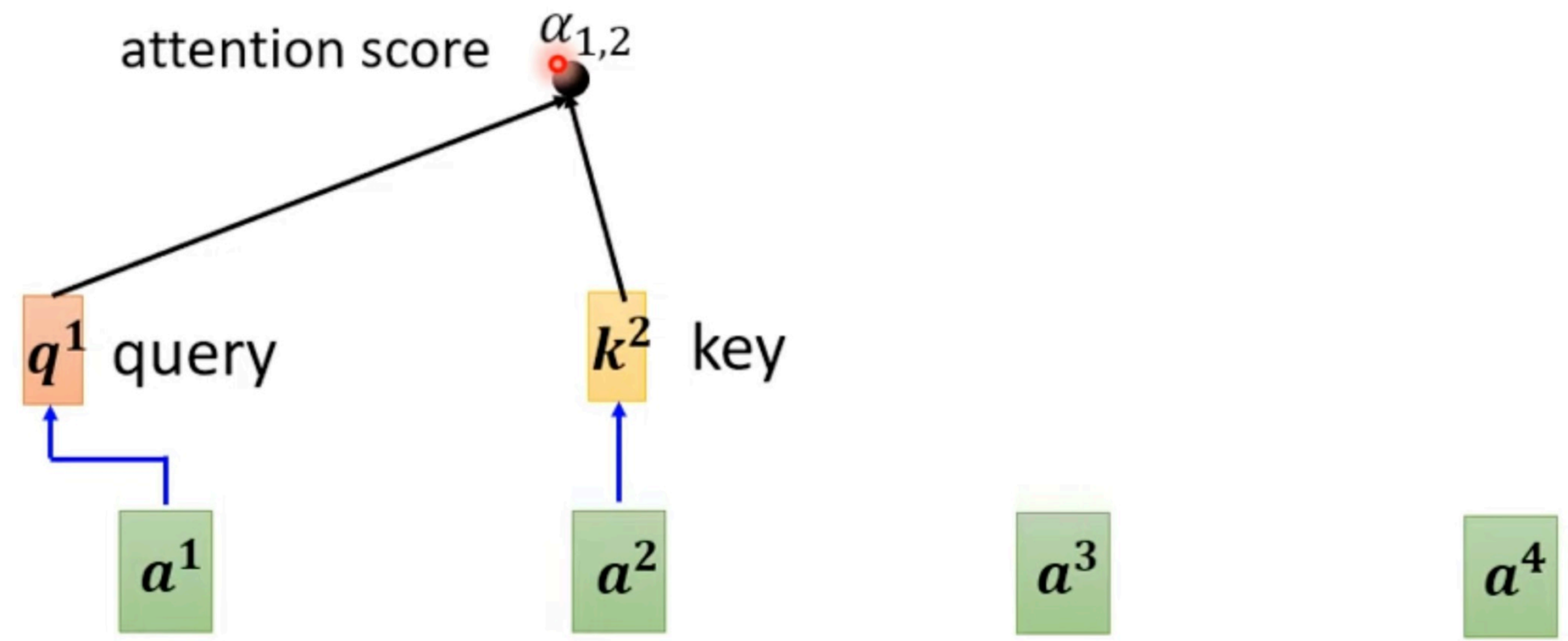
Additive



總之有非常多不同的方法

Self-attention

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

