



$$\left\{ \begin{array}{l} W_Q h_2^{(1)} \rightarrow \text{query vector} \\ W_K h_1^{(1)}, W_K h_2^{(1)}, \dots, W_K h_6^{(1)} \rightarrow \text{key vector.} \end{array} \right.$$

Attention weights:

$$\begin{aligned} [W_Q h_1^{(1)}]^T (W_K h_1^{(1)}) &\rightarrow \tilde{a}_1 / \sqrt{d_k} \\ [- \dots -]^T (W_K h_2^{(1)}) &\rightarrow \tilde{a}_2 / \sqrt{d_k} \\ &\vdots \\ [- \dots -]^T (W_K h_6^{(1)}) &\rightarrow \tilde{a}_6 / \sqrt{d_k} \end{aligned} \left. \vphantom{\begin{aligned} [W_Q h_1^{(1)}]^T (W_K h_1^{(1)}) \\ [- \dots -]^T (W_K h_2^{(1)}) \\ \vdots \\ [- \dots -]^T (W_K h_6^{(1)}) \end{aligned}} \right\} \text{softmax} \left. \vphantom{\begin{aligned} [W_Q h_1^{(1)}]^T (W_K h_1^{(1)}) \\ [- \dots -]^T (W_K h_2^{(1)}) \\ \vdots \\ [- \dots -]^T (W_K h_6^{(1)}) \end{aligned}} \right\} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_6 \end{pmatrix}$$

$$= a_1 \cdot (W_V h_1^{(1)}) + a_2 (W_V h_2^{(1)}) + \dots + a_6 (W_V h_6^{(1)})$$

(self) attention vector for one attention head