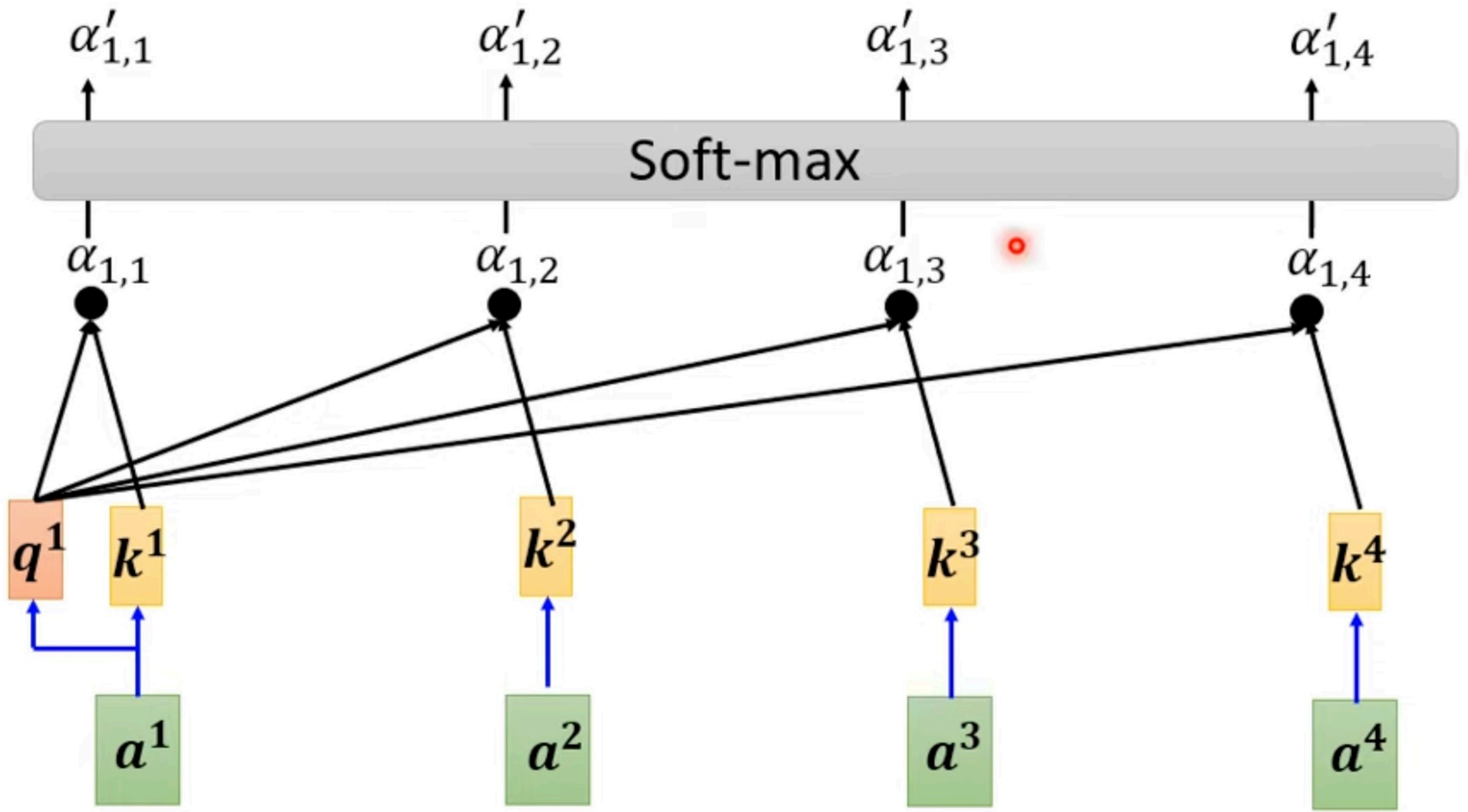


$$\boldsymbol{q^1} = W^q \boldsymbol{a^1} \qquad \boldsymbol{k^2} = W^k \boldsymbol{a^2}$$

叫做Attention的Score

Self-attention

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

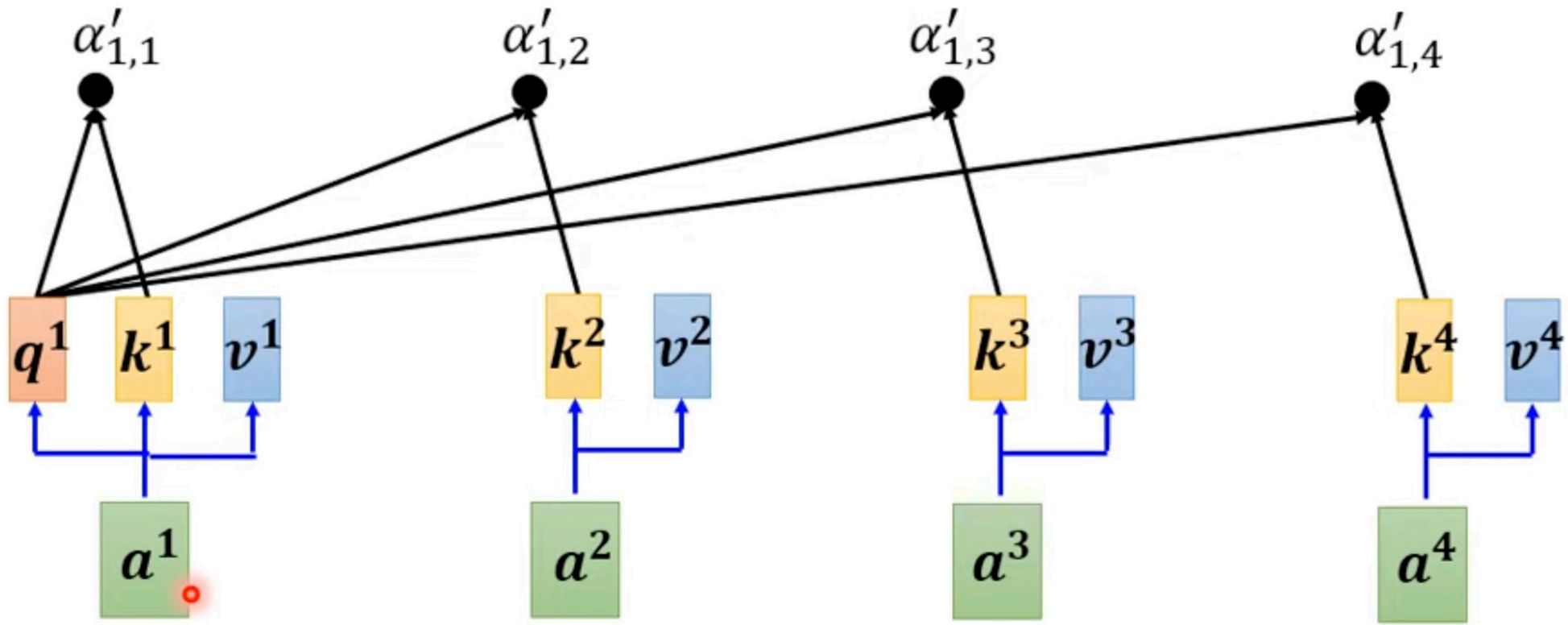


$q^1 = W^q a^1$ $k^2 = W^k a^2$ $k^3 = W^k a^3$ $k^4 = W^k a^4$

$k^1 = W^k a^1$

接下來這邊會作為一個Soft-Max

Self-attention Extract information based on attention scores

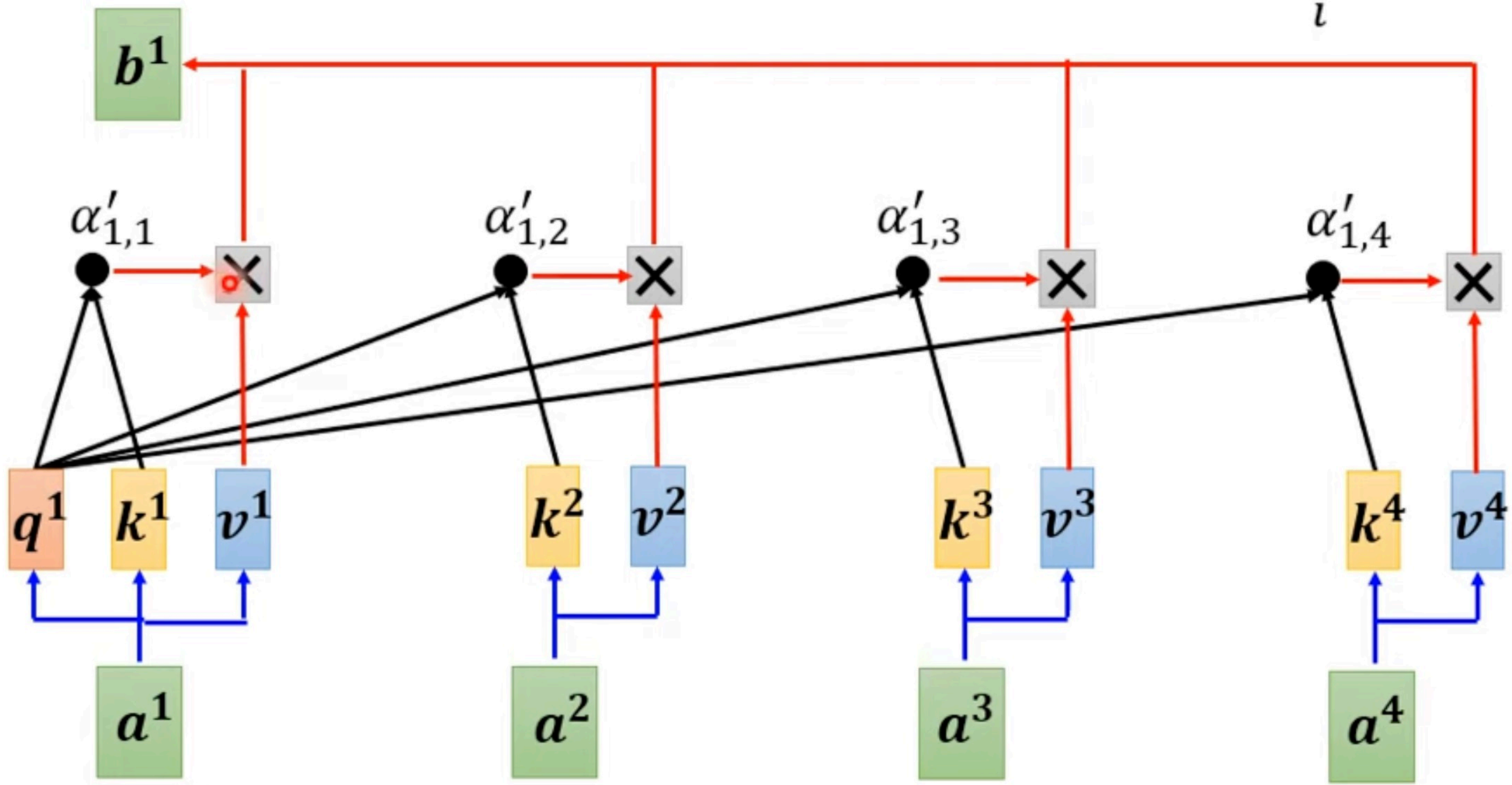


$v^1 = W^v a^1$ $v^2 = W^v a^2$ $v^3 = W^v a^3$ $v^4 = W^v a^4$

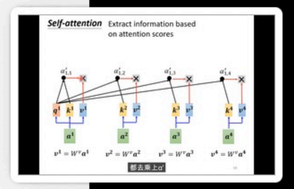
Self-attention

Extract information based on attention scores

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



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



Self-attention

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

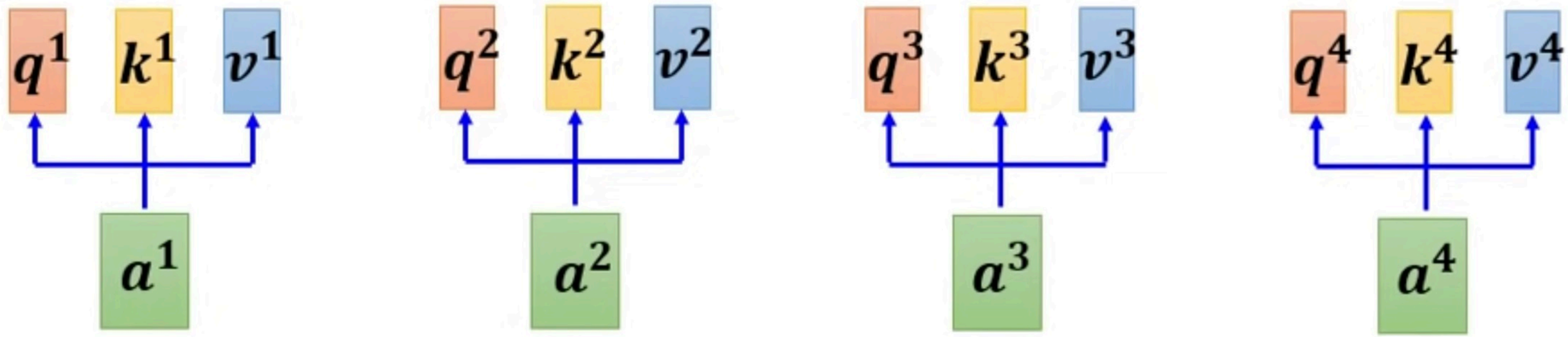
$$\begin{matrix} q^1 & q^2 & q^3 & q^4 \\ Q & & & \end{matrix} = W^q \begin{matrix} a^1 & a^2 & a^3 & a^4 \\ I & & & \end{matrix}$$

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

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

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

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

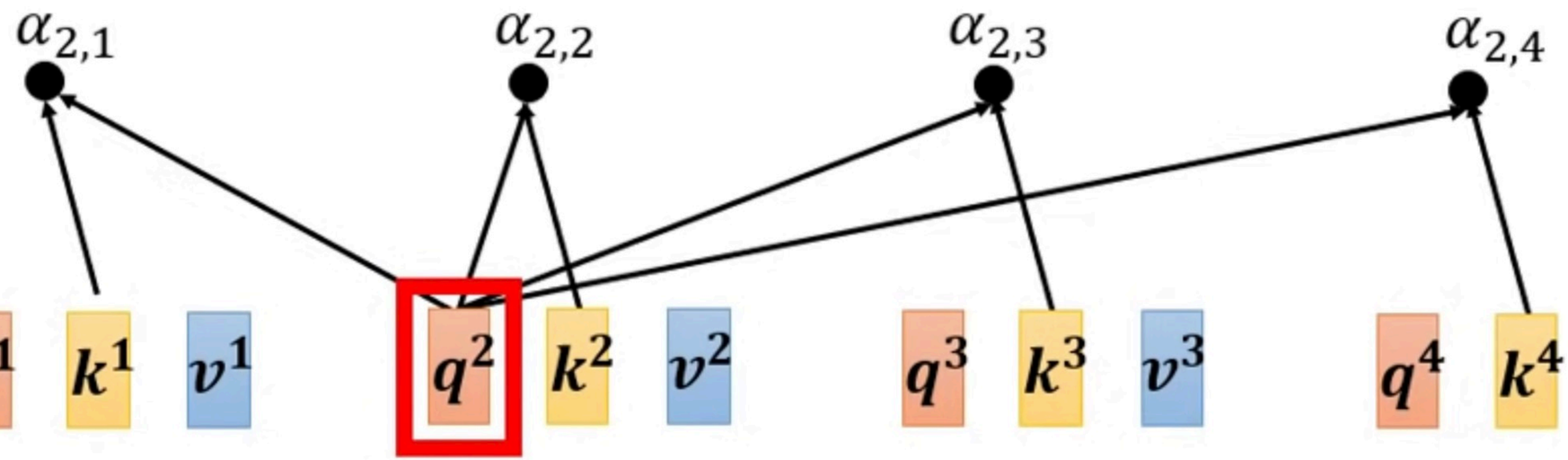


Self-attention

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

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

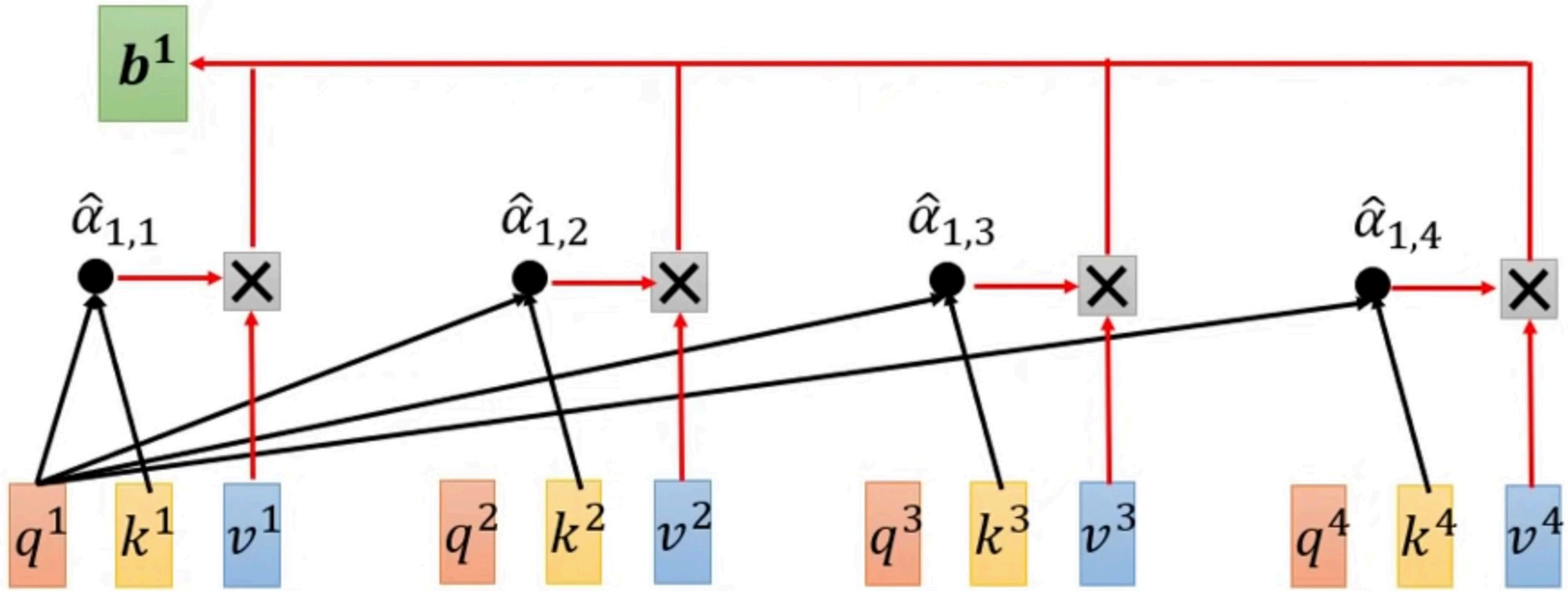
$$\begin{bmatrix}\alpha_{1,1}\\ \alpha_{1,2}\\ \alpha_{1,3}\\ \alpha_{1,4}\end{bmatrix} = \begin{bmatrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{bmatrix} \boldsymbol{q^1}$$



$$\begin{bmatrix}\alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1}\\ \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2}\\ \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3}\\ \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4}\end{bmatrix} \quad A' \qquad \text{softmax} \qquad \begin{bmatrix}\alpha_{1,1} & \alpha_{2,1} & \alpha_{3,1} & \alpha_{4,1}\\ \alpha_{1,2} & \alpha_{2,2} & \alpha_{3,2} & \alpha_{4,2}\\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & \alpha_{4,3}\\ \alpha_{1,4} & \alpha_{2,4} & \alpha_{3,4} & \alpha_{4,4}\end{bmatrix} \quad A$$

$$= \begin{bmatrix}\boldsymbol{k^1}\\ \boldsymbol{k^2}\\ \boldsymbol{k^3}\\ \boldsymbol{k^4}\end{bmatrix} \quad K^T \qquad \begin{matrix}\boldsymbol{q^1} & \boldsymbol{q^2} & \boldsymbol{q^3} & \boldsymbol{q^4}\end{matrix} \quad Q$$

Self-attention



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

Self-attention

$$Q = W^q I$$

$$K = W^k I$$

$$V = W^v I$$

$$A = K^T Q$$

$A \leftarrow A$

Attention Matrix

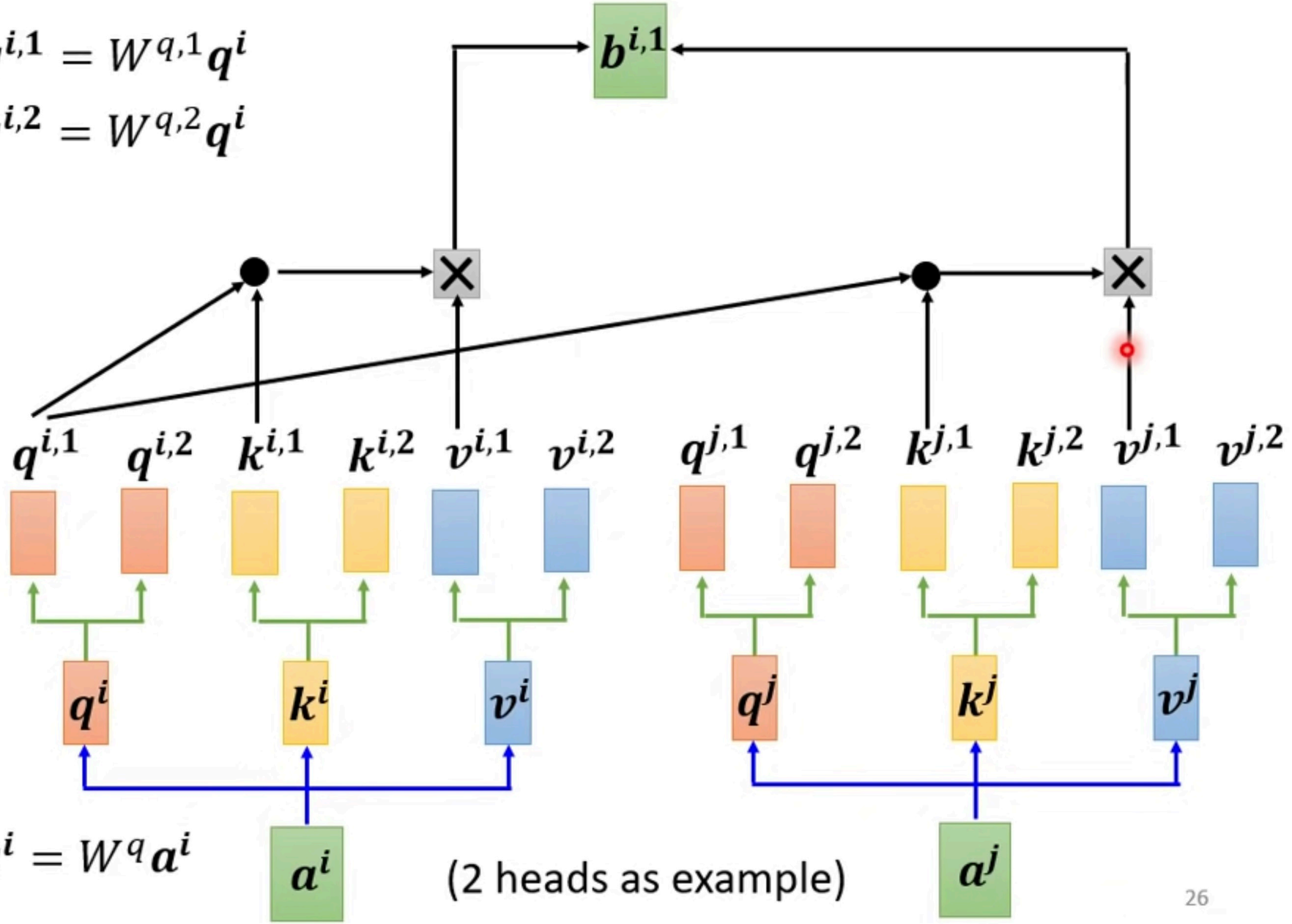
$$O = V A'$$

Multi-head Self-attention

Different types of relevance

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

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

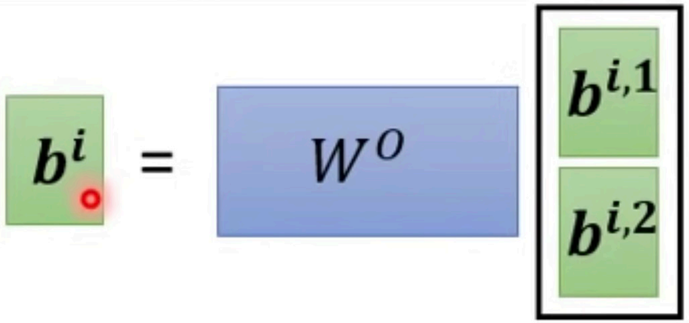


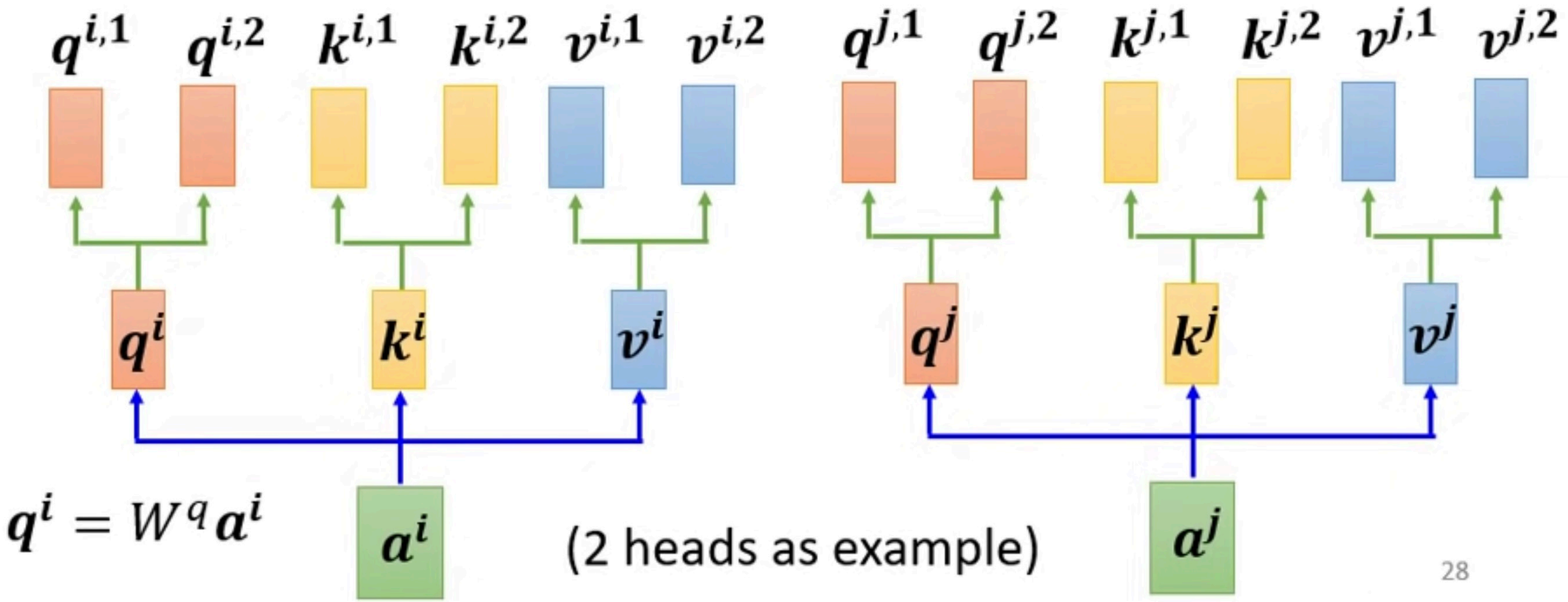
$$q^i = W^q a^i$$

(2 heads as example)

Multi-head Self-attention

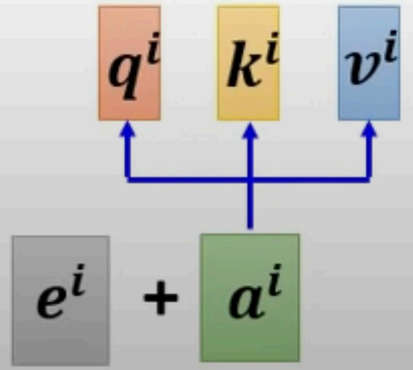
Different types of relevance





Positional Encoding

- No position information in self-attention.
- Each position has a unique positional vector e^i
- **hand-crafted**
- **learned from data**



Each column represents a positional vector e^i